ANALYSIS & PDE

Volume 18 No. 7 2025

msp

Analysis & PDE

msp.org/apde

PRODUCTION

production@msp.org

Silvio Levy, Scientific Editor

Cover image: Eric J. Heller: "Linear Ramp"

See inside back cover or msp.org/apde for submission instructions.

The subscription price for 2025 is US $475/year for the electronic version, and $735/year (+$70, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscriber address should be sent to MSP.

Analysis & PDE (ISSN 1948-206X electronic, 2157-5045 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online.

APDE peer review and production are managed by EditFlow® from MSP.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

ANALYSIS AND PDE
Vol. 18 (2025), No. 7, pp. 1567–1613
DOI: 10.2140/apde.2025.18.1567

REGULARIZED BRASCAMP–LIEB INEQUALITIES

NEAL BEZ AND SHOHEI NAKAMURA

Given any (forward) Brascamp–Lieb inequality on euclidean space, a famous theorem of Lieb guarantees that gaussian near-maximizers always exist. Recently, Barthe and Wolff used mass transportation techniques to establish a counterpart to Lieb's theorem for all nondegenerate cases of the inverse Brascamp–Lieb inequality. Here we build on work of Chen, Dafnis and Paouris and employ heat-flow techniques to understand the inverse Brascamp–Lieb inequality for certain regularized input functions, in particular extending the Barthe–Wolff theorem to such a setting. Inspiration arose from work of Bennett, Carbery, Christ and Tao for the forward inequality, and we recover their generalized Lieb's theorem using a clever limiting argument of Wolff. In fact, we use Wolff's idea to deduce regularized inequalities in the broader framework of the forward-reverse Brascamp–Lieb inequality, in particular allowing us to recover the gaussian saturation property in this framework first obtained by Courtade, Cuff, Liu and Verdú.

1. Introduction

The main subject of this paper is the inverse Brascamp–Lieb inequality

$$\int_{\mathbb{R}^n} e^{-\pi\langle x,\mathcal{Q}x\rangle} \prod_{j=1}^m f_j(B_j x)^{c_j}\,dx \geq C \prod_{j=1}^m \bigg(\int_{\mathbb{R}^{n_j}} f_j\bigg)^{c_j} \tag{1-1}$$

associated with a self-adjoint linear transformation $\mathcal{Q}:\mathbb{R}^n\to\mathbb{R}^n$, and families of linear transformations $B_j:\mathbb{R}^n\to\mathbb{R}^{n_j}$ and exponents $c_j\in\mathbb{R}$. Largely thanks to recent work of Barthe and Wolff [2014; 2022] and Chen, Dafnis and Paouris [Chen et al. 2015], much is known about inverse Brascamp–Lieb inequalities for *general* input functions f_j. For example, gaussian near-minimizers always exist for (1-1) in all nondegenerate situations; this was established in full generality in [Barthe and Wolff 2022] (using mass transportation) and in certain special cases in [Chen et al. 2015] (using heat flow in the spirit of the work of Carlen, Lieb and Loss [Carlen et al. 2004] and Bennett, Carbery, Christ and Tao [Bennett et al. 2008] on the forward Brascamp–Lieb inequality). Such a result is a counterpart to Lieb's famous theorem for the forward version of (1-1). In the present work, we investigate what happens when the input functions f_j are "regularized" in the sense that they coincide with the evolution of positive finite Borel measures under certain heat equations, and our main result is a regularized version of the aforementioned theorem of Barthe and Wolff.

Our motivation to pursue this project arose from several directions. Firstly, a regularized version of the forward Brascamp–Lieb inequality was considered in [Bennett et al. 2008] and, in particular, they

This work was supported by JSPS Kakenhi grant numbers 18KK0073, 19H00644, 19H01796 and 23H01080 (Bez), and Grant-in-Aid for JSPS Research Fellow number 17J01766 and JSPS Kakenhi grant numbers 19K03546, 19H01796 and 21K13806 (Nakamura).

MSC2020: primary 42B37; secondary 44A12, 52A40.

Keywords: Brascamp–Lieb inequality, heat flow.

were able to provide a completely new proof of Lieb's theorem based on heat flow (see the forthcoming Theorem 2.3 for a version which incorporates a gaussian kernel). This approach to Lieb's theorem was an important source of inspiration to us and we shall see that our analysis of (1-1) for regularized inputs will naturally yield the inverse version of Lieb's theorem in full generality. In this sense, we follow [Chen et al. 2015] by adopting heat-flow techniques for the inverse inequality, extend their gaussian saturation result to full generality, and thus rederive the gaussian saturation result of [Barthe and Wolff 2022] by heat-flow regularization. The special case considered in [Chen et al. 2015] is a certain "geometric" version of the inverse Brascamp–Lieb inequality (see Section 3 for further details) and the extension to the general case is far from straightforward.

More broadly speaking, the act of regularizing an inequality by restricting the input functions to a smaller subclass of sufficiently well-behaved functions is natural and ubiquitous. One potential benefit of doing so is to raise the possibility of establishing the existence of extremizers for the inequality amongst the restricted subclass, and in doing so may open up a fruitful approach to analyze the general form of the inequality. As we shall see more precisely later, this is the philosophy behind the aforementioned heat-flow proof of Lieb's theorem in [Bennett et al. 2008], where general input functions are treated as limits of solutions to heat equations at time zero.

Another virtue of restricting the inputs to regularized functions is that one may obtain an improved form of the inequality. Such a perspective is often taken up in fields such as convex geometry, differential geometry, probability and information theory, in which case the regularization is often described in terms of semi/uniform log-concavity and log-convexity. For example, a famous conjecture of Talagrand predicts that Markov's inequality can be improved if one restricts to regularized functions; see, for example, [Eldan and Lee 2018]. This conjecture was originally stated for the discrete cube and as far as we are aware it is still an open problem in that form, but its analogue for gaussian space has been affirmatively solved in [Eldan and Lee 2018].

In differential geometry, one can use functional inequalities involving entropy in order to describe the curvature of a Riemannian manifold; this is based on fundamental work of Bakry and Emery and we refer the reader to the book by Bakry, Gentil and Ledoux [Bakry et al. 2014] for more details. One example of such an inequality is the (local) log-Sobolev inequality on a weighted Riemannian manifold [Bakry et al. 2014, Theorem 5.5.2]. More recently, several papers have highlighted the importance of improving the log-Sobolev inequality by restricting inputs to regularized functions, which can be regarded as a certain "curvature improvement". As far as we are aware, the first example of this kind can be found in the work of Fathi, Indrei and Ledoux [Fathi et al. 2016], where they improved the best constant for the log-Sobolev inequality under regularization in terms of the Poincaré constant. Along this line, very recently Aishwarya and Rotem [2023] established an improvement of the convexity of entropy under regularization in terms of uniform log-concavity, and related work can also be found in work of Eldan, Lehec and Shenfeld [Eldan et al. 2020] and Bez, Nakamura and Tsuji [Bez et al. 2023]. Another important entropic inequality in information theory is the Shannon–Stam inequality, which is a special case of the entropic Brascamp–Lieb inequality (see [Carlen et al. 2004; Carlen and Cordero-Erausquin 2009]). Inspired by important work of Ball, Barthe and Naor [Ball et al. 2003] and Ball and Nguyen [2012] on

entropy jump inequalities, the stability problem for the Shannon–Stam inequality has recently attracted attention. In this direction, the regularized framework has played a role; for instance, Courtade, Fathi and Pananjady [Courtade et al. 2018] and Eldan and Mikulincer [2020] established stability estimates for the Shannon–Stam inequality for uniformly log-concave random variables.

In convex geometry, regularization appears in terms of the convexity of the boundary of a convex body. We mention [Schmuckenschläger 1995] (see also [Klartag and Milman 2008]) where regularized convex bodies (2-uniformly convex bodies) were investigated and, in particular, Bourgain's celebrated hyperplane conjecture for such convex bodies was established. Furthermore, the second author, together with Tsuji, pointed out a close link between the Brascamp–Lieb inequality and the volume product of a convex body [Nakamura and Tsuji 2022; 2024]. In particular, they developed ideas which arose out of consideration of the regularized Brascamp–Lieb inequality, and succeeded in confirming Mahler's conjecture for certain regularized convex bodies (see [Nakamura and Tsuji 2022]).

Finally we mention that a regularized version of the forward Brascamp–Lieb inequality of a different (rougher) nature has recently been obtained by Maldague [2022] (see also [Zorin-Kranich 2020]), and has found applications to multilinear Kakeya-type inequalities and decoupling theory for the Fourier transform in work of Guo, Oh, Zhang and Zorin-Kranich [Guo et al. 2023]. The regularization in [Maldague 2022] has a rough cut-off instead of a gaussian weight factor and the input functions f_j are constant on cubes in the unit cube lattice.

Before presenting our main results in Section 2, to help set the scene we give an introduction and some background regarding Brascamp–Lieb inequalities. In line with the historical development of the subject, we begin with the forward Brascamp–Lieb inequality. After this, we discuss recent developments regarding the inverse inequality (1-1). Finally, we introduce the so-called forward-reverse Brascamp–Lieb inequality which originated in work of Liu, Courtade, Cuff and Verdú [Liu et al. 2018] and whose theory was significantly developed in [Courtade and Liu 2021]. The forward-reverse Brascamp–Lieb framework encompasses both the forward and inverse Brascamp–Lieb inequalities, as well as the reverse Brascamp–Lieb inequality due to Barthe [1997; 1998b]. However, it is in fact the case that the gaussian saturation property for the inverse inequality (1-1) implies the gaussian saturation property for the forward-reverse Brascamp–Lieb inequality. This follows from a clever argument of Wolff (e.g., see [Courtade and Liu 2021, Section 4]), and we shall make use of Wolff's idea to deduce regularized forward-reverse Brascamp–Lieb inequalities from our main result for regularized inverse Brascamp–Lieb inequalities (thus, for example, recovering regularized forward Brascamp–Lieb inequalities due to [Bennett et al. 2008]).

1.1. ***The forward Brascamp–Lieb inequality.*** Inequalities of the form

$$\int_{\mathbb{R}^n} \prod_{j=1}^m f_j(B_j x)^{c_j}\, dx \le C \prod_{j=1}^m \Big(\int_{\mathbb{R}^{n_j}} f_j \Big)^{c_j} \tag{1-2}$$

for integrable functions $f_j : \mathbb{R}^{n_j} \to \mathbb{R}_+$ are widely known as *Brascamp–Lieb inequalities*.[1] Here, the linear transformations $B_j : \mathbb{R}^n \to \mathbb{R}^{n_j}$ and positive exponents c_j are given, and the pair $(\boldsymbol{B}, \boldsymbol{c})$ is referred

[1]We often add the term "forward" to clarify that we are referring to (1-2) rather than (1-1).

to as a *Brascamp–Lieb datum*, where $\boldsymbol{B} = (B_j)_{j=1}^m$ and $\boldsymbol{c} = (c_j)_{j=1}^m$. The best (i.e., smallest) constant C is called the *Brascamp–Lieb constant* and is defined by

$$\mathrm{F}(\boldsymbol{B}, \boldsymbol{c}) = \sup_{\boldsymbol{f}} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}; \boldsymbol{f}) \in (0, \infty],$$

where

$$\mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}; \boldsymbol{f}) = \frac{\int_{\mathbb{R}^n} \prod_{j=1}^m f_j(B_j x)^{c_j}\, dx}{\prod_{j=1}^m \big(\int_{\mathbb{R}^{n_j}} f_j\big)^{c_j}}$$

is the *Brascamp–Lieb functional*, and the supremum is taken over measurable $f_j : \mathbb{R}^{n_j} \to \mathbb{R}_+$ such that $\int_{\mathbb{R}^{n_j}} f_j \in (0, \infty)$.

The multilinear Hölder, Loomis–Whitney and Young convolution inequalities are standard examples of Brascamp–Lieb inequalities. From a historical perspective, the pursuit of the best constant for the Young convolution inequality was influential in the emergence of the Brascamp–Lieb inequality. Indeed, Beckner [1975] and Brascamp and Lieb [1976] independently identified that the best constant in nonboundary cases of the Young convolution inequality is attained (essentially uniquely so — see [Brascamp and Lieb 1976]) on certain isotropic centred gaussians, and the systematic study of the more general inequality (1-2) traces back to [loc. cit.]. With the main focus of the current paper in mind, we also remark that [loc. cit.] considered the inverse form of the Young convolution inequality in which the direction of the inequality is reversed (and negative exponents c_j are admissible). They were able to obtain the best (i.e., largest) constant, a characterization of maximizers and, as a delightful application, were able to rederive the Prékopa–Leindler inequality via a clever limiting argument.

Nowadays the theory of the Brascamp–Lieb inequality is well developed and finds applications across a vastly diverse range of fields. As but one example, we mention again the close link to multilinear Kakeya-type inequalities and decoupling theory for the Fourier transform, both of which have found staggering applications in the last 15 years to problems in harmonic analysis, geometric analysis, dispersive PDEs, and number theory; see, for example [Bourgain 2017; Bourgain et al. 2016; Bourgain and Guth 2011; Guo and Zhang 2019; Guo and Zorin-Kranich 2020]. Whilst it feels somewhat discourteous not to include details of other kinds of applications of the Brascamp–Lieb inequality (and its variants), there are a number of papers which already contain thorough discussions of this nature and we encourage the reader to try [Bennett and Bez 2021; Bennett et al. 2020; Durcik and Thiele 2021; Gardner 2002; Zhang 2022].

Unlike (nonboundary cases of) the Young convolution inequality, the classes of maximizers for the multilinear Hölder and Loomis–Whitney inequalities are of a wider nature. Nevertheless, it is the case that isotropic centred gaussians are amongst the maximizers for each of these inequalities and this naturally raises the question of whether this is a general phenomenon for (1-2). Remarkably, Lieb [1990] established that centred gaussian *near maximizers* always exist for (1-2). We state this result next, along with a "localized" version incorporating a centred gaussian weight factor. For this, we introduce a little more notation.

We extend the notion of Brascamp–Lieb datum to the triple $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$, where $\mathcal{Q} : \mathbb{R}^n \to \mathbb{R}^n$ is a given self-adjoint transformation. Associated to such a datum is the Brascamp–Lieb constant

$$\mathrm{F}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}) = \sup_{\boldsymbol{f}} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{f}) \in (0, \infty],$$

which is given in terms of the Brascamp–Lieb functional

$$\mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{f}) = \frac{\int_{\mathbb{R}^n} e^{-\pi\langle x, \mathcal{Q}x\rangle} \prod_{j=1}^m f_j(B_j x)^{c_j}\, dx}{\prod_{j=1}^m \big(\int_{\mathbb{R}^{n_j}} f_j\big)^{c_j}}.$$

In the above, the supremum is taken over measurable $f_j : \mathbb{R}^{n_j} \to \mathbb{R}_+$ such that $\int_{\mathbb{R}^{n_j}} f_j \in (0, \infty)$. Also, associated to a positive definite transformation A on a given euclidean space, we define the (normalized) centred gaussian g_A by

$$g_A(x) := (\det A)^{1/2} e^{-\pi\langle x, Ax\rangle}. \tag{1-3}$$

On gaussian inputs $\boldsymbol{f}$ with $f_j = g_{A_j}$ for each $j = 1, \dots, m$, we slightly abuse notation for the Brascamp–Lieb functional and write

$$\mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}) := \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{f}).$$

Theorem 1.1 (Lieb). *Let $(\boldsymbol{B}, \boldsymbol{c})$ be a Brascamp–Lieb datum. Then we have*

$$\mathrm{F}(\boldsymbol{B}, \boldsymbol{c}) = \sup_{\boldsymbol{A}} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}; \boldsymbol{A}),$$

where the supremum is taken over positive definite transformations A_j. More generally

$$\mathrm{F}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}) = \sup_{\boldsymbol{A}} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}) \tag{1-4}$$

whenever $\mathcal{Q} : \mathbb{R}^n \to \mathbb{R}^n$ is positive semidefinite.

Lieb's theorem above was proved in full generality in [Lieb 1990]. Earlier, Brascamp and Lieb [1976] had established certain special cases including, for instance, rank-1 linear transformations B_j. A computation reveals that

$$\mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}) = \left(\frac{\prod_{j=1}^m (\det A_j)^{c_j}}{\det\big(\mathcal{Q} + \sum_{j=1}^m c_j B_j^* A_j B_j\big)} \right)^{1/2}$$

if $\mathcal{Q} + \sum_{j=1}^m c_j B_j^* A_j B_j$ is positive definite ($\mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}) = \infty$ otherwise) and therefore Lieb's theorem dramatically reduces the complexity of understanding the Brascamp–Lieb constant. For example, Lieb's theorem played a pivotal role in the proof in [Bennett et al. 2017] of the continuity of the Brascamp constant $\boldsymbol{B} \mapsto \mathrm{F}(\boldsymbol{B}, \boldsymbol{c})$ and consequently in recent developments in understanding nonlinear variants of the Brascamp–Lieb inequality in which the underlying transformations B_j are allowed to be nonlinear [Bennett et al. 2020].

We end this very brief overview of the Brascamp–Lieb inequality (1-2) by stating a theorem from [Bennett et al. 2008; 2010] which provides a characterization of the *finiteness* of the Brascamp–Lieb constant (see also [Gressman 2025] for new perspectives in this direction).

Theorem 1.2 (Bennett–Carbery–Christ–Tao). *Let $(\boldsymbol{B}, \boldsymbol{c})$ be a Brascamp–Lieb datum. Then $\mathrm{F}(\boldsymbol{B}, \boldsymbol{c})$ is finite if and only if*

$$n = \sum_{j=1}^m c_j n_j \tag{1-5}$$

and

$$\dim V \le \sum_{j=1}^{m} c_j \dim B_j V \quad \textit{for all subspaces } V \textit{ of } \mathbb{R}^n. \tag{1-6}$$

The above result along with Lieb's theorem form two pillars in the theory of the Brascamp–Lieb inequality. The scaling condition (1-5) is easily shown to be necessary for the Brascamp–Lieb constant to be finite. Also, as further necessary conditions for finiteness, by considering $V = \mathbb{R}^n$ and $V = \bigcap_{j=1}^m \ker B_j$, one obtains from (1-5) and (1-6) that each B_j must be surjective and that $\bigcap_{j=1}^m \ker B_j = \{0\}$; such data are usually referred to as *nondegenerate*.

Remark. Maldague's regularized version of (1-2) quantifies the finiteness characterization in Theorem 1.2. In particular, it is shown in [Maldague 2022] that when the integral on the left-hand side of (1-2) is truncated to a ball of radius $R > 0$ and the input functions are restricted to be constant on cubes in a lattice of size $r \in (0, R)$, the constant behaves like $R^{\kappa} r^{-\tilde{\kappa}}$, where $\kappa = \sup_{V \le \mathbb{R}^n} \big(\dim V - \sum_{j=1}^m c_j \dim B_j V\big)$ and $\tilde{\kappa} := \kappa - \big(n - \sum_{j=1}^m c_j n_j\big)$. On the other hand, the regularized version of (1-2) proved in [Bennett et al. 2008, Corollary 8.15] may be viewed as an extension of Lieb's theorem, and our main result (Theorem 2.1 below) is an extension of the inverse version of Lieb's theorem.

1.2. ***The inverse Brascamp–Lieb inequality.*** For the inverse Brascamp–Lieb inequality (1-1), it is appropriate to introduce a different notion of nondegeneracy compared with the forward version of the inequality. The nondegeneracy condition was identified by Barthe and Wolff [2022] and is as follows. Assume each $B_j : \mathbb{R}^n \to \mathbb{R}^{n_j}$ is a surjection, and $\boldsymbol{c}$ is arranged so that

$$c_1, \dots, c_{m_+} > 0 > c_{m_+ + 1}, \dots, c_m$$

for some $0 \le m_+ \le m$, where we interpret the case $m_+ = 0$ to mean that all of $c_1, \dots, c_m$ are negative. Correspondingly, we define the linear transformation $\boldsymbol{B}_+ : \mathbb{R}^n \to \bigoplus_{j=1}^{m_+} \mathbb{R}^{n_j}$ by

$$\boldsymbol{B}_+ x := (B_1 x, \dots, B_{m_+} x).$$

Then, if we denote the number of positive eigenvalues of the self-adjoint transformation $\mathcal{Q} : \mathbb{R}^n \to \mathbb{R}^n$ by $s^+(\mathcal{Q})$, the Brascamp–Lieb datum[2] $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$ is said to be *nondegenerate* if

$$\mathcal{Q}|_{\ker \boldsymbol{B}_+} > 0 \quad \text{and} \quad n \ge s^+(\mathcal{Q}) + \sum_{j=1}^{m_+} n_j \tag{1-7}$$

hold. Here $\mathcal{Q}|_{\ker \boldsymbol{B}_+}$ is the restriction of $\mathcal{Q} : \mathbb{R}^n \to \mathbb{R}^n$ to the subspace $\ker \boldsymbol{B}_+$. Note that if $\mathcal{Q} = 0$ then the nondegeneracy condition (1-7) is equivalent to fact that $\boldsymbol{B}_+$ is a bijection; in other words,

$$\bigcap_{j=1}^{m_+} \ker B_j = \{0\} \quad \text{and} \quad \operatorname{Im} \boldsymbol{B}_+ = \bigoplus_{j=1}^{m_+} \mathbb{R}^{n_j}. \tag{1-8}$$

[2]Strictly speaking, this terminology has already been used for the forward Brascamp–Lieb inequality where we imposed the condition $c_j > 0$ for all j. It would be more accurate to use terminology such as "inverse Brascamp–Lieb datum", but it will always be clear from the context which notion of Brascamp–Lieb datum is being used.

Next, in analogy with the notation introduced above for the forward Brascamp–Lieb inequality, we define

$$\mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}) := \inf_{\boldsymbol{f}} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{f})$$

to be the best (i.e., largest) constant for which (1-1) holds, where the infimum is taken over measurable $f_j : \mathbb{R}^{n_j} \to \mathbb{R}_+$ such that $\int_{\mathbb{R}^{n_j}} f_j \in (0, \infty)$. Strictly speaking, we are extending our definition of the Brascamp–Lieb functional $\boldsymbol{f} \mapsto \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{f})$ to the case of real exponents $\boldsymbol{c} \subseteq \mathbb{R}^m$, which we do with the understanding that $0 \cdot \infty = 0$.

It is not immediately apparent that (1-7) is a natural nondegeneracy condition to impose and we refer the reader to the careful discussion in [Barthe and Wolff 2022, Section 2] for the details. Here we simply extract from that work that $\mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}) = 0$ or ∞ when the nondegeneracy condition is dropped.

The following inverse version of Lieb's theorem was proved in [loc. cit., Theorem 2.9].

Theorem 1.3 (Barthe–Wolff). *For any nondegenerate Brascamp–Lieb datum* $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$, *we have*

$$\mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}) = \inf_{\boldsymbol{A}} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}).$$

For completeness, we note that a counterpart to Theorem 1.2 regarding the strict positivity of $\mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$ was proved by Barthe–Wolff; see [loc. cit., Theorem 1.5] for the case $\mathcal{Q} = 0$ and [loc. cit., Theorem 8.9] for the general case.

1.3. ***The forward-reverse Brascamp–Lieb inequality.*** Next we introduce the broad class of inequalities first considered in [Liu et al. 2018] called *forward-reverse Brascamp–Lieb inequalities*. To do so, we fix linear transformations $T_j : E \to E(j),\ j = 1, \dots, J$, with $E(j) = \mathbb{R}^{n(j)}$ and where the base space is given by $E = \bigoplus_{i=1}^{I} E_i$ with $E_i = \mathbb{R}^{n_i},\ i = 1, \dots, I$. The collection of such linear transformations is written $\boldsymbol{T} = (T_j)_{j=1}^{J}$, and also we introduce the notation π_i for the orthogonal projection from E to E_i. Finally, we fix two collections of positive real numbers $(d_i)_{i=1}^{I}$ and $(d(j))_{j=1}^{J}$, and write $\boldsymbol{d} = ((d_i)_{i=1}^{I}, (d(j))_{j=1}^{J})$.

The forward-reverse Brascamp–Lieb inequality associated with the datum $(\boldsymbol{T}, \boldsymbol{d})$ takes the form

$$\prod_{i=1}^{I} \Big(\int_{E_i} f_i \Big)^{d_i} \le C \prod_{j=1}^{J} \Big(\int_{E(j)} h_j \Big)^{d(j)} \tag{1-9}$$

for input functions $f_i : E_i \to \mathbb{R}_+,\ h_j : E(j) \to \mathbb{R}_+$ which satisfy

$$\prod_{i=1}^{I} f_i(\pi_i x)^{d_i} \le \prod_{j=1}^{J} h_j(T_j x)^{d(j)} \quad \text{for all } x \in E. \tag{1-10}$$

As described in [Courtade and Liu 2021], this framework encompasses the forward, reverse and inverse forms of the Brascamp–Lieb inequality in the case where there is no gaussian kernel. For example, taking $I = 1$ and $d_1 = 1$, it is clear that the optimal choice in (1-10) is

$$f = \prod_{j=1}^{J} (h_j \circ T_j)^{d(j)},$$

in which case (1-9) reduces to (1-2).

Also, the reverse form of the Brascamp–Lieb inequality, originating in [Barthe 1998b], corresponds to the case $J = 1$ and $d(1) = 1$. In other words, given linear mappings $B_i : \mathbb{R}^n \to \mathbb{R}^{n_i}$ and positive real numbers $d_i,\ i = 1, \dots, I$, the reverse Brascamp–Lieb inequality takes the form

$$C \prod_{i=1}^{I} \Big(\int_{\mathbb{R}^{n_i}} f_i \Big)^{d_i} \leq \int_{\mathbb{R}^n} h_1 \tag{1-11}$$

for functions satisfying

$$\prod_{i=1}^{I} f_i(x_i)^{d_i} \leq h_1\Big(\sum_{i=1}^{I} d_i B_i^* x_i \Big) \quad \text{for } (x_1, \dots, x_I) \in \bigoplus_{i=1}^{I} \mathbb{R}^{n_i}. \tag{1-12}$$

As has been pointed out elsewhere, arguably it would be more natural to refer to inequalities of the form (1-11) as *dual* (rather than reverse) Brascamp–Lieb inequalities in light of the fact that

$$\mathrm{R}(\boldsymbol{B}, \boldsymbol{d})\mathrm{F}(\boldsymbol{B}, \boldsymbol{d}) = 1, \tag{1-13}$$

where $\mathrm{R}(\boldsymbol{B}, \boldsymbol{d})$ denotes the best (i.e., largest) constant C such that (1-11) holds under (1-12). This remarkable fact was proved in [Barthe 1998b].

We also remark that the Prékopa–Leindler inequality is the special case of (1-11) with $C = 1$ obtained by taking $I = 2$, $B_1 = B_2 = \mathrm{id}_{\mathbb{R}^n}$, and $d_1 + d_2 = 1$. From the Prékopa–Leindler inequality one can derive the famous Brunn–Minkowski inequality

$$\mathrm{vol}_n(X + Y)^{1/n} \geq \mathrm{vol}_n(X)^{1/n} + \mathrm{vol}_n(Y)^{1/n}$$

for appropriate $X, Y \subseteq \mathbb{R}^n$. The stimulus to introduce the general form of the inequality (1-11) appears to have arisen from convex geometry and we refer the reader to [Ball 1991; Barthe 1998a; 1998b] for further discussion and applications.

Finally, to deduce the inverse Brascamp–Lieb inequality, take a forward-reverse Brascamp–Lieb datum $(\boldsymbol{T}, \boldsymbol{d})$ and set

$$h_{J+1}(x) = \prod_{i=1}^{I} f_i(\pi_i x)^{d_i} \prod_{j=1}^{J} h_j(T_j x)^{-d(j)}.$$

Then (1-10) holds if we replace J by $J + 1$, take T_{J+1} to be the identity transformation, and $d_{J+1} = 1$. The resulting inequality (1-9) reduces to the inverse Brascamp–Lieb inequality

$$\prod_{i=1}^{I} \Big(\int_{E^i} f_i \Big)^{d_i} \prod_{j=1}^{J} \Big(\int_{E(j)} h_j \Big)^{-d(j)} \leq C \int_E \prod_{i=1}^{I} f_i(\pi_i x)^{d_i} \prod_{j=1}^{J} h_j(T_j x)^{-d(j)}\, dx.$$

With the nondegeneracy condition in mind (i.e., $\boldsymbol{B}_+$ is bijective), this framework is as general as the one presented in the previous section in the case of no gaussian kernel.[3]

The analogue of Lieb's theorem holds in the setting of the forward-reverse Brascamp–Lieb inequality, as shown in [Liu et al. 2018, Theorem 2].

[3]We also refer the reader to the Appendix for a more complete discussion along these lines.

Theorem 1.4 (Liu–Courtade–Cuff–Verdú). *For any datum* $(\boldsymbol{T}, \boldsymbol{d})$, *if the input functions* $f_i : E_i \to \mathbb{R}_+$, $h_j : E(j) \to \mathbb{R}_+$ *satisfy* (1-10), *then they also satisfy* (1-9) *with constant* C *given by*

$$\sup_{\substack{A_i>0\\ A(j)>0}} \prod_{i=1}^{I} (\det A_i)^{-d_i/2} \prod_{j=1}^{J} (\det A(j))^{d(j)/2}.$$

The proof of Theorem 1.4 in [Liu et al. 2018] rests on an equivalent entropic formulation of the forward-reverse Brascamp–Lieb inequality[4] and ideas similar to those in [Geng and Nair 2014] and Lieb's original proof [1990] of the gaussian saturation property for the forward Brascamp–Lieb inequality. A different proof of Theorem 1.4 was found in [Courtade and Liu 2021], again utilizing entropic duality, but incorporating ideas from [Bennett et al. 2008; Lehec 2014]. We also remark that [Courtade and Liu 2021] embarked on a systematic study of the forward-reverse Brascamp–Lieb inequality and, for example, established a characterization of the finiteness of the constant in the spirit of Theorem 1.2 (see [loc. cit., Theorem 1.27]) and extended Barthe's duality relation (1-13) to all forward-reverse Brascamp–Lieb data (see [loc. cit., Theorem 1.3]).

Remark. Given the above discussion, it is clear that one may use Theorem 1.4 to deduce Theorem 1.3 in the case where there is no gaussian kernel. Surprisingly, the converse is true and this fact follows from a clever limiting argument due to Paweł Wolff, which is explained in [loc. cit., Section 4.1]. In the presence of a gaussian kernel, as far as we are aware, a version of such an equivalence has not appeared in the literature (see [loc. cit., Section 4] for some partial results along these lines).

In the following section we state our main result — an extension of Theorem 1.3 to certain regularized input functions — and some consequences. For example, we can recover all versions of the gaussian saturation property stated above (i.e., Theorems 1.1, 1.3 and 1.4). In fact, we present a framework[5] for the forward-reverse Brascamp–Lieb inequality which allows for gaussian kernels, and, in particular, we shall see that the gaussian saturation property in this framework is equivalent to that of Barthe–Wolff in Theorem 1.3. Despite this equivalence between forward-reverse and inverse Brascamp–Lieb inequalities, our heat-flow monotonicity approach seems best suited to the inverse Brascamp–Lieb inequality and so the inverse inequality should be regarded as the focal point of the paper. In fact, it is not at all clear to us whether one can expect a heat-flow monotonicity approach to succeed in the setting of the general forward-reverse Brascamp–Lieb inequality. Heat-flow arguments of a different nature have been successfully implemented in the case of the reverse Brascamp–Lieb inequality with geometric data — see [Barthe and Cordero-Erausquin 2004] (rank-1 transformations) and [Barthe and Huet 2009] (general rank transformations), as well as [Borell 1993; 2000; 2003] which seem to have been a source of inspiration for [Barthe and Cordero-Erausquin 2004; Barthe and Huet 2009]. For example, it is shown in [Barthe and Huet 2009, Theorem 4] that, for geometric data, the relation (1-12) is preserved if one replaces all functions with their evolution under classical heat flow $e^{t\Delta}$ *for all* $t > 0$; the reverse Brascamp–Lieb

[4]In the context of the forward Brascamp–Lieb inequality, duality with subadditivity of entropy can be found in [Carlen et al. 2004; Carlen and Cordero-Erausquin 2009].

[5]This framework was suggested to us by an anonymous referee to whom we are extremely grateful.

inequality for geometric data then follows by taking a limit $t \to \infty$. It would be very interesting to see if such an approach can be extended to general data in the forward-reverse framework, but it is certainly not clear to us whether such an approach can be profitable in proving, say, Theorem 2.2.

2. Main results

2.1. ***Regularized inverse Brascamp–Lieb inequalities.*** From now on, we denote the class of $n \times n$ real and self-adjoint transformations by $S(\mathbb{R}^n)$ and

$$S_+(\mathbb{R}^n) := \{A \in S(\mathbb{R}^n) : A > 0\}$$

for the subclass of positive definite transformations.

For $G \in S_+(\mathbb{R}^n)$ consider the solution $u : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}_+$ to the heat equation

$$\partial_t u = \frac{1}{4\pi} \operatorname{div}(G^{-1}\nabla u), \quad u(0) = \mu, \tag{2-1}$$

where μ is a positive finite Borel measure with nonzero mass. Then we call $f(x) = u(1, x)$ a *type G function*. Explicitly $u(1, x)$ can be written in convolution form

$$u(1, x) = g_G * \mu(x) = (\det G)^{1/2} \int_{\mathbb{R}^n} e^{-\pi \langle x-y, G(x-y)\rangle}\, d\mu(y).$$

We consider the inverse Brascamp–Lieb inequality restricted to regularized input functions of this type, and thus we introduce the notation

$$\mathcal{T}(G) := \{g_G * \mu : \mu \text{ is a positive finite Borel measure with nonzero mass}\}.$$

Note that the class $\mathcal{T}(G)$ is clearly subset of the class of all nonnegative and integrable functions, and its members are smooth and strictly positive. Also, it is not difficult to see that we have the nesting property,

$$G_1, G_2 \in S_+(\mathbb{R}^n), \quad G_1 \le G_2 \quad \Longrightarrow \quad \mathcal{T}(G_1) \subseteq \mathcal{T}(G_2). \tag{2-2}$$

It also formally makes sense[6] to view

$$\mathcal{T}(\infty) = \left\{ f : \mathbb{R}^n \to \mathbb{R}_+ : \int_{\mathbb{R}^n} f < \infty \right\} \tag{2-3}$$

as the class of all inputs.

For each Brascamp–Lieb datum $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$ and $\boldsymbol{G} = (G_j)_{j=1}^m$, with $G_j \in S_+(\mathbb{R}^{n_j})$, we refer to $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G})$ as a *generalized Brascamp–Lieb datum*[7] and define

$$\mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G}) := \inf_{\boldsymbol{f} \in \mathcal{T}(\boldsymbol{G})} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{f}) \tag{2-4}$$

to be the best (i.e., largest) constant such that (1-1) holds for input functions $f_j \in \mathcal{T}(G_j)$; accordingly, the notation $\boldsymbol{f} \in \mathcal{T}(\boldsymbol{G})$ means $f_j \in \mathcal{T}(G_j)$ for each $j = 1, \dots, m$.

[6]For example, we know that $g_{\lambda \mathrm{id}}$ converges to the Dirac delta supported at the origin as $\lambda \to \infty$.

[7]When $\mathcal{Q} = 0$, we shall simply say that $(\boldsymbol{B}, \boldsymbol{c}, \boldsymbol{G})$ is a *generalized Brascamp–Lieb datum*.

For generalized Brascamp–Lieb data $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G})$, we add a further condition to the nondegeneracy condition (1-7). We shall say that $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G})$ is nondegenerate if (1-7) holds and

$$\mathcal{Q} + \sum_{j=1}^{m_+} c_j B_j^* G_j B_j > 0. \tag{2-5}$$

This appears to be a reasonable condition in the following sense. If $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$ is nondegenerate (that is, (1-7) holds), then as discussed in the proof of [Barthe and Wolff 2022, Proposition 2.2], there exists $A_1, \dots, A_{m_+} > 0$ such that

$$\mathcal{Q} + \sum_{j=1}^{m_+} c_j B_j^* A_j B_j > 0$$

and hence $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G})$ is nondegenerate whenever $G_j \geq A_j$ for $j = 1, \dots, m_+$. From such considerations, our main result below recovers the Barthe–Wolff result in Theorem 1.3 as a limiting case; see the remark after Lemma 3.7 for further details. In addition, we note that $\inf_{\boldsymbol{A} \leq \boldsymbol{G}} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}) = \infty$ if (2-5) is not satisfied.

Theorem 2.1. *For any nondegenerate generalized Brascamp–Lieb datum* $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G})$, *we have*

$$\mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G}) = \inf_{\boldsymbol{A} \leq \boldsymbol{G}} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}). \tag{2-6}$$

As we have already mentioned, we use a heat-flow approach to prove Theorem 2.1. This may be viewed as a significant extension of the analysis in [Chen et al. 2015], which handled a special class of Brascamp–Lieb data (so-called geometric data — see the discussion at the beginning of Section 3).

2.2. ***Regularized forward-reverse Brascamp–Lieb inequalities.*** We shall see that Theorem 2.1 yields a gaussian saturation property for a regularized version of the forward-reverse Brascamp–Lieb inequality with gaussian kernels. To present this application, as far as possible, we use similar notation to that in Section 1.3.

We fix a collection of linear transformations $\boldsymbol{T} = (T_j)_{j=1}^J$. Here, $T_j : E \to E(j)$, $j = 1, \dots, J$, with $E(j) = \mathbb{R}^{n(j)}$ and where the base space is given by $E = \bigoplus_{i=0}^I E_i$ with $E_i = \mathbb{R}^{n_i}$, $i = 0, 1, \dots, I$. Next, we take two collections of positive real number $\boldsymbol{d} = ((d_i)_{i=1}^I, (d(j))_{j=1}^J)$. In addition, fix $Q_L \in S_+(E_0)$, Q_R to be a positive semidefinite transformation on E, and write $Q = (Q_L, Q_R)$. Finally, we fix two collections of positive definite transformations $\boldsymbol{G} = ((G_i)_{i=1}^I, (G(j))_{j=1}^J)$, where $G_i \in S_+(E_i)$ and $G(j) \in S_+(E(j))$.

The datum $(\boldsymbol{T}, \boldsymbol{d}, Q, \boldsymbol{G})$ is said to be nondegenerate if

$$E_0 \oplus \{0\} \oplus \cdots \oplus \{0\} \subseteq \ker Q_R, \quad \pi_0^* Q_L \pi_0 + \sum_{i=1}^I d_i \pi_i^* G_i \pi_i > Q_R \quad \text{on } E, \tag{2-7}$$

where, as before, π_i denotes the orthogonal projection from E to E_i. The second condition is reasonable since it is necessary for (2-9) below to hold with[8] $f_i = \gamma_{G_i}$, $h_j = \gamma_{G(j)}$, where $\gamma_A(x) := e^{-\pi \langle x, Ax \rangle}$. In fact,

[8] In this setting it is slightly more convenient to work with nonnormalized gaussians.

if the second condition in (2-7) fails to be true, then there is no $A_i \le G_i,\ A(j) \le G(j)$ satisfying (2-9) with $f_i = \gamma_{A_i},\ h_j = \gamma_{A(j)}$. Indeed, (2-9) with $f_i = \gamma_{A_i},\ h_j = \gamma_{A(j)}$ is equivalent to

$$\pi_0^* Q_L \pi_0 + \sum_{i=1}^{I} d_i \pi_i^* A_i \pi_i \ge Q_R + \sum_{j=1}^{J} d(j) T_j^* A(j) T_j \quad \text{on } E, \tag{2-8}$$

which in particular implies

$$\pi_0^* Q_L \pi_0 + \sum_{i=1}^{I} d_i \pi_i^* A_i \pi_i > Q_R.$$

Under such a nondegeneracy condition, we have the following generalization of Theorem 1.4.

Theorem 2.2. *For any nondegenerate data $\mathfrak{D} = (\boldsymbol{T}, \boldsymbol{d}, Q, \boldsymbol{G})$, we have the following. If $f_i \in \mathcal{T}(G_i)$ and $h_j \in \mathcal{T}(G(j))$ satisfy*

$$e^{-\pi\langle \pi_0 x, Q_L \pi_0 x\rangle} \prod_{i=1}^{I} f_i(\pi_i x)^{d_i} \le e^{-\pi\langle x, Q_R x\rangle} \prod_{j=1}^{J} h_j(T_j x)^{d(j)} \quad \textit{for all } x \in E, \tag{2-9}$$

then they also satisfy

$$\prod_{i=1}^{I} \Big(\int_{E_i} f_i \Big)^{d_i} \le \mathrm{FR}(\mathfrak{D}) \prod_{j=1}^{J} \Big(\int_{E(j)} h_j \Big)^{d(j)}, \tag{2-10}$$

where the constant is given by

$$\mathrm{FR}(\mathfrak{D}) = \sup\Bigg\{ \prod_{i=1}^{I} (\det A_i)^{-d_i/2} \prod_{j=1}^{J} (\det A(j))^{d(j)/2} : A_i \le G_i,\ A(j) \le G(j) \textit{ satisfy } (2\text{-}8) \Bigg\}.$$

We shall deduce the above theorem from Theorem 2.1 by means of Wolff's limiting argument alluded to at the end of Section 1. In fact, the theorems are equivalent and we shall prove this in the Appendix.

Capitalizing on the fact that the forward Brascamp–Lieb inequality is a special case of the forward-reverse Brascamp–Lieb inequality, we can quickly deduce the following generalized version of Lieb's theorem.

Theorem 2.3. *For any generalized Brascamp–Lieb data $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G})$ with $c_j > 0$ for each j, and positive semidefinite $\mathcal{Q}$, we have*

$$\sup_{\boldsymbol{f} \in \mathcal{T}(\boldsymbol{G})} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{f}) = \sup_{\boldsymbol{A} \le \boldsymbol{G}} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}).$$

A limiting argument shows that Theorem 2.3 implies Theorem 1.1, so in this sense, we see that Lieb's result is also a consequence of Theorem 2.1. Also, the case $\mathcal{Q} = 0$ was proved in [Bennett et al. 2008, Corollary 8.11] and we shall follow the approach taken in [loc. cit.] in proving Theorem 2.1.

In a similar manner, one may also quickly obtain a regularized version of Barthe's reverse Brascamp–Lieb inequality (1-11) from Theorem 2.2 (by taking $J = 1$ and $d(1) = 1$).

Remark. Valdimarsson [2007] obtained a regularized version of (1-11) in which the input functions took the form $f_j(x) = \exp(-\pi\langle x, G_j^{-1} x\rangle - H_j(x))$ with H_j a convex function (so-called "inverse class G_j"). This particular set-up appears to be independent from ours in Theorem 2.2. Note that Valdimarsson's

setting of inverse class G_j input functions allowed him to extend the ideas of [Barthe 1998b] and obtain a generalization of the duality relation (1-13) involving $\mathrm{F}(\boldsymbol{B}, \boldsymbol{c}, \boldsymbol{G})$. It is unclear to us whether a duality result is possible for the regularization we consider in Theorem 2.2 and it seems it may be natural to adapt the framework somehow to include inverse class G functions in some appropriate way.

In addition, a certain regularized version of the forward-reverse Brascamp–Lieb inequality (in its dual entropic representation) was considered in [Liu et al. 2018, Section 4]. Again, we believe that this has no direct connection with the kind of regularization that we study in the present paper.

Organization. The remainder of the paper is primarily devoted to proving Theorem 2.1. Section 3 contains several preliminary observations, mostly related to the heat-flow monotonicity approach that we will use to prove Theorem 2.1. Although the main body of the proof of Theorem 2.1 is given in Section 5, the key heat-flow result needed for the proof has been isolated in Theorem 4.1 in Section 4; we take the opportunity to present this result in the form of a "closure property" for sub/supersolutions to certain heat equations and thus contribute to the emerging theory of such closure properties in, for example, [Aoki et al. 2020; Bennett and Bez 2009; 2019]. After the proof of Theorem 2.1 in Section 5, in Section 6 we present some further applications and remarks. Finally, in the Appendix, we prove the equivalence between Theorems 2.1 and 2.2.

3. Preliminaries

It will be convenient to introduce the notation

$$\mathrm{I}_{\mathrm{g}}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G}) = \inf_{\boldsymbol{A} \le \boldsymbol{G}} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A})$$

for the best constant such that (1-1) holds for gaussian input functions $f_j = g_{A_j}$ with $A_j \le G_j$ for each j. When $\boldsymbol{G} = (\infty, \dots, \infty)$ or $\mathcal{Q} = 0$ we simply omit it from the above notation.

3.1. ***Geometric data and a key linear algebraic lemma.*** Before embarking on the full proof of Theorem 2.1, as a highly instructive preliminary first step, let us consider the so-called *geometric* case and, for simplicity, we set $\mathcal{Q} = 0$ and $\boldsymbol{G} = (\infty, \dots, \infty)$. The geometric condition on the data is

$$B_j B_j^* = \mathrm{id} \quad (j = 1, \dots, m) \quad \text{and} \quad \sum_{j=1}^{m} c_j B_j^* B_j = \mathrm{id}. \tag{3-1}$$

In a such a case, it is clear that $\ker \boldsymbol{B}_+ = \{0\}$ and it follows that the nondegeneracy condition (1-7) is equivalent to the surjectivity of $\boldsymbol{B}_+$.

Theorem 3.1. *Suppose the Brascamp–Lieb datum $(\boldsymbol{B}, \boldsymbol{c})$ satisfies* (3-1) *and is nondegenerate in the sense that $\boldsymbol{B}_+$ is surjective. Then*

$$\mathrm{I}(\boldsymbol{B}, \boldsymbol{c}) = \mathrm{I}_{\mathrm{g}}(\boldsymbol{B}, \boldsymbol{c}) = \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}; \boldsymbol{A}) = 1,$$

where A_j is the identity transformation on $\mathbb{R}^{n_j}$ for $j = 1, \dots, m$.

Theorem 3.1 was proved in [Barthe and Wolff 2022, Theorem 4.7] based on a mass transportation argument. A very closely related result was established using heat flow in [Chen et al. 2015, Theorem 2];

in particular, it was shown that

$$\int_{\mathbb{R}^n} \prod_{j=1}^m f_j(B_j x)^{c_j}\,dx \geq \prod_{j=1}^m \Big(\int_{\mathbb{R}^{n_j}} f_j\Big)^{c_j} \tag{3-2}$$

holds under the assumption that $B_j B_j^* = \mathrm{id}$ for each j, $n = \sum_{j=1}^m c_j n_j$ and

$$\boldsymbol{B}\boldsymbol{B}^* \geq C^{-1} \tag{3-3}$$

hold, where $\boldsymbol{B} = (B_1, \ldots, B_m) : \mathbb{R}^n \to \prod_{j=1}^m \mathbb{R}^{n_j}$, and

$$C := \mathrm{diag}(c_1 \mathrm{id}_{\mathbb{R}^{n_1}}, \ldots, c_m \mathrm{id}_{\mathbb{R}^{n_m}}).$$

We refer the reader to [Barthe and Wolff 2022, Section 4] for a precise clarification of how one may obtain [Chen et al. 2015, Theorem 2] from their work.

As observed in [loc. cit.], assumption (3-3) is clearly equivalent to

$$\Big|\sum_{j=1}^m c_j B_j^* \boldsymbol{v}_j\Big|^2 \geq \sum_{j=1}^m c_j |\boldsymbol{v}_j|^2 \quad \text{for all } \boldsymbol{v}_j \in \mathbb{R}^{n_j}, \tag{3-4}$$

and this fact was key to their heat-flow proof of (3-2). Here we exhibit a sketch proof of Theorem 3.1, following the heat flow argument used in [loc. cit.] (which itself is based on similar heat flow proofs of the forward Brascamp–Lieb inequality in [Bennett et al. 2008; Carlen et al. 2004; Valdimarsson 2008]), which will help to understand our argument in the case of more general nondegenerate Brascamp–Lieb data $(\boldsymbol{B}, \boldsymbol{c}, \boldsymbol{G}, \mathcal{Q})$ in Section 5. The geometric case is particularly well suited to a heat-flow argument since gaussian maximizers always exist in such a case (in fact, as stated above, isotropic gaussians are maximizers). Including some details in the specific case of geometric data will also motivate the crucial linear algebraic result (a generalization of (3-4) in the forthcoming Lemma 3.2) which underpins the heat-flow argument in our proof of Theorem 2.1.

A heat-flow approach to Theorem 3.1. From (3-1), it is easy to check that

$$\mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}; \boldsymbol{A}) = 1$$

if each A_j is the identity transformation, so it suffices to prove (3-2) holds for sufficiently nice nonnegative f_j. (We remark that identifying a sufficiently nice class of test functions will be taken up in Section 3.4.) To this end, we define the quantity

$$\mathfrak{U}(t) := \int_{\mathbb{R}^n} U(t, x)\,dx, \tag{3-5}$$

where $u_j : (0, \infty) \times \mathbb{R}^{n_j} \to (0, \infty)$ solves the heat equation

$$\partial_t u_j = \Delta u_j, \quad u_j(0) = f_j, \tag{3-6}$$

and $U := \prod_{j=1}^m (u_j \circ B_j)^{c_j}$. Formally, we have

$$\mathfrak{U}(0) = \int_{\mathbb{R}^n} \prod_{j=1}^m f_j(B_j x)^{c_j}\,dx, \quad \lim_{t\to\infty} \mathfrak{U}(t) = \prod_{j=1}^m \Big(\int_{\mathbb{R}^{n_j}} f_j\Big)^{c_j},$$

where the argument for the limit at infinity made use of (3-1). Thus our aim is to show that $\mathfrak{Q}$ is nonincreasing in time. To this end, we note that

$$\partial_t U(t,x) = U(t,x)\sum_{j=1}^m c_j \partial_t \log u_j(t, B_j x) = U(t,x)\sum_{j=1}^m c_j[|\boldsymbol{v}_j|^2 + \operatorname{div}(\boldsymbol{v}_j)](t, B_j x),$$

where $\boldsymbol{v}_j := \nabla \log u_j$. By the divergence theorem, it follows (at least formally) that

$$\mathfrak{Q}'(t) = \int_{\mathbb{R}^n} U(t,x)\bigg[\sum_{j=1}^m c_j |\boldsymbol{v}_j(t, B_j x)|^2 - \bigg|\sum_{j=1}^m c_j B_j^* \boldsymbol{v}_j(t, B_j x)\bigg|^2\bigg]\,dx.$$

Thus, the argument above has reduced the proof of Theorem 3.1 to showing (3-4). This fact is a special case of Lemma 3.2 below (with $A_j = \mathrm{id}$ for each $j = 1, \dots, m$); the general case in Lemma 3.2 will be crucial for the argument in Section 5 to prove Theorem 2.1. Before stating the lemma, we remark that in the case of the forward Brascamp–Lieb inequality with $c_j > 0$ for each j, for geometric data $(\boldsymbol{B}, \boldsymbol{c})$ we have that (3-4) holds *in reverse*; to see this, if we define

$$\bar{\boldsymbol{w}} := \sum_{j=1}^m c_j B_j^* \boldsymbol{v}_j,$$

then by the Cauchy–Schwarz inequality and the geometric condition $\sum_{j=1}^m c_j B_j^* B_j = \mathrm{id}$, we get

$$|\bar{\boldsymbol{w}}|^2 = \sum_{j=1}^m \langle \sqrt{c_j} B_j \bar{\boldsymbol{w}}, \sqrt{c_j}\boldsymbol{v}_j\rangle \le \bigg(\sum_{j=1}^m c_j |B_j \bar{\boldsymbol{w}}|^2\bigg)^{1/2}\bigg(\sum_{j=1}^m c_j |\boldsymbol{v}_j|^2\bigg)^{1/2} = |\bar{\boldsymbol{w}}|\bigg(\sum_{j=1}^m c_j |\boldsymbol{v}_j|^2\bigg)^{1/2},$$

which rearranges to give (3-4) in reverse. Interestingly, it was shown in [Barthe and Wolff 2022] that one can obtain (3-4) from $\sum_{j=1}^m c_j |\boldsymbol{v}_j|^2 \le \big|\sum_{j=1}^m c_j B_j^* \boldsymbol{v}_j\big|^2$ $(c_j > 0)$; more generally, a proof along such lines was used to derive the more general inequality in (3-8) below.

Lemma 3.2. *Suppose the Brascamp–Lieb datum $(\boldsymbol{B}, \boldsymbol{c})$ is such that $\boldsymbol{B}_+$ is surjective, and $A_j \in S_+(\mathbb{R}^{n_j})$, $j = 1, \dots, m$, are such that the transformation $M := \sum_{j=1}^m c_j B_j^* A_j B_j$ is positive definite. Let $\boldsymbol{v}_j \in \mathbb{R}^{n_j}$ for $j = 1, \dots, m$, and let $x_* \in \mathbb{R}^n$ be any nonzero element of $\boldsymbol{B}_+^{-1}(A_1^{-1}\boldsymbol{v}_1, \dots, A_{m_+}^{-1}\boldsymbol{v}_{m_+})$. Then we have the identity*

$$\langle \bar{\boldsymbol{w}}, M^{-1}\bar{\boldsymbol{w}}\rangle - \sum_{j=1}^m c_j \langle \boldsymbol{v}_j, A_j^{-1}\boldsymbol{v}_j\rangle = \langle \bar{\boldsymbol{w}}', M\bar{\boldsymbol{w}}'\rangle + \sum_{j=m_++1}^m |c_j|\langle \boldsymbol{v}_j', A_j \boldsymbol{v}_j'\rangle, \tag{3-7}$$

where $\bar{\boldsymbol{w}} := \sum_{j=1}^m c_j B_j^* \boldsymbol{v}_j$, $\bar{\boldsymbol{w}}' := x_* - M^{-1}\bar{\boldsymbol{w}}$, and $\boldsymbol{v}_j' := B_j x_* - A_j^{-1}\boldsymbol{v}_j$ for $j = m_+ + 1, \dots, m$. In particular, we have

$$\sum_{j=1}^m c_j \langle \boldsymbol{v}_j, A_j^{-1}\boldsymbol{v}_j\rangle \le \langle \bar{\boldsymbol{w}}, M^{-1}\bar{\boldsymbol{w}}\rangle \tag{3-8}$$

for any $\boldsymbol{v}_j \in \mathbb{R}^{n_j}$, $j = 1, \dots, m$.

Inequality (3-8) can be obtained directly from [Barthe and Wolff 2022, Lemma 3.5]. Since the above result is pivotal to the proof of Theorem 2.1, we include our own brief justification.

Proof of Lemma 3.2. First note that

$$B_j x_* = A_j^{-1} \boldsymbol{v}_j \quad (1 \le j \le m_+) \tag{3-9}$$

and with this in mind we will expand the right-hand side of (3-7). For $j = m_+ + 1, \dots, m$, we have

$$\begin{aligned}\langle \boldsymbol{v}_j', A_j \boldsymbol{v}_j' \rangle &= \langle B_j x_*, A_j B_j x_* \rangle - 2\langle B_j x_*, \boldsymbol{v}_j \rangle + \langle A_j^{-1} \boldsymbol{v}_j, \boldsymbol{v}_j \rangle,\\ \langle \bar{\boldsymbol{w}}', M \bar{\boldsymbol{w}}' \rangle &= \langle x_*, M x_* \rangle - 2\langle x_*, \bar{\boldsymbol{w}} \rangle + \langle M^{-1} \bar{\boldsymbol{w}}, \bar{\boldsymbol{w}} \rangle.\end{aligned}$$

So, if we set $M_+ := \sum_{j=1}^{m_+} c_j B_j^* A_j B_j$, the right-hand side of (3-7) can be written $I_a + I_b + I_c$ where

$$I_a = \langle x_*, M_+ x_* \rangle, \quad I_b = -2\langle x_*, \bar{\boldsymbol{w}} \rangle - 2 \sum_{j=m_++1}^{m} |c_j| \langle B_j x_*, \boldsymbol{v}_j \rangle, \quad I_c = \langle M^{-1} \bar{\boldsymbol{w}}, \bar{\boldsymbol{w}} \rangle + \sum_{j=m_++1}^{m} |c_j| \langle A_j^{-1} \boldsymbol{v}_j, \boldsymbol{v}_j \rangle.$$

From (3-9) we have

$$\begin{aligned} I_a &= \Big\langle x_*, \sum_{j=1}^{m_+} c_j B_j^* \boldsymbol{v}_j \Big\rangle = \sum_{j=1}^{m_+} c_j \langle B_j x_*, \boldsymbol{v}_j \rangle = \sum_{j=1}^{m_+} c_j \langle A_j^{-1} \boldsymbol{v}_j, \boldsymbol{v}_j \rangle,\\ I_b &= -2\Big\langle x_*, \sum_{j=1}^{m_+} c_j B_j^* \boldsymbol{v}_j \Big\rangle = -2 \sum_{j=1}^{m_+} c_j \langle B_j x_*, \boldsymbol{v}_j \rangle = -2 \sum_{j=1}^{m_+} c_j \langle A_j^{-1} \boldsymbol{v}_j, \boldsymbol{v}_j \rangle. \end{aligned}$$

So, putting things together, $I_a + I_b + I_c$ clearly coincides with the left-hand side of (3-7). □

In order to make rigorous the formal considerations in our above sketch proof of Theorem 3.1, it is important to identify an appropriately nice class of functions f_j on which we may reduce matters. Similar considerations will also be necessary for our proof of Theorem 2.1 in Section 5 and the content of Section 3.4 below has been prepared primarily for this purpose.

3.2. ***A log-convexity estimate.*** Taking trace on both sides of the geometric condition $\sum_{j=1}^m c_j B_j^* B_j = \mathrm{id}$, we see that $\sum_{j=1}^m c_j n_j = n$ follows. In fact, in the case $\boldsymbol{G} = (\infty, \dots, \infty)$ and $\mathcal{Q} = 0$, the condition $\sum_{j=1}^m c_j n_j = n$ is easily seen to be necessary for the strict positivity of the inverse Brascamp–Lieb constant by standard scaling considerations. For general $\boldsymbol{G}$, such a scaling condition is no longer a requirement and it is natural to modify the quantity $\mathfrak{Q}$ considered in (3-5) and mitigate for the loss of scaling by considering quantities of the form

$$\mathfrak{Q}(t) = t^{-\alpha} \int_{\mathbb{R}^n} \prod_{j=1}^m u_j(t, B_j x)^{c_j}\, dx,$$

where $\alpha := \frac{1}{2}\big(n - \sum_{j=1}^m c_j n_j\big)$. In such a case, the heat-flow monotonicity argument will make important use of the following log-convexity estimate.

Lemma 3.3 (log-convexity). *Let $G \in S_+(\mathbb{R}^n)$ and $u : \mathbb{R}^n \to (0, \infty)$ be of type G. Then*

$$D^2(\log u) \ge -2\pi G.$$

Obviously $D^2(\log g_G) = -2\pi G$ and thus Lemma 3.3 is equivalent to the log-convexity of the ratio u/g_G. For a proof of Lemma 3.3, we refer the reader to [Bennett et al. 2008, Lemma 8.6]. We also remark that

estimates of the type in Lemma 3.3 are also known as *Li–Yau gradient estimates* and hold much more generally in the framework of Riemannian manifolds; see [Li and Yau 1986].

3.3. ***The key decomposition of*** $\mathcal{Q}$**.** In order to handle the gaussian kernel, our proof of Theorem 2.1 is underpinned by the following decomposition of the transformation $\mathcal{Q}$ into positive and negative parts.

Proposition 3.4 [Barthe and Wolff 2022]. *The transformation* $\mathcal{Q} \in S(\mathbb{R}^n)$ *satisfies* (1-7) *if and only if there exist linear surjections* $B_0 : \mathbb{R}^n \to \mathbb{R}^{n_0}$ *and* $B_{m+1} : \mathbb{R}^n \to \mathbb{R}^{n_{m+1}}$, *and* $\mathcal{Q}_+ \in S_+(\mathbb{R}^{n_0})$ *and* $\mathcal{Q}_- \in S_+(\mathbb{R}^{n_{m+1}})$ *such that*

$$\mathcal{Q} = B_0^* \mathcal{Q}_+ B_0 - B_{m+1}^* \mathcal{Q}_- B_{m+1}, \tag{3-10}$$

the transformation $(B_0, \boldsymbol{B}_+) : \mathbb{R}^n \to \bigoplus_{j=0}^{m_+} \mathbb{R}^{n_j}$ *is bijective*, *and* $\ker \boldsymbol{B}_+ \subseteq \ker B_{m+1}$.

We refer the reader to [Barthe and Wolff 2022, Lemma 3.1] for a proof of the above decomposition. We note here that, assuming the nondegeneracy condition (1-7), Barthe and Wolff take B_0 to be the projection onto $\ker \boldsymbol{B}_+$ and we understand that in the case $\ker \boldsymbol{B}_+ = \{0\}$ we have $n_0 = 0$ and (3-10) becomes $\mathcal{Q} = -B_{m+1}^* \mathcal{Q}_- B_{m+1}$ (similarly when $n_{m+1} = 0$ we understand that $\mathcal{Q} = B_0^* \mathcal{Q}_+ B_0$). As we shall see in the forthcoming proof of Theorem 2.1, the decisive scenario is when $\boldsymbol{B}_+$ is bijective.

3.4. ***Nice classes of input functions.*** The main purpose of this remaining part of Section 3 is to identify appropriate classes of test functions which approximate general inputs and are sufficiently well behaved in order to facilitate, as far as possible, a proof of Theorem 2.1 which is free from burdensome technicalities.

Definition 3.5. The class $\mathcal{N}$ is defined to be those inputs $\boldsymbol{f} = (f_j)_{j=1}^m$ satisfying the conditions

$$f_j(x_j) \le C_{\boldsymbol{f}} \mathbf{1}_{|x_j| \le C_{\boldsymbol{f}}} \qquad (1 \le j \le m_+), \tag{3-11}$$

$$f_j(x_j) \ge C_{\boldsymbol{f}}^{-1} (1 + |x_j|^2)^{-n_j} \quad (m_+ + 1 \le j \le m) \tag{3-12}$$

for some constant $C_{\boldsymbol{f}} \in (0, \infty)$.

Lemma 3.6. *For any nondegenerate Brascamp–Lieb datum* $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$, *we have*

$$\mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}) = \inf_{\boldsymbol{f} \in \mathcal{N}} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{f}).$$

Proof. For $N \ge 1$ and an arbitrary input $\boldsymbol{f} = (f_j)_{j=1}^m$, define the input $\boldsymbol{f}^{(N)}$ by

$$f_j^{(N)}(x_j) := \begin{cases} \mathbf{1}_{|x_j| \le N} \mathbf{1}_{f_j \le N}(x_j) f_j(x_j), & 1 \le j \le m_+, \\ N^{-1}(1 + |x_j|^2)^{-n_j} + f_j(x_j), & m_+ + 1 \le j \le m. \end{cases}$$

Clearly $\boldsymbol{f}^{(N)} \in \mathcal{N}$. Also, it is obvious that $f_j^{(N)} \uparrow f_j$ for $1 \le j \le m_+$ and $f_j^{(N)} \downarrow f_j$ for $m_+ + 1 \le j \le m$ as $N \to \infty$. For $1 \le j \le m_+$, an application of the monotone convergence theorem shows $\int f_j^{(N)} \to \int f_j$. For $m_+ + 1 \le j \le m$, since $(1 + |x_j|^2)^{-n_j} \in L^1(\mathbb{R}^{n_j})$, we have

$$\lim_{N \to \infty} \int_{\mathbb{R}^{n_j}} f_j^{(N)} = \lim_{N \to \infty} \frac{1}{N} \int_{\mathbb{R}^{n_j}} (1 + |x_j|^2)^{-n_j} \, dx_j + \int_{\mathbb{R}^{n_j}} f_j = \int_{\mathbb{R}^{n_j}} f_j.$$

Also, since $(f_j^{(N)})^{c_j} \uparrow f_j^{c_j}$ for each $j = 1, \dots, m$, another application of the monotone convergence theorem gives

$$\lim_{N\to\infty} \int_{\mathbb{R}^n} e^{-\pi\langle x, \mathcal{Q}x\rangle} \prod_{j=1}^m f_j^{(N)}(B_j x)^{c_j}\, dx = \int_{\mathbb{R}^n} e^{-\pi\langle x, \mathcal{Q}x\rangle} \prod_{j=1}^m f_j(B_j x)^{c_j}\, dx.$$

The above shows

$$\mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{f}) = \lim_{N\to\infty} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{f}^{(N)}) \geq \inf_{\tilde{\boldsymbol{f}}\in\mathcal{N}} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \tilde{\boldsymbol{f}})$$

and thus $\mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}) \geq \inf_{\boldsymbol{f}\in\mathcal{N}} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{f})$. The reverse inequality is trivial and thus we conclude the proof of the lemma. □

Thanks to the previous approximation lemma, we can deduce the following result.

Lemma 3.7. *If $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G})$ is a nondegenerate generalized Brascamp–Lieb datum, then*

$$\lim_{\lambda\to\infty} \mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \lambda\boldsymbol{G}) = \mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}), \quad \lim_{\lambda\to\infty} \mathrm{I}_{\mathrm{g}}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \lambda\boldsymbol{G}) = \mathrm{I}_{\mathrm{g}}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}).$$

Remark. Lemma 3.7 allows us to show that Theorem 2.1 recovers Theorem 1.3 (and justifies the notation (2-3)). To deduce Theorem 1.3, we take $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$ such that (1-7) holds and use [Barthe and Wolff 2022, Proposition 2.2] to obtain the existence of $\lambda_0 > 0$ such that

$$\mathcal{Q} + \lambda_0 \sum_{j=1}^{m_+} c_j B_j^* B_j > 0.$$

Thus, setting $G_j := \lambda_0\,\mathrm{id}$ for each $j = 1, \dots, m$, it follows that $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \lambda\boldsymbol{G})$ is nondegenerate for all $\lambda \geq 1$. By using Lemma 3.7 and Theorem 2.1 we obtain

$$\mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}) = \lim_{\lambda\to\infty} \mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \lambda\boldsymbol{G}) = \lim_{\lambda\to\infty} \mathrm{I}_{\mathrm{g}}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \lambda\boldsymbol{G}) = \mathrm{I}_{\mathrm{g}}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$$

and we recover Theorem 1.3.

Proof of Lemma 3.7. To see $\lim_{\lambda\to\infty} \mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \lambda\boldsymbol{G}) = \mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$, it clearly suffices to show

$$\lim_{\lambda\to\infty} \mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \lambda\boldsymbol{G}) \leq \mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}). \tag{3-13}$$

To see this, we argue that

$$\lim_{\lambda\to\infty} \mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \lambda\boldsymbol{G}) \leq \lim_{\lambda\to\infty} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; g_{\lambda\boldsymbol{G}} * \boldsymbol{f}) = \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{f})$$

for arbitrary $\boldsymbol{f} \in \mathcal{N}$ and use Lemma 3.6; thus it suffices to show

$$\lim_{\lambda\to\infty} \int_{\mathbb{R}^n} e^{-\pi\langle x, \mathcal{Q}x\rangle} \prod_{j=1}^m g_{\lambda G_j} * f_j(B_j x)^{c_j}\, dx = \int_{\mathbb{R}^n} e^{-\pi\langle x, \mathcal{Q}x\rangle} \prod_{j=1}^m f_j(B_j x)^{c_j}\, dx. \tag{3-14}$$

To see this, we use the fact that $f \in \mathcal{N}$ to get the bounds

$$\begin{aligned} g_{\lambda G_j} * f_j(x_j) &\leq C e^{-\pi\langle x_j, (\lambda/4) G_j x_j\rangle} \quad (1 \leq j \leq m_+), \\ g_{\lambda G_j} * f_j(x_j) &\geq C^{-1}(1 + |x_j|^2)^{-n_j} \quad (m_+ + 1 \leq j \leq m) \end{aligned}$$

for all $x_j \in \mathbb{R}^{n_j}$, and where $C \in (0, \infty)$ is a constant depending on G_j and $C_{\boldsymbol{f}}$. Then

$$e^{-\pi\langle x, \mathcal{Q}x\rangle} \prod_{j=1}^{m} g_{\lambda G_j} * f_j(B_j x)^{c_j} \le C e^{-\pi\langle x, (\mathcal{Q}+(\lambda/4)M_+)x\rangle}(1+|x|^2)^N$$

for all $x \in \mathbb{R}^n$, where $M_+ := \sum_{j=1}^{m_+} c_j B_j^* G_j B_j$, $N := \sum_{j=m_++1}^{m} |c_j| n_j$, and C is a finite constant depending on $\boldsymbol{B}, \boldsymbol{G}$ and $\boldsymbol{f}$, but independent of λ. Thanks to the nondegeneracy condition (2-5), it follows that $\mathcal{Q} + (\lambda/4)M_+$ is positive definite for $\lambda \ge 4$. Hence we may apply the dominated convergence theorem to deduce (3-14).

Similarly, we can prove $\lim_{\lambda\to\infty} \mathrm{I_g}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \lambda\boldsymbol{G}) = \mathrm{I_g}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$. □

Now we introduce $\boldsymbol{G} = (G_j)_{j=1}^m$, where $G_j \in S_+(\mathbb{R}^{n_j})$, and assume the generalized Brascamp–Lieb datum $(\boldsymbol{B}, \boldsymbol{c}, \boldsymbol{G}, \mathcal{Q})$ is nondegenerate. The following class of functions will play a prominent role in the heat flow proof of Theorem 2.1.

Definition 3.8. The class $\mathcal{N}(\boldsymbol{G})$ is defined to be those inputs $\boldsymbol{f} = (f_j)_{j=1}^m \in \mathcal{T}(\boldsymbol{G})$ of the form $f_j = g_{G_j} * h_j$ with $h_j \in \mathcal{N}$, $j = 1, \dots, m$.

Regarding this function class, we first note that the following analogue of Lemma 3.6 holds.

Lemma 3.9. *We have*

$$\mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G}) = \inf_{\boldsymbol{f}\in\mathcal{N}(\boldsymbol{G})} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{f}).$$

One can prove Lemma 3.9 in a similar manner to the proof of Lemma 3.6 and so we omit the details.

Although we considered isotropic heat flow (3-6) for geometric data, for more general data it will be appropriate to consider certain anisotropic heat flows. Associated with such flows, we have the following pointwise estimates whose role will be to make rigorous the forthcoming heat-flow proof of Theorem 2.1 for data such that gaussian maximizers exist.

Lemma 3.10. *Let $(\boldsymbol{B}, \boldsymbol{c})$ be a Brascamp–Lieb datum. Suppose $\boldsymbol{f} \in \mathcal{N}(\boldsymbol{G})$ and $\boldsymbol{A} \le \boldsymbol{G}$. For each $j = 1, \dots, m$, suppose further that $u_j : (1, \infty) \times \mathbb{R}^{n_j} \to (0, \infty)$ is a solution to the heat equation*

$$\partial_t u_j = \frac{1}{4\pi} \operatorname{div}(A_j^{-1} \nabla u_j), \quad u_j(1, x_j) = f_j(x_j). \tag{3-15}$$

Fix $t > 1$ and $\varepsilon > 0$. Then

$$|\nabla \log u_j(t, B_j x)| \le C_*(t)(1+|x|^2)^{n_j} \qquad (j = 1, \dots, m), \tag{3-16}$$

$$|\partial_t \log u_j(t, B_j x)| \le C_*(t)(1+|x|^2)^{\max\{2, n_j\}} \quad (j = 1, \dots, m), \tag{3-17}$$

and

$$\prod_{j=1}^{m} u_j(t, B_j x)^{c_j} \le C_*(t, \varepsilon)(1+|x|^2)^N e^{-(1-\varepsilon)\pi\langle x, t^{-1}M_+ x\rangle} \tag{3-18}$$

for all $x \in \mathbb{R}^n$. Here, $M_+ := \sum_{j=1}^{m_+} c_j B_j^* A_j B_j$ and $N := \sum_{j=m_++1}^{m} |c_j| n_j$. Also, $C_*(t)$ denotes a strictly positive and finite constant which depends on t, $\boldsymbol{B}$, $\boldsymbol{c}$, $\boldsymbol{A}$, $\boldsymbol{G}$, $\boldsymbol{f}$ and is locally uniformly bounded in t, and $C_*(t, \varepsilon)$ denotes such a constant which also depends on ε.

Proof. Since $\boldsymbol{f} \in \mathcal{N}(\boldsymbol{G})$, we may write $f_j = g_{G_j} * h_j$, where $h_j \in \mathcal{N}$. For $j \in \{1, \dots, m_+\}$ we let $R_j > 0$ be such that the support of h_j is contained in the ball of radius R_j with centre at the origin.

It is easily verified (using, say, the Fourier transform) that u_j can be explicitly written down as

$$u_j(t, x_j) = g_{P_j(t)} * h_j(x_j), \tag{3-19}$$

where $P_j(t) := (G_j^{-1} + (t-1)A_j^{-1})^{-1}$.

First we prove (3-18). If $j \in \{1, \dots, m_+\}$, then one can easily check that

$$\begin{aligned} u_j(t, x_j) &= (\det P_j(t))^{1/2} \int_{\mathbb{R}^{n_j}} e^{-\pi \langle x_j - y_j, P_j(t)(x_j - y_j)\rangle} h_j(y_j)\, dy_j \\ &\le (\det P_j(t))^{1/2} \bigg(\int_{\mathbb{R}^{n_j}} h_j \bigg) e^{-(1-\varepsilon)\pi \langle x_j, P_j(t) x_j\rangle} \end{aligned}$$

for $|x_j| \ge \frac{1}{\varepsilon\pi} \|P(t)^{1/2}\| \|P(t)^{-1/2}\| R_j$. It follows that

$$u_j(t, x_j) \le C_*(t, \varepsilon) e^{-(1-\varepsilon)\pi \langle x_j, P_j(t) x_j\rangle}$$

for all $x_j \in \mathbb{R}^{n_j}$. By assumption we have $A_j \le G_j$ and as a consequence $P_j(t) \ge A_j/t$. Hence

$$u_j(t, x_j) \le C_*(t, \varepsilon) e^{-(1-\varepsilon)\pi \langle x_j, t^{-1} A_j x_j\rangle}$$

for all $x_j \in \mathbb{R}^{n_j}$.

For $j \in \{m_+ + 1, \dots, m\}$, we have

$$u_j(t, x_j) \ge \int_{\mathbb{R}^{n_j}} (1 + |x_j - P_j(t)^{-1/2} y_j|^2)^{-n_j} e^{-\pi |y_j|^2}\, dy_j \ge C_*(t)^{-1} (1 + |x_j|^2)^{-n_j}$$

for all $x_j \in \mathbb{R}^{n_j}$, where the second lower bound follows by restricting the domain of integration to $|y_j| \le \|P_j(t)^{-1/2}\|^{-1}$. From the above, we clearly obtain (3-18).

Next we check (3-16). For $j \in \{1, \dots, m_+\}$, we use the compactness of the support of h_j to obtain

$$\begin{aligned} |\nabla u_j(t, x_j)| &\le 2\pi \|P_j(t)\| (\det P_j(t))^{1/2} \int_{\mathbb{R}^{n_j}} |x_j - y_j| e^{-\pi \langle x_j - y_j, P_j(t)(x_j - y_j)\rangle} h_j(y_j)\, dy_j \\ &\le C_*(t)(1 + |x_j|) u_j(t, x_j) \end{aligned}$$

for all $x_j \in \mathbb{R}^{n_j}$, and this clearly suffices for (3-16) for such j. For $j \in \{m_+ + 1, \dots, m\}$, we use the fact that $P_j(t) > 0$ to obtain $|w_j| e^{-\pi \langle w_j, P_j(t) w_j\rangle} \le C_*(t)$ for all $w_j \in \mathbb{R}^{n_j}$, and therefore $|\nabla u_j(t, x_j)| \le C_*(t)$ for all $x_j \in \mathbb{R}^{n_j}$. It follows that

$$|\nabla \log u_j(t, x_j)| \le C_*(t)(1 + |x_j|^2)^{n_j}$$

for all $x_j \in \mathbb{R}^{n_j}$.

Finally we note that one can essentially follow the above argument for (3-16) in order to show (3-17) and so we omit the details. □

We shall also need the large time asymptotics of the solution to (3-15). From the explicit form of the solution (3-19) we easily see that

$$\lim_{t \to \infty} t^{n_j/2} u_j(t, \sqrt{t} x_j) = \bigg(\int_{\mathbb{R}^{n_j}} f_j \bigg) g_{A_j}(x_j) \tag{3-20}$$

for each $x_j \in \mathbb{R}^{n_j}$.

4. Closure properties of sub/supersolutions of heat equations

The heat-flow proof of Theorem 3.1 in the previous section rested on the nonincreasingness of

$$\mathfrak{O}(t) = \int_{\mathbb{R}^n} U(t,x)\,dx,$$

where U is the "anisotropic geometric mean" given by

$$U(t,x) = \prod_{j=1}^{m} (u_j \circ B_j)^{c_j}$$

and u_j is a solution to the heat equation $\partial_t u_j = \Delta u_j$ with nonnegative initial data. Although it is not immediate from the argument we presented in the previous section, one can show that U satisfies $\partial_t U \le \Delta U$; that is, U is a *subsolution* of the corresponding heat equation. Formally at least, the monotonicity of $\mathfrak{O}$ can be seen in this way since

$$\mathfrak{O}'(t) = \int_{\mathbb{R}^n} \partial_t U(t,x)\,dx \le \int_{\mathbb{R}^n} \Delta U(t,x)\,dx = 0. \tag{4-1}$$

In fact, in the case of geometric Brascamp–Lieb data, it is more generally true that if u_j are nonnegative *subsolutions* for $j = 1, \dots, m_+$ and nonnegative *supersolutions* for $j = m_+ + 1, \dots, m$, then $\partial_t U \le \Delta U$; in other words

$$c_j(\partial_t u_j - \Delta u_j) \le 0 \quad (j = 1, \dots, m) \quad \Longrightarrow \quad \partial_t U - \Delta U \le 0. \tag{4-2}$$

This observation is a kind of reverse counterpart to the observation, or "closure property",

$$\partial_t u_j - \Delta u_j \ge 0 \quad (j = 1, \dots, m) \quad \Longrightarrow \quad \partial_t U - \Delta U \ge 0 \tag{4-3}$$

recently presented in [Bennett and Bez 2019]; in the manner described above, the closure property (4-3) generates the forward version of the geometric Brascamp–Lieb inequality (i.e., $\mathrm{F}(\boldsymbol{B}, \boldsymbol{c}) = 1$ for geometric Brascamp–Lieb data). The perspective taken in the article [loc. cit.] is that it is natural to place oneself in the framework of sub/supersolutions to heat equations since collections of sub/supersolutions are closed under a wide variety of operations (and this is not the case for bona fide solutions). As a result, one is able to generate, in a *systematic manner* by combining various closure properties, monotone quantities which underpin a number of fundamental inequalities in geometric analysis and neighbouring fields.

In this section, we build on [loc. cit.] and present a new closure property which generalizes (4-2) to general Brascamp–Lieb data; later we apply our closure property as a key step in our proof of Theorem 2.1.

For any Brascamp–Lieb datum $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$ and gaussian input $\boldsymbol{A} = (A_j)_{j=1}^m$, with $A_j \in S_+(\mathbb{R}^{n_j})$, we write

$$M(\boldsymbol{B}, \boldsymbol{c}; \boldsymbol{A}) := \sum_{j=1}^{m} c_j B_j^* A_j B_j$$

and

$$\widetilde{M}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}) := M(\boldsymbol{B}, \boldsymbol{c}; \boldsymbol{A}) + \mathcal{Q}.$$

When there is no danger of any confusion, we shall often omit the dependence on the data and simply write $M(\boldsymbol{A})$ and $\widetilde{M}(\boldsymbol{A})$ (sometimes we may also drop the dependence on $\boldsymbol{A}$).

Theorem 4.1. *Let $(\boldsymbol{B}, \boldsymbol{c}, \boldsymbol{G}, \mathcal{Q})$ be a nondegenerate generalized Brascamp–Lieb datum and suppose $\widetilde{M}(\boldsymbol{A})$ is positive definite. Assume further that*

$$c_j(A_j^{-1} - B_j\widetilde{M}(\boldsymbol{A})^{-1}B_j^*) \le 0, \tag{4-4}$$

$$(A_j^{-1} - B_j\widetilde{M}(\boldsymbol{A})^{-1}B_j^*)(G_j - A_j) = 0 \tag{4-5}$$

for $j = 1, \dots, m$. Given $u_j : (1,\infty)\times\mathbb{R}^{n_j} \to (0,\infty)$ for $j = 1, \dots, m$, let $U : (1,\infty)\times\mathbb{R}^n \to (0,\infty)$ be given by

$$U(t,x) := t^{-\alpha}\prod_{j=0}^{m+1} u_j(t, B_j x)^{c_j}.$$

Here $\alpha := \frac{1}{2}\big(n - \sum_{j=0}^{m+1} c_j n_j\big)$, $c_0 = 1$, $c_{m+1} = -1$, and u_0 and u_{m+1} are given by

$$\partial_t u_0 = \frac{1}{4\pi}\operatorname{div}(\mathcal{Q}_+^{-1}\nabla u_0), \qquad u_0(1) = g_{\mathcal{Q}_+},$$

$$\partial_t u_{m+1} = \frac{1}{4\pi}\operatorname{div}(\mathcal{Q}_-^{-1}\nabla u_{m+1}), \quad u_{m+1}(1) = g_{\mathcal{Q}_-}.$$

If

$$c_j\Big(\partial_t u_j - \frac{1}{4\pi}\operatorname{div}(A_j^{-1}\nabla u_j)\Big) \le 0 \quad \text{and} \quad D^2(\log u_j) \ge -\frac{2\pi G_j}{t} \tag{4-6}$$

for $j = 1, \dots, m$, then

$$\partial_t U - \frac{1}{4\pi}\operatorname{div}(\widetilde{M}(\boldsymbol{A})^{-1}\nabla U) \le 0. \tag{4-7}$$

Remarks. (1) When we apply Theorem 4.1, $\boldsymbol{A}$ will be such that the gaussian input $(g_{A_j})_{j=1}^m$ minimizes the inverse Brascamp–Lieb constant over all gaussian inputs of type $\boldsymbol{G}$. For such a minimizer, we shall see (in Lemma 5.3) that conditions of the form (4-4) and (4-5) are satisfied.

(2) Suppose $u_j : (1,\infty)\times\mathbb{R}^{n_j} \to (0,\infty)$ is a *solution* to the heat equation

$$\partial_t u_j = \frac{1}{4\pi}\operatorname{div}(A_j^{-1}\nabla u_j), \quad u_j(1, x_j) = f_j(x_j),$$

where $f_j \in \mathcal{T}(G_j)$. Then we know from (3-19) that $u_j \in \mathcal{T}((G_j^{-1} + (t-1)A_j^{-1})^{-1})$. Thus, if $A_j \le G_j$ then we have $u_j \in \mathcal{T}(t^{-1}G_j)$ and it follows from Lemma 3.3 that $D^2(\log u_j) \ge -2\pi t^{-1}G_j$. In this sense, the log-convexity component in the assumption (4-6) is reasonable. At the end of this subsection, we add a few further remarks on this.

(3) Theorem 4.1 is very much in the spirit of [Bennett and Bez 2019, Theorem 3.7] in which it is shown that U is a supersolution in the case $\mathcal{Q} = 0$, $c_j > 0$, and $A_j = G_j$ for each j; that is, in such a setting, if $A_j^{-1} - B_j M(\boldsymbol{A})^{-1}B_j^* \ge 0$ for each j, then we have $\partial_t U - \frac{1}{4\pi}\operatorname{div}(M(\boldsymbol{A})^{-1}\nabla U) \ge 0$ under the assumption that

$$\partial_t u_j - \frac{1}{4\pi}\operatorname{div}(A_j^{-1}\nabla u_j) \ge 0 \quad \text{and} \quad D^2(\log u_j) \ge -\frac{2\pi A_j}{t}.$$

In fact, [loc. cit., Theorem 3.7] also provides the conclusion that U obeys the log-convexity inequality

$$D^2(\log U) \ge -\frac{2\pi M(\boldsymbol{A})}{2t},$$

which can be easily deduced from the identity

$$D^2(\log U) = \sum_{j=1}^{m} c_j B_j^* D^2(\log u_j) B_j \tag{4-8}$$

and the positivity of each c_j. In the setting of mixed signatures for the c_j in Theorem 4.1, a closure property related to the log-convexity assumption seems less apparent.

Proof of Theorem 4.1. For simplicity, we write $\widetilde{M} = \widetilde{M}(\boldsymbol{A})$ in this proof. Let us define $A_0 := \mathcal{Q}_+$ and $A_{m+1} := \mathcal{Q}_-$, in which case the decomposition (3-10) can be seen as $\mathcal{Q} = c_0 B_0^* A_0 B_0 + c_{m+1} B_{m+1}^* A_{m+1} B_{m+1}$. By straightforward computations

$$\frac{\partial_t U}{U}(t, x) = -\frac{\alpha}{t} + \sum_{j=0}^{m+1} c_j \frac{\partial_t u_j}{u_j}(t, B_j x). \tag{4-9}$$

Similarly, we have

$$\nabla U(t, x) = U(t, x) \sum_{j=0}^{m+1} c_j B_j^* \boldsymbol{v}_j(t, B_j x), \tag{4-10}$$

where $\boldsymbol{v}_j := \nabla \log u_j$, from which it follows that

$$\frac{\operatorname{div}(\widetilde{M}^{-1}\nabla U)}{U}(t, x) = \sum_{j=0}^{m+1} c_j \operatorname{div}(B_j \widetilde{M}^{-1} B_j^* \boldsymbol{v}_j) + \langle \bar{\boldsymbol{w}}, \widetilde{M}^{-1} \bar{\boldsymbol{w}} \rangle,$$

where $\bar{\boldsymbol{w}} := \sum_{j=0}^{m+1} c_j B_j^* \boldsymbol{v}_j$. Here, and in what follows, on the right-hand side we are suppressing the argument of the functions; precisely speaking, we should write $u_j(t, B_j x)$, $\boldsymbol{v}_j(t, B_j x)$ and $\bar{\boldsymbol{w}}(t, x)$.

From (4-6) and definitions of u_0, u_{m+1}, we have

$$\begin{aligned}\frac{\partial_t U}{U}(t, x) &\le -\frac{\alpha}{t} + \frac{1}{4\pi} \sum_{j=0}^{m+1} c_j \frac{\operatorname{div}(A_j^{-1}\nabla u_j)}{u_j} \\ &= -\frac{\alpha}{t} + \frac{1}{4\pi} \sum_{j=0}^{m+1} c_j \operatorname{div}(A_j^{-1} \boldsymbol{v}_j) + \frac{1}{4\pi} \sum_{j=0}^{m+1} c_j \langle \boldsymbol{v}_j, A_j^{-1} \boldsymbol{v}_j \rangle\end{aligned}$$

and therefore

$$U^{-1}\Big(\partial_t U - \frac{1}{4\pi} \operatorname{div}(\widetilde{M}^{-1}\nabla U)\Big)(t, x) \le \frac{1}{4\pi}(I + II),$$

where

$$I := -\langle \bar{\boldsymbol{w}}, \widetilde{M}^{-1} \bar{\boldsymbol{w}} \rangle + \sum_{j=0}^{m+1} c_j \langle \boldsymbol{v}_j, A_j^{-1} \boldsymbol{v}_j \rangle, \quad II := -\frac{4\pi\alpha}{t} + \sum_{j=0}^{m+1} c_j \operatorname{div}((A_j^{-1} - B_j \widetilde{M}^{-1} B_j^*) \boldsymbol{v}_j).$$

Since $(B_0, \boldsymbol{B}_+)$ is surjective from the nondegeneracy assumption, we know from (3-8) in Lemma 3.2 that $I \le 0$. For II, we write

$$II = -\frac{4\pi\alpha}{t} + \sum_{j=0}^{m+1} \operatorname{tr}\big(c_j (A_j^{-1} - B_j \widetilde{M}^{-1} B_j^*) D^2 \log u_j\big).$$

From (4-4), (4-6) and (4-5), it follows that for $j = 1, \dots, m$,

$$\begin{aligned}\operatorname{tr}(c_j(A_j^{-1} - B_j\widetilde{M}^{-1}B_j^*)D^2\log u_j) &\le -\frac{2\pi}{t}\operatorname{tr}(c_j(A_j^{-1} - B_j\widetilde{M}^{-1}B_j^*)G_j)\\ &= -\frac{2\pi c_j}{t}\operatorname{tr}((A_j^{-1} - B_j\widetilde{M}^{-1}B_j^*)A_j) = -\frac{2\pi c_j}{t}(n_j - \operatorname{tr}(\widetilde{M}^{-1}B_j^*A_jB_j)).\end{aligned}$$

On the other hand, for $j = 0, m+1$, from the explicit formula (3-19) we know that

$$D^2(\log u_j) = -\frac{2\pi A_j}{t}.$$

Hence, regardless of the sign of $A_j^{-1} - B_j\widetilde{M}^{-1}B_j^*$, we have

$$\begin{aligned}\operatorname{tr}(c_j(A_j^{-1} - B_j\widetilde{M}^{-1}B_j^*)D^2\log u_j) &= -\frac{2\pi}{t}\operatorname{tr}(c_j(A_j^{-1} - B_j\widetilde{M}^{-1}B_j^*)A_j)\\ &= -\frac{2\pi c_j}{t}(n_j - \operatorname{tr}(\widetilde{M}^{-1}B_j^*A_jB_j))\end{aligned}$$

for $j = 0, m+1$. Therefore, from the definition of $\widetilde{M}$, we get

$$II \le -\frac{4\pi\alpha}{t} - \frac{2\pi}{t}\sum_{j=0}^{m+1} c_j(n_j - \operatorname{tr}(\widetilde{M}^{-1}B_j^*A_jB_j)) = 0. \qquad \square$$

We end this subsection with two further observations related to Remarks (2) and (3) after the statement of Theorem 4.1. The first concerns the analogue of Theorem 4.1 for which $c_j > 0$ for all $j = 1, \dots, m$ (in which case we drop the nondegenerate condition and assume $\mathcal{Q} \ge 0$, namely $n_{m+1} = 0$). It is clear from the above proof (and (4-8)) that in such a case, if we assume

$$A_j^{-1} - B_j\widetilde{M}(\boldsymbol{A})^{-1}B_j^* \ge 0$$

and (4-5), then

$$\partial_t u_j - \frac{1}{4\pi}\operatorname{div}(A_j^{-1}\nabla u_j) \ge 0 \quad \text{and} \quad D^2(\log u_j) \ge -\frac{2\pi G_j}{t} \tag{4-11}$$

for $j = 1, \dots, m$ imply

$$\partial_t U - \frac{1}{4\pi}\operatorname{div}(\widetilde{M}(\boldsymbol{A})^{-1}\nabla U) \ge 0 \quad \text{and} \quad D^2(\log U) \ge -\frac{2\pi\widetilde{M}(\boldsymbol{G})}{2t}. \tag{4-12}$$

This observation extends [Bennett and Bez 2019, Theorem 3.7] to the setting $\boldsymbol{A} \le \boldsymbol{G}$ and $\mathcal{Q} \ge 0$.

Our second observation here concerns the log-convexity assumption in (4-6), and we claim that it would be reasonable to assume

$$D^2(\log u_j) \ge -2\pi(G_j^{-1} + (t-1)A_j^{-1})^{-1}. \tag{4-13}$$

Indeed, as we have noted several times, solutions of the heat equation (3-15) with $u_j(1, \cdot) \in \mathcal{T}(G_j)$ satisfy $u_j(t, \cdot) \in \mathcal{T}((G_j^{-1} + (t-1)A_j^{-1})^{-1})$, and hence (4-13) by Lemma 3.3. We claim that, when $c_j > 0$ for all $j = 1, \dots, m$, then (4-13) is also closed under the operation $(u_1, \dots, u_m) \mapsto U$ in the sense that

$$D^2(\log U) \ge -2\pi(\widetilde{M}(\boldsymbol{G})^{-1} + (t-1)\widetilde{M}(\boldsymbol{A})^{-1})^{-1}. \tag{4-14}$$

To see this, following the notation in [Hiai 2010] we write

$$X : Y := (X^{-1} + Y^{-1})^{-1}$$

for the harmonic mean of the positive semidefinite transformations X and Y, and note the fundamental facts (see, for example, [loc. cit., Corollary 3.1.6]):

(I) $S^*(X : Y)S \le (S^*XS) : (S^*YS)$.

(II) $(X_1 : Y_1) + (X_2 : Y_2) \le (X_1 + X_2) : (Y_1 + Y_2)$.

If we first use (4-8) and (4-13), we get

$$D^2(\log U) \ge -2\pi \sum_{j=0}^{m} c_j B_j^*(G_j : A_{j,t})B_j,$$

where $c_0 = 1$, $A_0 = G_0 = \mathcal{Q}_+$ and $A_{j,t} := (t-1)^{-1}A_j$. However

$$\begin{aligned}\left(\sum_{j=0}^{m} c_j B_j^* G_j B_j\right) : \left(\sum_{j=0}^{m} c_j B_j^* A_{j,t} B_j\right) &\ge \sum_{j=0}^{m} (c_j B_j^* G_j B_j : c_j B_j^* A_{j,t} B_j) \\ &= \sum_{j=0}^{m} c_j (B_j^* G_j B_j : B_j^* A_{j,t} B_j) \ge \sum_{j=0}^{m} c_j B_j^*(G_j : A_{j,t})B_j,\end{aligned}$$

where we have successively applied (II) (in an iterative way for sums of m transformations), linearity, and (I). This yields (4-14).

We incorporate the preceding observation into the following (independent) result in the spirit of [Bennett and Bez 2019, Theorem 3.7].

Theorem 4.2. *Let $(\boldsymbol{B}, \boldsymbol{c})$ be a Brascamp–Lieb datum with $c_j > 0$ for all $j = 1, \dots, m$ and $\mathcal{Q} = B_0^* \mathcal{Q}_+ B_0 \ge 0$. Suppose $\boldsymbol{A} = (A_j)_{j=1}^m$, $\boldsymbol{G} = (G_j)_{j=1}^m$, with $A_j, G_j \in S_+(\mathbb{R}^{n_j})$, are such that $\widetilde{M}(\boldsymbol{A})$ and $\widetilde{M}(\boldsymbol{G})$ are positive definite. Assume further that*

$$\begin{aligned} A_j^{-1} - B_j \widetilde{M}(\boldsymbol{A})^{-1} B_j^* &\ge 0, \\ (A_j^{-1} - B_j \widetilde{M}(\boldsymbol{A})^{-1} B_j^*)(G_j - A_j) &= 0 \end{aligned}$$

for $j = 1, \dots, m$. Given $u_j : (1, \infty) \times \mathbb{R}^{n_j} \to (0, \infty)$ for $j = 1, \dots, m$, let $U : (1, \infty) \times \mathbb{R}^n \to (0, \infty)$ be given by

$$U(t, x) := t^{-\alpha} \prod_{j=0}^{m} u_j(t, B_j x)^{c_j},$$

where $\alpha := \frac{1}{2}\big(n - \sum_{j=0}^m c_j n_j\big)$, $c_0 = 1$, and u_0 be as in Theorem 4.1. If

$$\partial_t u_j \ge \frac{1}{4\pi} \operatorname{div}(A_j^{-1} \nabla u_j)) \quad \text{and} \quad D^2(\log u_j) \ge -2\pi (G_j^{-1} + (t-1)A_j^{-1})^{-1}$$

for $j = 1, \dots, m$, then

$$\partial_t U \ge \frac{1}{4\pi} \operatorname{div}(\widetilde{M}(\boldsymbol{A})^{-1} \nabla U) \quad \text{and} \quad D^2(\log U) \ge -2\pi (\widetilde{M}(\boldsymbol{G})^{-1} + (t-1)\widetilde{M}(\boldsymbol{A})^{-1})^{-1}.$$

5. Proof Theorem 2.1

The proof makes use of the key decomposition in Proposition 3.4.

Section 5.1. Inspired by ideas in [Bennett et al. 2008], we begin by establishing Theorem 2.1 in the so-called "gaussian extremizable" case (see Definition 5.1 below for the precise definition) via a heat-flow monotonicity argument; see Theorem 5.2. At this stage, we appeal to Theorem 4.1.

Section 5.2. In order to effectively reduce to the gaussian extremizable case, we introduce the notion of "amplifying data" (Definition 5.4), for which gaussian extremizers always exist, and then complete the proof of Theorem 2.1 by showing that arbitrary nondegenerate Brascamp–Lieb data can be approximated in an appropriate sense by amplifying data.

Before beginning the proof of Theorem 2.1, we recall the notation

$$M(\boldsymbol{B}, \boldsymbol{c}; \boldsymbol{A}) := \sum_{j=1}^{m} c_j B_j^* A_j B_j$$

for any Brascamp–Lieb datum $(\boldsymbol{B}, \boldsymbol{c})$ and input $\boldsymbol{A} = (A_j)_{j=1}^m$, with $A_j \in S_+(\mathbb{R}^{n_j})$. We shall write $\boldsymbol{A}_+ := (A_j)_{j=1}^{m_+}$ and $\boldsymbol{A}_- := (A_j)_{j=m_++1}^m$, and similarly for $\boldsymbol{c}_\pm$ and $\boldsymbol{G}_\pm$. Using this notation, for example, the nondegeneracy condition (2-5) becomes

$$M(\boldsymbol{B}_+, \boldsymbol{c}_+; \boldsymbol{G}_+) + \mathcal{Q} > 0.$$

Also it is natural to introduce the class

$$\Lambda(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}) := \{\boldsymbol{A} = (A_j)_{j=1}^m : A_j \in S_+(\mathbb{R}^{n_j}),\, M(\boldsymbol{B}, \boldsymbol{c}; \boldsymbol{A}) + \mathcal{Q} > 0\}. \tag{5-1}$$

For example, by definition, we have $\mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}) < \infty$ if and only if $\boldsymbol{A} \in \Lambda(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$. Also we note that the nondegenerate condition (1-7) ensures that $\Lambda(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$ is nonempty; see [Barthe and Wolff 2022, Proposition 2.2]. In fact, we have already used this observation in the remark after Lemma 3.7 in showing that Theorem 2.1 implies Theorem 1.3.

5.1. ***Gaussian extremizable data.*** First we introduce the definition of gaussian extremizable data and then proceed to show that Theorem 2.1 holds for such data.

Definition 5.1 (gaussian extremizable data). The generalized Brascamp–Lieb datum $(\boldsymbol{B}, \boldsymbol{c}, \boldsymbol{G}, \mathcal{Q})$ is said to be *gaussian extremizable* if

$$\mathrm{I_g}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G}) = \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A})$$

for some $\boldsymbol{A} \in \Lambda(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$ with $\boldsymbol{A} \le \boldsymbol{G}$, and (with a slight abuse of terminology) we refer to such $\boldsymbol{A}$ as a *gaussian extremizer*.

Theorem 5.2. *Let $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G})$ be a nondegenerate generalized Brascamp–Lieb datum. Then the following are equivalent*:

(1) *$(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G})$ is gaussian extremizable.*

(2) *There exists $\boldsymbol{A} \in \Lambda(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$ with $\boldsymbol{A} \le \boldsymbol{G}$ satisfying* (4-4) *and* (4-5) *for $j = 1, \dots, m$.*

(3) *There exists* $\boldsymbol{A} \in \Lambda(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$ *with* $\boldsymbol{A} \le \boldsymbol{G}$, *such that*

$$\mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G}) = \mathrm{I}_{\mathrm{g}}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G}) = \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}) \in (0, \infty).$$

The first thing to notice is that the implication (3) $\Rightarrow$ (1) is just a consequence of the definition of gaussian extremizability. Also, the implication (1) $\Rightarrow$ (2) is a consequence of the following.

Lemma 5.3. *Suppose the nondegenerate generalized Brascamp–Lieb datum* $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G})$ *is gaussian extremizable. Then for any gaussian extremizer* $\boldsymbol{A} \in \Lambda(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$ *with* $\boldsymbol{A} \le \boldsymbol{G}$ *we have*

$$c_j(A_j^{-1} - B_j \widetilde{M}^{-1} B_j^*) \le 0, \tag{5-2}$$

$$(A_j^{-1} - B_j \widetilde{M}^{-1} B_j^*)(G_j - A_j) = 0 \tag{5-3}$$

for all $j = 1, \dots, m$, *where* $\widetilde{M} := M(\boldsymbol{B}, \boldsymbol{c}; \boldsymbol{A}) + \mathcal{Q}$.

Proof of Lemma 5.3. We follow the basic strategy behind the argument in the proof of [Bennett et al. 2008, Corollary 8.11] with certain minor simplifications. We also remark that $\widetilde{M} > 0$ since $\boldsymbol{A} \in \Lambda(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$.

The gaussian extremizability assumption implies

$$\Phi(\boldsymbol{X}) \ge \Phi(\boldsymbol{A}) \quad \text{for all } \boldsymbol{X} \le \boldsymbol{G}, \tag{5-4}$$

where

$$\Phi(\boldsymbol{X}) := \sum_{j=1}^m c_j \log\det X_j - \log\det\Big(\sum_{j=1}^m c_j B_j^* X_j B_j + \mathcal{Q}\Big).$$

We also have the identity

$$\frac{d}{d\varepsilon}\Phi(A_1, \dots, A_j + \varepsilon D_j, \dots, A_m)|_{\varepsilon=0+} = c_j \operatorname{tr}((A_j^{-1} - B_j \widetilde{M}^{-1} B_j^*) D_j) \tag{5-5}$$

for any linear map $D_j : \mathbb{R}^{n_j} \to \mathbb{R}^{n_j}$, which can easily be checked using the fact that $\log\det(I + \varepsilon A) = (\operatorname{tr} A)\varepsilon + O(\varepsilon^2)$.

Now we fix $j \in \{1, \dots, m\}$. Since $A_j \le G_j$, for arbitrary $N_j \le 0$ we clearly have

$$0 < A_j + \varepsilon N_j \le G_j \quad \text{for all sufficiently small } \varepsilon > 0.$$

Thus it follows from (5-4) that

$$\frac{d}{d\varepsilon}\Phi(A_1, \dots, A_j + \varepsilon N_j, \dots, A_m)|_{\varepsilon=0+} \ge 0. \tag{5-6}$$

From (5-5) we obtain

$$\operatorname{tr}(c_j(A_j^{-1} - B_j \widetilde{M}^{-1} B_j^*) N_j) \ge 0 \quad \text{for all } N_j \le 0 \tag{5-7}$$

and (5-2) follows.

To show (5-3), we write

$$P_j := -c_j(A_j^{-1} - B_j \widetilde{M}^{-1} B_j^*), \quad Q_j := G_j - A_j$$

and we begin with the seemingly weaker claim

$$\operatorname{tr}(P_j Q_j) = 0. \tag{5-8}$$

To see (5-8), since $Q_j \geq 0$ we clearly have

$$0 < A_j + \varepsilon Q_j \leq G_j \quad \text{for all } \varepsilon \in (0, 1).$$

Therefore, from (5-4) and (5-5) we have

$$\operatorname{tr}(P_j Q_j) \leq 0.$$

On the other hand, we may apply (5-7) with $N_j = -Q_j$ to see

$$\operatorname{tr}(P_j Q_j) \geq 0$$

and hence (5-8).

To obtain (5-3) from (5-8), we note that the cyclic property of the trace gives $\operatorname{tr}(R_j^* R_j) = 0$, where $R_j := P_j^{1/2} Q_j^{1/2}$. This means $R_j = 0$ and hence $P_j Q_j = P_j^{1/2} R_j Q_j^{1/2} = 0$, concluding our proof of (5-3). □

Proof of Theorem 5.2. With Lemma 5.3 in mind, to prove Theorem 5.2 it remains to show the implication (2) ⇒ (3), and thus we assume (2). Making use of the decomposition (3-10) of $\mathcal{Q}$, given an arbitrary $\boldsymbol{A} \in \Lambda(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$ with $\boldsymbol{A} \leq \boldsymbol{G}$ satisfying (4-4) and (4-5), it suffices by Lemma 3.9 to show

$$\int_{\mathbb{R}^n} \prod_{j=0}^{m+1} f_j(B_j x)^{c_j}\, dx \geq \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}) \prod_{j=1}^{m} \bigg(\int_{\mathbb{R}^{n_j}} f_j \bigg)^{c_j} \tag{5-9}$$

for arbitrary $\boldsymbol{f} \in \mathcal{N}(\boldsymbol{G})$, where $c_0 = 1$, $c_{m+1} = -1$, and

$$f_0(x_0) := e^{-\pi \langle x_0, \mathcal{Q}_+ x_0 \rangle}, \quad f_{m+1}(x_{m+1}) := e^{-\pi \langle x_{m+1}, \mathcal{Q}_- x_{m+1} \rangle}.$$

Our strategy is to use a heat-flow monotonicity argument and to set things up we regard the left-hand side of (5-9) as $Q(1)$, where

$$\mathfrak{Q}(t) := \int_{\mathbb{R}^n} U(t, x)\, dx \tag{5-10}$$

and

$$U(t, x) := t^{-\alpha} \prod_{j=0}^{m+1} u_j(t, B_j x)^{c_j},$$

with

$$\alpha := \frac{1}{2} \bigg(n - \sum_{j=0}^{m+1} c_j n_j \bigg).$$

Here, the function u_j satisfies the heat equation

$$\partial_t u_j = \frac{1}{4\pi} \operatorname{div}(A_j^{-1} \nabla u_j), \quad u_j(1, x_j) = f_j(x_j) \tag{5-11}$$

for $j = 0, \ldots, m+1$, where

$$A_0 := \mathcal{Q}_+, \quad A_{m+1} := \mathcal{Q}_-.$$

In order to prove (5-9), it suffices to show that $\mathfrak{Q}$ given by (5-10) is nonincreasing on $(1, \infty)$. Indeed, from the nonincreasingness of $\mathfrak{Q}$ we may obtain

$$\int_{\mathbb{R}^n} \prod_{j=0}^{m+1} f_j(B_j x)^{c_j}\, dx = Q(1) \geq \liminf_{t \to \infty} \mathfrak{Q}(t).$$

Furthermore, by an elementary change of variables we may write

$$\mathfrak{Q}(t) = \int_{\mathbb{R}^n} \prod_{j=0}^{m+1} \big(t^{n_j/2} u_j(t, \sqrt{t} B_j y)\big)^{c_j}\, dy$$

and thus Fatou's lemma and (3-20) imply

$$\liminf_{t\to\infty} \mathfrak{Q}(t) \geq \int_{\mathbb{R}^n} \prod_{j=0}^{m+1} \bigg((\det A_j)^{1/2} e^{-\pi \langle B_j y, A_j B_j y\rangle} \int_{\mathbb{R}^{n_j}} f_j \bigg)^{c_j} dy = \bigg(\frac{\prod_{j=1}^m (\det A_j)^{c_j}}{\det \widetilde{M}} \bigg)^{1/2} \prod_{j=1}^m \bigg(\int_{\mathbb{R}^{n_j}} f_j \bigg)^{c_j}.$$

In order to verify that $\mathfrak{Q}$ is nonincreasing we shall make use of Theorem 4.1. Thanks to the assumption (2) we know that (4-4) and (4-5) are satisfied for $j = 1, \dots, m$. Hence we deduce from Theorem 4.1 that U satisfies

$$\partial_t U \leq \frac{1}{4\pi} \operatorname{div}(\widetilde{M}^{-1} \nabla U)$$

and we would like to rigorously argue along the lines of (4-1) in order to show that $\mathfrak{Q}$ is nonincreasing. First, we justify the integration by parts step, and then show that one can interchange the time derivative and the integral.

Take $\varepsilon > 0$ sufficiently small (specified momentarily) and note the bound

$$\prod_{j=1}^m u_j(t, B_j x)^{c_j} \leq C_*(t, \varepsilon)(1 + |x|^2)^N e^{-(1-\varepsilon)\pi \langle x, t^{-1} M_+ x\rangle}$$

given by (3-18) in Lemma 3.10, where $N := \sum_{j=m_+ +1}^m |c_j| n_j$, $M_+ := M(\boldsymbol{B}_+, \boldsymbol{c}_+; \boldsymbol{A}_+)$, and $C_*(t, \varepsilon)$ is a constant depending on t, $\boldsymbol{B}$, $\boldsymbol{c}$, $\boldsymbol{A}$, $\boldsymbol{G}$, and $\boldsymbol{f}$, which is locally uniformly bounded in t. Since

$$u_j(t, x_j) = t^{-n_j/2} e^{-\pi \langle x_j, t^{-1} A_j x_j\rangle}$$

for $j = 0, m+1$ where we recall that $A_0 = \mathcal{Q}_+$, $A_{m+1} = \mathcal{Q}_-$, $c_0 = 1$, and $c_{m+1} = -1$, we have

$$U(t, x) \leq C_*(t, \varepsilon)(1 + |x|^2)^N e^{-\pi \langle x, t^{-1} P(\varepsilon) x\rangle}, \tag{5-12}$$

with

$$P(\varepsilon) := (1 - \varepsilon) M_+ + \mathcal{Q}.$$

Since $\boldsymbol{A} \in \Lambda(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$ it follows that $M_+ + \mathcal{Q} > 0$ and thus $P(\varepsilon) > 0$ by choosing $\varepsilon > 0$ sufficiently small depending on $\boldsymbol{B}$, $\boldsymbol{c}$ and $\boldsymbol{A}$; consequently, U is rapidly decreasing in space locally uniformly in time. By (4-10) and (3-16) we have

$$|\nabla U(t, x)| \leq C_*(t, \varepsilon) U(t, x) \sum_{j=0}^{m+1} |c_j| (1 + |x|)^{n_j}$$

and therefore $|\nabla U(t, x)|$ is rapidly decreasing in space locally uniformly in time. Hence, from the divergence theorem we have

$$\lim_{R\to\infty} \int_{|x|\leq R} \operatorname{div}(\widetilde{M}^{-1} \nabla U)(t, x)\, dx = 0$$

for each fixed $t > 1$.

In order to see that $\mathfrak{Q}'(t) = \int \partial_t U(t,x)\,dx$, we use the identity (4-9) along with the bounds (3-17) and (5-12) to see that $|\partial_t U(t,x)|$ is rapidly decreasing in space locally uniformly in time.

From the above, we have

$$\mathfrak{Q}'(t) = \int_{\mathbb{R}^n} \partial_t U(t,x)\,dx \le \frac{1}{4\pi} \int_{\mathbb{R}^n} \operatorname{div}(\widetilde{M}^{-1}\nabla U)(t,x)\,dx = 0,$$

and we have the desired monotonicity of $\mathfrak{Q}$ on $(1,\infty)$. □

Remarks. (1) One may establish Theorem 5.2 in a similar manner to our sketch proof of Theorem 3.1 in Section 3. The same line of reasoning was used in [Bennett et al. 2008, Proposition 8.9] in the case of the forward Brascamp–Lieb inequality and their heat-flow argument was abstracted in [loc. cit., Lemma 2.6]. By proceeding via Theorem 4.1 we have kept our proof self contained, and, as explained in the previous section, we believe that the closure property in Theorem 4.1 is of wider independent interest.

(2) An inspection of the arguments in Section 5.1 reveals that the full strength of the nondegeneracy assumption (i.e., both (1-7) and (2-5)) was not required, and the condition (2-5) can be dropped at this stage. The condition (2-5) will, however, be important for the arguments in the forthcoming Section 5.2.

(3) One can make use of the above argument with Theorem 4.2 instead of Theorem 4.1 to derive the analogous statement to Theorem 5.2 for the forward Brascamp–Lieb inequality as follows. Let $\mathcal{Q} \ge 0$, $\boldsymbol{G} = (G_j)_{j=1}^m$, with $G_j \in S_+(\mathbb{R}^{n_j})$, and the Brascamp–Lieb datum $(\boldsymbol{B}, \boldsymbol{c})$ be nondegenerate in the sense of [loc. cit.] (namely $c_j > 0$, B_j is surjective for all $j = 1,\dots,m$, and $\bigcap_{j=1}^m \ker B_j = \{0\}$). Then we have

$$\mathrm{F}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G}) = \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A})$$

for some $\boldsymbol{A} \le \boldsymbol{G}$ if and only if $\boldsymbol{A}$ satisfies

$$A_j^{-1} - B_j \widetilde{M}(\boldsymbol{A}) B_j^* \ge 0, \quad (A_j^{-1} - B_j \widetilde{M}(\boldsymbol{A}) B_j^*)(G_j - A_j) = 0.$$

5.2. ***Approximation by amplifying data and the proof of Theorem 2.1.*** In the case of the forward Brascamp–Lieb inequality, in order to reduce to the gaussian-extremizable case, the argument in [Bennett et al. 2008] naturally made use of so-called *localized* data, which means $B_j = \mathrm{id}$ and $c_j = 1$ for some $j \in \{1,\dots,m\}$. In the framework of the inverse inequality, and in particular when $\boldsymbol{B}_+$ is bijective, the condition for localized data corresponds to the case $m_+ = 1$ and such a restrictive class of data cannot be expected to play an important role in reducing to the gaussian-extremizable case. Instead, we introduce the following notion.

Definition 5.4 (amplifying data). The generalized Brascamp–Lieb datum $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G})$ is said to be *amplifying* if

$$B_j = \mathrm{id}_{\mathbb{R}^n} \quad \text{and} \quad |c_j| > \max\{c_1,\dots,c_{m_+}\} - 1$$

hold for some $j \in \{m_+ + 1,\dots,m\}$.

The crucial properties of amplifying data are that they are gaussian extremizable and are able to approximate (in an appropriate sense) any nondegenerate data; we establish these facts in Lemmas 5.5 and 5.6 below.

Lemma 5.5. *Suppose the nondegenerate generalized Brascamp–Lieb datum* $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G})$ *is amplifying. Then* $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G})$ *is gaussian extremizable and in particular, from Theorem 5.2, we have*

$$\mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G}) = \mathrm{I}_{\boldsymbol{g}}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G}).$$

Proof. Without loss of generality, we may assume

$$B_m = \mathrm{id}_{\mathbb{R}^n} \quad \text{and} \quad |c_m| > \max\{c_1, \dots, c_{m_+}\} - 1.$$

In this proof, we suppress the dependence on $(\boldsymbol{B}, \boldsymbol{c})$ and write

$$M(\boldsymbol{A}) = \sum_{j=1}^{m} c_j B_j^* A_j B_j.$$

Our first simple but important remark is that

$$\mathrm{I}_{\boldsymbol{g}}(\boldsymbol{B}, \boldsymbol{c}, \boldsymbol{G}, \mathcal{Q}) < \infty. \tag{5-13}$$

To see this, first note that $B_0^* \mathcal{Q}_+ B_0 + M(\boldsymbol{G}_+) - B_{m+1}^* \mathcal{Q}_- B_{m+1} > 0$ follows immediately from (2-5). Thus, if we consider $\boldsymbol{A} \le \boldsymbol{G}$ such that $\boldsymbol{A}_+ = \boldsymbol{G}_+$ and the components of $\boldsymbol{A}_-$ are chosen to be sufficiently small depending on $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G})$, then

$$M(\boldsymbol{A}) + \mathcal{Q} = B_0^* \mathcal{Q}_+ B_0 + M(\boldsymbol{A}_+) - B_{m+1}^* \mathcal{Q}_- B_{m+1} + M(\boldsymbol{A}_-) > 0.$$

Hence we clearly have (5-13).

Next, by a continuous extension

$$\mathrm{I}_{\mathrm{g}}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G}) = \inf_{\substack{0 \le A_j \le G_j \\ j=1,\dots,m}} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A})$$

and thus there exists $\boldsymbol{A}^\star$ such that $0 \le A_j^\star \le G_j$ for $j = 1, \dots, m$ and

$$\mathrm{I}_{\mathrm{g}}(\boldsymbol{B}, \boldsymbol{c}, \boldsymbol{G}, \mathcal{Q}) = \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}^\star).$$

Our proof will be complete once we show that $A_j^\star > 0$ for each $j = 1, \dots, m$ and $\boldsymbol{A}^\star \in \Lambda(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$. The latter claim is easily dealt with since $\boldsymbol{A}^\star \notin \Lambda(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$ means $\mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}^\star) = \infty$ and this contradicts (5-13).

Our remaining goal is to show $A_j^\star > 0$ for each $j = 1, \dots, m$. We suppose, for a contradiction, that $\det A_\ell^\star = 0$ for some $\ell \in \{1, \dots, m\}$ and consider any $\boldsymbol{A}^{(\varepsilon)} \le \boldsymbol{G}$ which converges to $\boldsymbol{A}^\star$. We shall show that

$$\lim_{\varepsilon \to 0} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}^{(\varepsilon)}) = \infty \tag{5-14}$$

and thus contradict (5-13). In order to show (5-14), it clearly suffices to consider those $\varepsilon > 0$ such that $M(\boldsymbol{A}^{(\varepsilon)}) + \mathcal{Q} > 0$, in which case we have

$$\mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}^{(\varepsilon)}) = \frac{\prod_{j=1}^{m_+} (\det A_j^{(\varepsilon)})^{c_j} \prod_{k=m_+ +1}^{m} (\det A_k^{(\varepsilon)})^{-|c_k|}}{\det(M(\boldsymbol{A}^{(\varepsilon)}) + \mathcal{Q})}. \tag{5-15}$$

Suppose first $\det A_\ell^\star = 0$ with $\ell \in \{1, \dots, m_+\}$ and, without loss of generality, we suppose $\ell = 1$. In this case we estimate from below

$$\mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}^{(\varepsilon)}) \ge C(\boldsymbol{G}) \frac{(\det A_m^{(\varepsilon)})^{-|c_m|}}{\det(B_0^* \mathcal{Q}_+ B_0 + M(\boldsymbol{A}_+^{(\varepsilon)}))} \prod_{j=1}^{m_+} (\det A_j^{(\varepsilon)})^{c_j},$$

where $C(\boldsymbol{G})$ is a positive constant depending only on $\boldsymbol{G}$. Here we have also used the fact that the determinant respects the ordering of semidefinite positive matrices. Since $B_m = \mathrm{id}_{\mathbb{R}^n}$, we have

$$B_0^* \mathcal{Q}_+ B_0 + M(\boldsymbol{A}_+^{(\varepsilon)}) - |c_m| A_m^{(\varepsilon)} \geq M(\boldsymbol{A}^{(\varepsilon)}) + \mathcal{Q} > 0$$

and therefore $\det(B_0^* \mathcal{Q}_+ B_0 + M(\boldsymbol{A}_+^{(\varepsilon)})) \geq |c_m|^n \det A_m^{(\varepsilon)}$. Also, the bijectivity of $(B_0, \boldsymbol{B}_+)$ implies

$$\det(B_0^* \mathcal{Q}_+ B_0 + M(\boldsymbol{A}_+^{(\varepsilon)})) = \det \mathcal{Q}_+ \prod_{j=1}^{m_+} c_j^{n_j} \det A_j^{(\varepsilon)}.$$

Therefore

$$\mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}^{(\varepsilon)}) \geq C(\boldsymbol{c}, \boldsymbol{G}) (\det \mathcal{Q}_+)^{-|c_m|-1} \prod_{j=1}^{m_+} (\det A_j^{(\varepsilon)})^{c_j - 1 - |c_m|}$$

and hence, using that $c_j - 1 - |c_m| < 0$ for $j = 1, \dots, m_+$, we have

$$\mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; \boldsymbol{A}^{(\varepsilon)}) \geq C(\boldsymbol{c}, \boldsymbol{G}, \mathcal{Q}_+) (\det A_1^{(\varepsilon)})^{c_1 - 1 - |c_m|}$$

for appropriate positive constants $C(\boldsymbol{c}, \boldsymbol{G})$ and $C(\boldsymbol{c}, \boldsymbol{G}, \mathcal{Q}_+)$. Since $\det A_1^{(\varepsilon)} \to 0$ and $c_1 - 1 - |c_m| < 0$ we obtain (5-14).

In the remaining case we have $\det A_\ell^\star = 0$ with $\ell \in \{m_+ + 1, \dots, m\}$ and we may suppose $\det A_j^\star > 0$ for each $j \in \{1, \dots, m_+\}$ (otherwise the above argument applies). In this case it is clear that $\prod_{j=1}^{m_+} (\det A_j^{(\varepsilon)})^{c_j} \geq C(\boldsymbol{A}^\star)$ for sufficiently small $\varepsilon > 0$ and appropriate positive constant $C(\boldsymbol{A}^\star)$. Thus it is clear that the numerator in (5-15) tends to infinity as ε tends to zero. Also, $\det(M(\boldsymbol{A}^{(\varepsilon)}) + \mathcal{Q}) \leq 2 \det(M(\boldsymbol{A}^\star) + \mathcal{Q}) < \infty$ for all $\varepsilon > 0$ sufficiently small, and hence (5-14) trivially follows. □

The following lemma ensures the approximation of arbitrary generalized Brascamp–Lieb datum by amplifying data.

Lemma 5.6. *Suppose $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G})$ is a nondegenerate generalized Brascamp–Lieb datum. Then, for any $c_+ > 0$ we have*

$$\mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G}) = \lim_{\lambda \downarrow 0} \lambda^{nc_+/2} \mathrm{I}\big((\boldsymbol{B}, \mathrm{id}_{\mathbb{R}^n}), (\boldsymbol{c}, -c_+), \mathcal{Q}, (\boldsymbol{G}, \lambda \mathrm{id}_{\mathbb{R}^n})\big), \tag{5-16}$$

$$\mathrm{I}_{\mathrm{g}}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G}) = \lim_{\lambda \downarrow 0} \lambda^{nc_+/2} \mathrm{I}_{\mathrm{g}}\big((\boldsymbol{B}, \mathrm{id}_{\mathbb{R}^n}), (\boldsymbol{c}, -c_+), \mathcal{Q}, (\boldsymbol{G}, \lambda \mathrm{id}_{\mathbb{R}^n})\big). \tag{5-17}$$

Proof. For arbitrary $\boldsymbol{f} \in \mathcal{N}(\boldsymbol{G})$, since $(B_0, \boldsymbol{B}_+)$ is bijective, we may apply the dominated convergence theorem to see

$$\begin{aligned}
\int_{\mathbb{R}^n} e^{-\pi \langle x, \mathcal{Q} x \rangle} \prod_{j=1}^m f_j(B_j x)^{c_j} \, dx &= \lim_{\lambda \downarrow 0} \int_{\mathbb{R}^n} e^{-\pi \langle x, \mathcal{Q} x \rangle} \prod_{j=1}^m f_j(B_j x)^{c_j} e^{\pi c_+ \langle x, \lambda \mathrm{id}_{\mathbb{R}^n} x \rangle} \, dx \\
&= \lim_{\lambda \downarrow 0} \lambda^{nc_+/2} \int_{\mathbb{R}^n} e^{-\pi \langle x, \mathcal{Q} x \rangle} \prod_{j=1}^m f_j(B_j x)^{c_j} g_{\lambda \mathrm{id}_{\mathbb{R}^n}}(x)^{-c_+} \, dx \\
&\geq \lim_{\lambda \downarrow 0} \lambda^{nc_+/2} \mathrm{I}((\boldsymbol{B}, \mathrm{id}_{\mathbb{R}^n}), (\boldsymbol{c}, -c_+), \mathcal{Q}, (\boldsymbol{G}, \lambda \mathrm{id}_{\mathbb{R}^n})) \prod_{j=1}^m \bigg(\int_{\mathbb{R}^{n_j}} f_j \bigg)^{c_j},
\end{aligned}$$

which, thanks to Lemma 3.9, shows

$$\mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G}) \geq \lim_{\lambda \downarrow 0} \lambda^{nc_+/2} \mathrm{I}((\boldsymbol{B}, \mathrm{id}_{\mathbb{R}^n}), (\boldsymbol{c}, -c_+), \mathcal{Q}, (\boldsymbol{G}, \lambda \mathrm{id}_{\mathbb{R}^n})).$$

In order to obtain the converse inequality, for any $\lambda > 0$ and any $(\boldsymbol{f}, f_{m+1}) \in \mathcal{T}(\boldsymbol{G}) \times \mathcal{T}(\lambda \mathrm{id}_{\mathbb{R}^n})$, we need to bound

$$\int_{\mathbb{R}^n} e^{-\pi \langle x, \mathcal{Q}x \rangle} \prod_{j=1}^{m} f_j(B_j x)^{c_j} f_{m+1}(x)^{-c_+} \, dx$$

from below. Since $f_{m+1} \in \mathcal{T}(\lambda \mathrm{id}_{\mathbb{R}^n})$, we can write $f_{m+1} = g_{\lambda \mathrm{id}_{\mathbb{R}^n}} * d\mu_{m+1}$ for some positive and finite Borel measure $d\mu_{m+1}$ and so we have

$$f_{m+1}(x) = \lambda^{n/2} \int_{\mathbb{R}^n} e^{-\lambda |x-y|^2} \, d\mu_{m+1}(y) \leq \lambda^{n/2} \int_{\mathbb{R}^n} f_{m+1}$$

uniformly in x. This yields

$$\int_{\mathbb{R}^n} e^{-\pi \langle x, \mathcal{Q}x \rangle} \prod_{j=1}^{m} f_j(B_j x)^{c_j} f_{m+1}(x)^{-c_+} \, dx \geq \lambda^{-nc_+/2} \mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G}) \prod_{j=1}^{m} \left(\int_{\mathbb{R}^{n_j}} f_j \right)^{c_j} \left(\int_{\mathbb{R}^n} f_{m+1} \right)^{-c_+},$$

which shows

$$\mathrm{I}((\boldsymbol{B}, \mathrm{id}_{\mathbb{R}^n}), (\boldsymbol{c}, -c_+), \mathcal{Q}, (\boldsymbol{G}, \lambda \mathrm{id}_{\mathbb{R}^n})) \geq \lambda^{-nc_+/2} \mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G}),$$

and we conclude (5-16). A similar argument yields (5-17). □

We are now in a position to remove the gaussian extremizability assumption in Theorem 5.2 and thus establish Theorem 2.1.

Proof of Theorem 2.1. Choose any $c_+ > 0$ such that $c_+ > \max\{c_1, \ldots, c_{m_+}\} - 1$. Then, for any $\lambda > 0$ it is clear that the augmented data $((\boldsymbol{B}, \mathrm{id}_{\mathbb{R}^n}), (\boldsymbol{c}, -c_+), \mathcal{Q}, (\boldsymbol{G}, \lambda \mathrm{id}_{\mathbb{R}^n}))$ is amplifying. Also, since $(\boldsymbol{B}, \mathrm{id}_{\mathbb{R}^n})_+ = \boldsymbol{B}_+$ the augmented data is nondegenerate and hence we may apply Lemma 5.5 to give

$$\mathrm{I}\big((\boldsymbol{B}, \mathrm{id}_{\mathbb{R}^n}), (\boldsymbol{c}, -c_+), \mathcal{Q}, (\boldsymbol{G}, \lambda \mathrm{id}_{\mathbb{R}^n})\big) = \mathrm{I}_{\mathrm{g}}\big((\boldsymbol{B}, \mathrm{id}_{\mathbb{R}^n}), (\boldsymbol{c}, -c_+), \mathcal{Q}, (\boldsymbol{G}, \lambda \mathrm{id}_{\mathbb{R}^n})\big).$$

Multiplying both sides by $\lambda^{nc_+/2}$ and taking the limit $\lambda \downarrow 0$, we obtain the desired conclusion from Lemma 5.6. □

6. Further applications and remarks

6.1. ***Regularized forms of the Young convolution inequality.*** For $c \in \mathbb{R}$, we introduce the constant

$$A_c := \left(\frac{|1-c|^{1-c}}{|c|^c} \right)^{1/2}.$$

The sharp form of the forward and inverse Young convolution inequality on $\mathbb{R}$ may be expressed as

$$\mathrm{F}(\boldsymbol{B}, \boldsymbol{c}) = A_{c_0} A_{c_1} A_{c_2} \quad (c_0, c_1, c_2 \in (0, 1]) \tag{6-1}$$

and

$$\mathrm{I}(\boldsymbol{B}, \boldsymbol{c}) = A_{c_0} A_{c_1} A_{c_2} \quad (c_j < 0 \text{ for some } j, \text{ and } c_k \in [1, \infty) \text{ for } k \neq j), \tag{6-2}$$

where, in both cases, the c_j satisfy the scaling condition $c_0+c_1+c_2=2$, and $B_j:\mathbb{R}^2\to\mathbb{R}$ are given by

$$B_0(x,y):=x,\quad B_1(x,y):=y,\quad B_2(x,y):=x-y. \tag{6-3}$$

The forward version (6-1) was established independently by Beckner [1975] and Brascamp and Lieb [1976]. In the same paper, Brascamp and Lieb established[9] (6-2).

Under the scaling condition $c_0+c_1+c_2=2$, one has an invariance of extremizers under the isotropic rescaling $f_j\to f_j(R\,\cdot)$ and so it follows that one cannot hope to improve the constants in (6-1) and (6-2) even if one only considers f_j of type[10] $1/\sigma_j$ for any $\sigma_j>0$. However, by considering such f_j, one may relax the scaling condition and below we present a result of this type. To state the result, we use the notation

$$\tilde{A}_{c,\sigma}:=\left(\frac{\sigma^{1-c}}{c}\right)^{1/2}\quad (c,\sigma>0).$$

Corollary 6.1. *Let $\boldsymbol{B}$ be given by* (6-3). *Given $c_0<1$ and $c_1,c_2>0$, suppose $\sigma_0,\sigma_1,\sigma_2>0$ satisfy*

$$\frac{\sigma_0}{1-c_0}=\frac{\sigma_1}{c_1}+\frac{\sigma_2}{c_2} \tag{6-4}$$

and set $\boldsymbol{G}=(\sigma_0^{-1},\sigma_1^{-1},\sigma_2^{-1})$.

(1) *Suppose $c_0,c_1,c_2\in(0,1)$. Then*

$$\mathrm{F}(\boldsymbol{B},\boldsymbol{c},\boldsymbol{G})=\frac{\tilde{A}_{c_1,\sigma_1}\tilde{A}_{c_2,\sigma_2}}{\tilde{A}_{1-c_0,\sigma_0}} \tag{6-5}$$

holds if and only if

$$\frac{c_0(1-c_0)}{\sigma_0}\geq\max\left\{\frac{c_1(1-c_1)}{\sigma_1},\frac{c_2(1-c_2)}{\sigma_2}\right\}. \tag{6-6}$$

(2) *Suppose $c_0<0,\ c_1,c_2\in[1,\infty)$. Then*

$$\mathrm{I}(\boldsymbol{B},\boldsymbol{c},\boldsymbol{G})=\frac{\tilde{A}_{c_1,\sigma_1}\tilde{A}_{c_2,\sigma_2}}{\tilde{A}_{1-c_0,\sigma_0}} \tag{6-7}$$

holds if and only if

$$\frac{c_0(1-c_0)}{\sigma_0}\leq\min\left\{\frac{c_1(1-c_1)}{\sigma_1},\frac{c_2(1-c_2)}{\sigma_2}\right\}. \tag{6-8}$$

Remarks. (1) Although the relaxation of the scaling condition is not explicit in the above statement, one can show that, when $c_0,c_1,c_2\in(0,1)$, if (6-6) holds for some $\sigma_j>0$ satisfying (6-4), then $c_0+c_1+c_2\geq 2$ holds. Similarly, when $c_0<0,\ c_1,c_2\in[1,\infty)$, it can be shown that if (6-8) holds for some $\sigma_j>0$ satisfying (6-4), then $c_0+c_1+c_2\leq 2$ holds.

[9]Strictly speaking, it was proved in [Brascamp and Lieb 1976] that the inequality $\|f_1*f_2\|_{p_0}\geq A_{c_0}A_{c_1}A_{c_2}\|f_1\|_{p_1}\|f_2\|_{p_2}$ holds when $p_0,p_1,p_2\in(0,1]$, where $c_0:=1-1/p_0,\ c_1:=1/p_1,\ c_2:=1/p_2$. As observed in [Barthe and Wolff 2022, Example 2.14], in dual form (6-2), it becomes apparent that there is a symmetry amongst c_0,c_1,c_2, or equivalently, p_0',p_1,p_2. As clarified in [loc. cit., Example 2.14], the condition on the c_j in (6-2) is necessary and corresponds to the bijectivity of $\boldsymbol{B}_+$.

[10]In this discussion, we use the notation $1/\sigma_j$ to facilitate a comparison with the related results in [Bennett and Bez 2009].

(2) Sharp forms of Young convolution inequalities have been previously considered in [Bennett and Bez 2009] (not in the dual setting, but in terms of the norm inequality). In particular, it follows from [loc. cit., Corollary 7] that (6-5) holds under the condition (6-6). However, the condition in [loc. cit.] is that

$$c_0 + c_1 + c_2 \geq 2$$

and, for some $\alpha_1, \alpha_2 \in [0, 1]$,

$$c_0 + \alpha_1 c_1 + \alpha_2 c_2 = 2 \quad \text{and} \quad c_1(1 - \alpha_1 c_1)\sigma_2 = c_2(1 - \alpha_2 c_2)\sigma_1.$$

Although is not immediately obvious, one can show the equivalence of this condition with (6-6). The inverse inequality (6-7) under the relaxed scaling condition was not explicitly stated in [loc. cit.], but the arguments there can be modified to show that (6-7) follows from (6-8).

(3) The meaning of (6-5) and (6-7) is that the regularized constants are attained "on the boundary". In other words, for example, (6-7) is equivalent to

$$\mathrm{I}(\boldsymbol{B}, \boldsymbol{c}, \boldsymbol{G}) = \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}; \boldsymbol{G}). \tag{6-9}$$

As one would expect, the proof will proceed via Theorem 5.2 and, essentially, our contribution in Corollary 6.1 is showing that the conditions for gaussian extremizability in Theorem 5.2 can be equivalently expressed in the simple form (6-8) (and similarly for the forward inequality).

Proof of Corollary 6.1. We give the details for the inverse case (since the forward case is similar) and hence assume $c_0 < 0,\ c_1, c_2 \in [1, \infty)$. As remarked above, the goal is to show (6-9) is equivalent to (6-8).

From $c_0 < 0,\ c_1, c_2 \in [1, \infty)$ it is immediate that the nondegenerate condition (1-8) is satisfied for our datum and hence we may apply Theorem 5.2 to see that (6-9) holds if and only if

$$\Gamma_0(\boldsymbol{G}) \geq 0, \quad \Gamma_j(\boldsymbol{G}) \leq 0 \quad (j = 1, 2), \tag{6-10}$$

where

$$\Gamma_j(a_0, a_1, a_2) := a_j^{-1} - \frac{\sum_{k=0,1,2:k\neq j} c_k a_k}{c_0 c_1 a_0 a_1 + c_0 c_2 a_0 a_2 + c_1 c_2 a_1 a_2} \tag{6-11}$$

for $a_0, a_1, a_2 > 0$. Indeed, a straightforward computation reveals that the condition (6-10) coincides with (4-4) (with $\boldsymbol{A} = \boldsymbol{G}$) for our datum.

From (6-4) it is easy to check that $\Gamma_0(\boldsymbol{G}) = 0$ and thus it suffices to show that

$$\Gamma_j(\boldsymbol{G}) \leq 0 \quad \Longleftrightarrow \quad \frac{c_0(1 - c_0)}{\sigma_0} \leq \frac{c_j(1 - c_j)}{\sigma_j} \tag{6-12}$$

for $j = 1, 2$. To see this for $j = 1$, first note that (6-4) yields

$$\frac{c_0 c_1}{\sigma_0 \sigma_1} + \frac{c_0 c_2}{\sigma_0 \sigma_2} + \frac{c_1 c_2}{\sigma_1 \sigma_2} = \frac{1}{1 - c_0} \frac{c_1 c_2}{\sigma_1 \sigma_2}$$

and hence $\Gamma_1(\boldsymbol{G}) \leq 0$ is equivalent to

$$\sigma_1 \leq (1 - c_0) \frac{\sigma_1 \sigma_2}{c_1 c_2} \left(\frac{c_0}{\sigma_0} + \frac{c_2}{\sigma_2} \right).$$

This, however, can be rearranged to $c_0(1-c_0)/\sigma_0 \leq c_1(1-c_1)/\sigma_1$ by making use of (6-4). The equivalence (6-12) for $j = 2$ can be verified in the same manner and this completes our proof. □

6.2. ***Regularized Prékopa–Leindler inequality.*** We begin by recalling the one-dimensional Prékopa–Leindler inequality which states that if c_1, c_2 satisfy the scaling condition

$$c_1 + c_2 = 1, \tag{6-13}$$

then, for all nonnegative $f_1,\, f_2 \in L^1(\mathbb{R})$,

$$\left(\int_{\mathbb{R}} f_1\right)^{c_1}\left(\int_{\mathbb{R}} f_2\right)^{c_2} \le \int_{\mathbb{R}} \operatorname*{ess\,sup}_{\substack{x_1,x_2\in\mathbb{R}\\ x=c_1x_1+c_2x_2}} f_1(x_1)^{c_1} f_2(x_2)^{c_2}\, dx. \tag{6-14}$$

The necessity of the scaling condition (6-13) is an easy consequence of the scaling argument. As in the case for the Young convolution inequality, one might expect to salvage the Prékopa–Leindler inequality in a scale-free case $c_1 + c_2 \neq 1$ by restricting inputs to regularized datum. For $c_1, c_2 \in (0, 1]$ and $\sigma_1, \sigma_2 > 0$, we define $\mathrm{PL}(\boldsymbol{c}, (\sigma_1^{-1}, \sigma_2^{-1})) \in [0, \infty]$ to be the sharp constant for the inequality

$$\left(\int_{\mathbb{R}} f_1\right)^{c_1}\left(\int_{\mathbb{R}} f_2\right)^{c_2} \le C\int_{\mathbb{R}} \operatorname*{ess\,sup}_{\substack{x_1,x_2\in\mathbb{R}\\ x=c_1x_1+c_2x_2}} f_1(x_1)^{c_1} f_2(x_2)^{c_2}\, dx, \tag{6-15}$$

where $(f_1,\, f_2) \in \mathcal{T}(\sigma_1^{-1}) \times \mathcal{T}(\sigma_2^{-1})$. From Theorem 2.2, then we have that, regardless of $c_1, c_2 \in (0, 1]$,

$$\mathrm{PL}(\boldsymbol{c}, (\sigma_1^{-1}, \sigma_2^{-1})) = \left(\inf_{\substack{0<a_j\le\sigma_j^{-1}\\ j=1,2}} \Phi_{\boldsymbol{c}}(a_1, a_2)\right)^{-1/2}, \tag{6-16}$$

where

$$\Phi_{\boldsymbol{c}}(a_1, a_2) := a_1^{c_1} a_2^{c_2}\left(\frac{c_1}{a_1} + \frac{c_2}{a_2}\right).$$

Moreover one can show that $\mathrm{PL}(\boldsymbol{c}, (\sigma_1^{-1}, \sigma_2^{-1})) < \infty$ as long as $c_1 + c_2 < 1$ and $\sigma_1, \sigma_2 > 0$. Here we use (6-16) to identify the exact constant under certain conditions on σ_1, σ_2.

Corollary 6.2. *Let $c_1, c_2 \in (0, 1)$ satisfy $c_1 + c_2 < 1$. For $\sigma_1, \sigma_2 > 0$, it holds that*

$$\mathrm{PL}(\boldsymbol{c}, (\sigma_1^{-1}, \sigma_2^{-1})) = \Phi_{\boldsymbol{c}}(\sigma_1^{-1}, \sigma_2^{-1})^{-1/2} = \left(\frac{\sigma_1^{c_1}\sigma_2^{c_2}}{c_1\sigma_1 + c_2\sigma_2}\right)^{1/2} \tag{6-17}$$

if and only if

$$c_1\sigma_1 + c_2\sigma_2 \le \min\{\sigma_1, \sigma_2\}. \tag{6-18}$$

Proof. We begin with the sufficiency part. Observe that

$$\partial_1\Phi_{\boldsymbol{c}}(a_1, a_2) = c_1 a_1^{c_1-1} a_2^{c_2}\left(-\frac{1-c_1}{a_1} + \frac{c_2}{a_2}\right), \quad \partial_2\Phi_{\boldsymbol{c}}(a_1, a_2) = c_2 a_1^{c_1} a_2^{c_2-1}\left(\frac{c_1}{a_1} - \frac{1-c_2}{a_2}\right)$$

and hence

$$\partial_1\Phi_{\boldsymbol{c}}(a_1, a_2) = \partial_2\Phi_{\boldsymbol{c}}(a_1, a_2) = 0 \quad \Longleftrightarrow \quad c_1 + c_2 = 1.$$

In particular, in the case $c_1+c_2 < 1$, there is no extremum on $\{(a_1, a_2) : 0 < a_j < \sigma_j^{-1}\}$. Hence the minimum of $\Phi_{\boldsymbol{c}}(a_1, a_2)$ is attained on the boundary of $[0, \sigma_1^{-1}] \times [0, \sigma_2^{-1}]$. Furthermore, since $\Phi_{\boldsymbol{c}}(a_1, a_2) = \infty$ if either $a_1 = 0$ or $a_2 = 0$, we see that

$$\inf_{0<a_j<\sigma_j^{-1}} \Phi_{\boldsymbol{c}}(a_1, a_2) = \min\left(\inf_{0<a_1<\sigma_1^{-1}} \Phi_{\boldsymbol{c}}(a_1, \sigma_2^{-1}), \inf_{0<a_2<\sigma_2^{-1}} \Phi_{\boldsymbol{c}}(\sigma_1^{-1}, a_2)\right).$$

Hence it suffices to investigate $\Phi_{\boldsymbol{c}}(a_1, \sigma_2^{-1})$ and $\Phi_{\boldsymbol{c}}(a_1, \sigma_2^{-1})$. For fixed $\sigma_1^{-1}, \sigma_2^{-1} > 0$, we see from the formula of $\partial_j \Phi_{\boldsymbol{c}}$ that

$$\partial_1 \Phi_{\boldsymbol{c}}(a_1, \sigma_2^{-1}) = 0 \quad \Longleftrightarrow \quad a_1 = a_1^* := \frac{1 - c_1}{c_2 \sigma_2}$$

and similarly

$$\partial_2 \Phi_{\boldsymbol{c}}(\sigma_1^{-1}, a_2) = 0 \quad \Longleftrightarrow \quad a_2 = a_2^* := \frac{1 - c_2}{c_1 \sigma_1}.$$

Now we appeal to our assumption $c_1\sigma_1 + c_2\sigma_2 \le \min(\sigma_1, \sigma_2)$. In fact, from this we see that

$$a_1^* \notin [0, \sigma_1^{-1}), \quad a_2^* \notin [0, \sigma_2^{-1}).$$

Namely, $\Phi_{\boldsymbol{c}}(a_1, \sigma_2^{-1})$ is monotone on $[0, \sigma_1^{-1}]$ and similarly $\Phi_{\boldsymbol{c}}(\sigma_1^{-1}, a_2)$ is monotone on $[0, \sigma_2^{-1}]$. We conclude that

$$\inf_{0 < a_1 < \sigma_1^{-1}} \Phi_{\boldsymbol{c}}(a_1, \sigma_2^{-1}) = \Phi_{\boldsymbol{c}}(\sigma_1^{-1}, \sigma_2^{-1}), \qquad \inf_{0 < a_2 < \sigma_2^{-1}} \Phi_{\boldsymbol{c}}(\sigma_1^{-1}, a_2) = \Phi_{\boldsymbol{c}}(\sigma_1^{-1}, \sigma_2^{-1}),$$

which implies (6-17).

To show the necessity of (6-18), we define

$$\Psi(a_1, a_2) := \log \Phi_{\boldsymbol{c}}(a_1, a_2) = \sum_{j=1,2} c_j \log a_j + \log(c_1 a_1^{-1} + c_2 a_2^{-1}), \quad a_1, a_2 > 0.$$

Then from the assumption (6-17) we see that

$$\frac{d}{d\varepsilon} \Psi(\sigma_1^{-1} - \varepsilon, \sigma_2^{-1})|_{\varepsilon = 0} := \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \big(\Psi(\sigma_1^{-1} - \varepsilon, \sigma_2^{-1}) - \Psi(\sigma_1^{-1}, \sigma_2^{-1}) \big) \ge 0.$$

On the other hand, we have

$$\frac{d}{d\varepsilon} \Psi(\sigma_1^{-1} - \varepsilon, \sigma_2^{-1})|_{\varepsilon = 0} = -c_1\sigma_1 + \sigma_1 - \frac{c_2}{c_1 \sigma_2^{-1} + c_2 \sigma_1^{-1}}$$

and hence it follows that

$$\frac{c_2}{c_1 \sigma_2^{-1} + c_2 \sigma_1^{-1}} - (1 - c_1)\sigma_1 \le 0,$$

which yields $c_1\sigma_1 + c_2\sigma_2 \le \sigma_1$. Similarly, by considering a perturbation in a_2, we obtain $c_1\sigma_1 + c_2\sigma_2 \le \sigma_2$. □

6.3. ***Regularized forms of hypercontractivity inequalities.*** We conclude this section with some remarks about the forward and inverse hypercontractivity inequalities of the form

$$\|e^{s\mathcal{L}}(F^{1/p})\|_{L^q(d\gamma)} \le \left(\int_{\mathbb{R}} F \, d\gamma \right)^{1/p} \quad (p > 1,\ s > 0) \tag{6-19}$$

for all nonnegative $F \in L^1(d\gamma)$, and

$$\|e^{s\mathcal{L}}(F^{1/p})\|_{L^q(d\gamma)} \ge \left(\int_{\mathbb{R}} F \, d\gamma \right)^{1/p} \quad (0 \ne p < 1,\ s > 0) \tag{6-20}$$

for all positive $F \in L^1(d\gamma)$. Here, $(e^{s\mathcal{L}})_{s>0}$ is the Ornstein–Uhlenbeck semigroup given by

$$e^{s\mathcal{L}}F(x) := \int_{\mathbb{R}} F(e^{-s}x + (1-e^{-2s})^{1/2}y)\, d\gamma(y),$$

where γ is the density function of the standard normal distribution

$$d\gamma(x) := g_{1/(2\pi)}(x)\, dx.$$

Also, the exponents $p, q \in \mathbb{R}$ and $s > 0$ in (6-19) and (6-20) satisfy the relation

$$e^{2s} = \frac{q-1}{p-1}, \tag{6-21}$$

and the constant 1 appearing in both inequalities is optimal. The forward inequality is due to [Nelson 1973] and the inverse inequality is due to [Borell 1982]; we refer the reader to, for example, [Bakry et al. 2014] for further details about the importance of inequalities of this kind.

There are a number of ways to obtain (6-19) and (6-20), one of which is to write

$$\|e^{s\mathcal{L}}(F^{1/p})\|_{L^q(d\gamma)} = C\|(Fg_{1/(2\pi)})^{1/p} * g_\star\|_{L^q(dx)},$$

where C is a constant and $g_\star$ is an isotropic gaussian (both explicitly computable). From this expression, one may obtain (6-19) and (6-20) from the sharp form of the forward and inverse Young convolution inequalities; see [Beckner 1975, Theorem 5]. Since one of the inputs is a fixed gaussian, the scale-invariance property of the Young convolution inequality for general inputs ceases to hold, and one may thus expect to improve the constant in (6-19) and (6-20) by restricting to inputs F which are regularized in an appropriate sense. One can obtain certain results of this type via Theorem 2.1 (and its forward counterpart) by using the representation

$$\|e^{s\mathcal{L}}(F^{1/p})\|_{L^q(d\gamma)} = C(p,s)\int_{\mathbb{R}^2} e^{-\pi\langle x, \mathcal{Q}x\rangle} \prod_{j=1,2} f_j(B_j x)^{c_j}\, dx,$$

where

$$C(p,s) := (2\pi)^{(1/2)(1/p+1/q')-1}(1-e^{-2s})^{-1/2},$$
$$c_1 := \frac{1}{p}, \quad c_2 = \frac{1}{q'}, \quad B_j(x_1,x_2) := x_j,$$

and

$$\mathcal{Q} := \frac{1}{2\pi(1-e^{-2s})}\begin{pmatrix} 1-(1-e^{-2s})\frac{1}{p} & -e^{-s} \\ -e^{-s} & 1-(1-e^{-2s})\frac{1}{q'} \end{pmatrix},$$

$$f_1(x_1) := F(x_1)g_{1/(2\pi)}(x_1), \quad f_2(x_2) := \frac{e^{s\mathcal{L}}(F^{1/p})(x_2)^q}{\|e^{s\mathcal{L}}(F^{1/p})\|^q_{L^q(d\gamma)}} g_{1/(2\pi)}(x_2).$$

We refrain from explicitly stating such results here since stronger results have been obtained in collaboration with Hiroshi Tsuji [Bez et al. 2023] and we refer the reader there for precise statements. For instance, we also proved that the best constant for the logarithmic Sobolev inequality and Talagrand's inequality can be improved by restricting inputs to certain regularized functions.

We also remark that it would be natural to consider bounds on more general gaussian kernel operators and to investigate the extent to which these are quantitatively improved upon by heat-flow regularization. For instance, it is already interesting to consider the gaussian kernel above for p, q that do not satisfy (6-21). For such data, the nondegeneracy condition (1-7) fails to hold and moreover one cannot expect a nontrivial Brascamp–Lieb constant; see [Barthe and Wolff 2022; Nakamura and Tsuji 2022]. From similar reasoning, such data does not appear to fit into the framework of the forward-reverse Brascamp–Lieb inequality in Theorem 2.2. However, the second author and Tsuji [Nakamura and Tsuji 2022; 2024] very recently observed that one can recover the Brascamp–Lieb inequality associated to such data if one restricts the inputs f_1, f_2 to be *even* functions, and moreover showed how such an improvement implies the functional Blaschke–Santaló inequality.[11]

Appendix: On the equivalence between Theorems 2.1 and 2.2

Theorem 2.2 ⇒ Theorem 2.1. Fix a nondegenerate datum $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G})$ in the sense of (1-7) and (2-5), where the B_i are mappings from $\mathbb{R}^n$ to $\mathbb{R}^{n_i}$, $i = 1, \dots, m$. To show Theorem 2.1, it suffices to show

$$\int_{\mathbb{R}^n} e^{-\pi\langle x, \mathcal{Q}x\rangle} \prod_{i=1}^{m_+} f_i(B_i x)^{c_i} \prod_{j=m_++1}^{m} h_j(B_j x)^{c_j}\, dx \geq \mathrm{I} \prod_{i=1}^{m_+} \left(\int_{\mathbb{R}^{n_i}} f_i\right)^{c_i} \prod_{j=m_++1}^{m} \left(\int_{\mathbb{R}^{n_j}} h_j\right)^{c_j} \tag{A-1}$$

for $f_i \in \mathcal{T}(G_i)$ and $h_j \in \mathcal{T}(G_j)$, where

$$\mathrm{I} := \inf_{\substack{A_i \leq G_i \\ i=1,\dots,m}} \frac{\prod_{i=1}^m (\det A_i)^{c_i/2}}{\det\left(\mathcal{Q} + \sum_{i=1}^m c_i B_i^* A_i B_i\right)^{1/2}}.$$

First we use Proposition 3.4 to get the decomposition

$$\mathcal{Q} = B_0^* \mathcal{Q}_+ B_0 - B_{m+1}^* \mathcal{Q}_- B_{m+1}.$$

Here, $\Phi := (B_0, \boldsymbol{B}_+) : \mathbb{R}^n \to E$ is bijective, where $E := \bigoplus_{i=0}^{m_+} \mathbb{R}^{n_i}$. Also, $\ker \boldsymbol{B}_+ \subseteq \ker B_{m+1}$, where $\mathcal{Q}_\pm$ are positive definite on $\mathbb{R}^{n_0}$ and $\mathbb{R}^{n_{m+1}}$ respectively. For reasons that will soon become apparent, we set

$$Q_L := \mathcal{Q}_+, \quad Q_R := (\Phi^{-1})^* B_{m+1}^* \mathcal{Q}_- B_{m+1} \Phi^{-1}.$$

Clearly we have

$$B_i \circ \Phi^{-1} = \pi_i, \quad i = 0, 1, \dots, m_+, \tag{A-2}$$

and, with this in mind, we see that (2-5) implies (in fact, is equivalent to)

$$(\Phi^{-1})^* \left(B_0^* \mathcal{Q}_+ B_0 - B_{m+1}^* \mathcal{Q}_- B_{m+1} + \sum_{i=1}^{m_+} c_i B_i^* G_i B_i \right) \Phi^{-1} > 0,$$

[11] For further information regarding the functional Blaschke–Santaló inequality, we refer the reader to [Artstein-Avidan et al. 2004; Ball 1986; Cordero-Erausquin et al. 2025; Courtade et al. 2024; Fathi 2018; Fradelizi and Meyer 2007; Fradelizi et al. 2023; Kolesnikov and Werner 2022; Lehec 2009a; 2009b]. For more detailed discussion about this new link to the volume product, including Mahler's conjecture, we refer the interested reader to [Nakamura and Tsuji 2022; 2024].

and hence also

$$\pi_0^* Q_L \pi_0 - Q_R + \sum_{i=1}^{m_+} c_i \pi_i^* G_i \pi_i > 0. \tag{A-3}$$

With the nondegeneracy condition (2-7) in mind, at this point we also observe that

$$\mathbb{R}^{n_0} \oplus \{0\} \oplus \cdots \oplus \{0\} \subseteq \ker Q_R \tag{A-4}$$

holds. To see this, it clearly suffices to check $B_{m+1}\Phi^{-1}(x_0, 0, \ldots, 0) = 0$. However,

$$\begin{aligned} \boldsymbol{B}_+ \Phi^{-1}(x_0, 0, \ldots, 0) &= (B_1\Phi^{-1}(x_0, 0, \ldots, 0), \ldots, B_{m_+}\Phi^{-1}(x_0, 0, \ldots, 0)) \\ &= (\pi_1(x_0, 0, \ldots, 0), \ldots, \pi_{m_+}(x_0, 0, \ldots, 0)) = 0 \end{aligned}$$

and since $\ker \boldsymbol{B}_+ \subseteq \ker B_{m+1}$, we obtain the desired conclusion.

Next, given $f_i \in \mathcal{T}(G_i),\ i = 1, \ldots, m_+$ and $h_j \in \mathcal{T}(G_j),\ j = m_+ + 1, \ldots, m$, we introduce the function $h_{m+1} : E \to \mathbb{R}_+$ by

$$h_{m+1}(x) := e^{-\pi\langle \pi_0 x, \mathcal{Q}_+ \pi_0 x\rangle} \prod_{i=1}^{m_+} f_i(B_i \circ \Phi^{-1}(x))^{c_i} \times e^{\pi\langle \Phi^{-1}(x), B_{m+1}^* \mathcal{Q}_- B_{m+1}\Phi^{-1}(x)\rangle} \prod_{j=m_++1}^{m} h_j(B_j \circ \Phi^{-1}(x))^{c_j}.$$

Setting $T_j := B_j \circ \Phi^{-1}$ for $j = m_+ + 1, \ldots, m$, and $T_{m+1} = \mathrm{id}$, we then have

$$h_{m+1}(T_{m+1}x) = e^{-\pi\langle \pi_0 x, Q_L \pi_0 x\rangle} \prod_{i=1}^{m_+} f_i(\pi_i x)^{c_i} \times e^{\pi\langle x, Q_R x\rangle} \prod_{j=m_++1}^{m} h_j(T_j x)^{c_j}$$

thanks to (A-2). In particular, (2-9) holds for inputs $(f_i)_{i=1}^{m_+}$ and $(h_j)_{j=m_++1}^{m+1}$, and exponents $\boldsymbol{d} = ((c_i)_{i=1}^{m_+}, (-c_{m_++1}, \ldots, -c_m, 1))$. Therefore, we may apply Theorem 2.2 with

$$\mathfrak{D} = ((T_j)_{j=m_++1}^{m+1}, \boldsymbol{d}, Q, ((G_i)_{i=1}^{m_+}, (G_{m_++1}, \ldots, G_m, \infty)))$$

to see

$$\prod_{i=1}^{m_+} \left(\int_{\mathbb{R}^{n_i}} f_i \right)^{c_i} \prod_{j=m_++1}^{m} \left(\int_{\mathbb{R}^{n_j}} h_j \right)^{c_j} \le \mathrm{FR}(\mathfrak{D}) \int_E h_{m+1}, \tag{A-5}$$

where

$$\mathrm{FR}(\mathfrak{D}) = \sup (\det A)^{1/2} \prod_{i=1}^{m} (\det A_i)^{-c_i/2}$$

and the supremum is taken over all $A_i \le G_i$, $i = 1, \ldots, m$, and $A > 0$ satisfying

$$\pi_0^* Q_L \pi_0 + \sum_{i=1}^{m_+} c_i \pi_i^* A_i \pi_i \ge Q_R + \sum_{j=m_++1}^{m} (-c_j) T_j^* A_j T_j + A. \tag{A-6}$$

Here we remark that, strictly speaking, we are using a variant of Theorem 2.2 which admits $h_{m+1} \in \mathcal{T}(\infty)$. Such a result can be quickly obtained from Theorem 2.2 by a limiting argument, and this explains why the above matrix A is an arbitrary positive definite matrix. The nondegeneracy condition is not affected by this limiting argument, and we have already verified it above in (A-3) and (A-4).

Notice that $x_i = \pi_i(x_0, \dots, x_{m_+}) = B_i \Phi^{-1}(x_0, \dots, x_{m_+})$ from (A-2). Hence, by the change of variable $\Phi^{-1}(x_0, \dots, x_{m_+}) = y$, we see that

$$\int_E h_{m+1}(x_0, \dots, x_{m_+})\, dx_0 \cdots dx_{m_+} = \det \Phi \int_{\mathbb{R}^n} e^{-\pi \langle y, B_0^* \mathcal{Q}_+ B_0 y \rangle} e^{\pi \langle y, B_{m+1}^* \mathcal{Q}_- B_{m+1} y \rangle} \prod_{i=1}^{m_+} f_i(B_i y)^{c_i} \prod_{j=m_++1}^{m} h_j(B_j y)^{c_j}\, dy.$$

Hence, it follows from (A-5) that

$$\int_{\mathbb{R}^n} e^{-\pi \langle x, \mathcal{Q} x \rangle} \prod_{i=1}^{m_+} f_i(B_i y)^{c_i} \prod_{j=m_++1}^{m} h_j(B_j y)^{c_j}\, dy \geq \mathrm{FR}(\mathfrak{D})^{-1} \det \Phi^{-1} \prod_{i=1}^{m_+} \left(\int_{\mathbb{R}^{n_i}} f_i \right)^{c_i} \prod_{j=m_++1}^{m} \left(\int_{\mathbb{R}^{n_j}} h_j \right)^{c_j}.$$

It remains to estimate $\mathrm{FR}(\mathfrak{D})$. To this end, we note that (A-6) is equivalent to

$$(\Phi^{-1})^* \left(\mathcal{Q} + \sum_{i=1}^{m} c_i B_i^* A_i B_i \right) \Phi^{-1} \geq A$$

and hence

$$\det A \leq \frac{1}{(\det \Phi)^2} \det \left(\mathcal{Q} + \sum_{i=1}^{m} c_i B_i^* A_i B_i \right).$$

This shows that

$$\mathrm{FR}(\mathfrak{D}) \leq \frac{1}{\det \Phi} \sup_{\substack{A_i \leq G_i \\ i=1,\dots,m}} \det \left(\mathcal{Q} + \sum_{i=1}^{m} c_i B_i^* A_i B_i \right)^{1/2} \prod_{i=1}^{m} (\det A_i)^{-c_i/2} = \frac{1}{\det \Phi} \times \frac{1}{\mathrm{I}},$$

which concludes the proof of (A-1). □

Theorem 2.1 ⇒ Theorem 2.2. We follow a limiting argument due to Wolff and presented in [Courtade and Liu 2021, Section 4]. First, let us show the following.

Claim A.1. *Suppose $\mathfrak{D} = (\boldsymbol{T}, \boldsymbol{d}, Q, \boldsymbol{G})$ is a generalized forward-reverse Brascamp–Lieb datum which is nondegenerate (notation from Section 2.2 will prevail). Consider the Brascamp–Lieb datum $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q})$ defined by*

$$\begin{aligned} \boldsymbol{B} &:= (\pi_1, \dots, \pi_I, T_1, \dots, T_J), \\ \boldsymbol{c} &:= (d_1, \dots, d_I, -d(1), \dots, -d(J)), \\ \mathcal{Q} &:= \pi_0^* Q_L \pi_0 - Q_R. \end{aligned}$$

If $f_i \in \mathcal{T}(G_i)$, $h_j \in \mathcal{T}(G(j))$ and $\tilde{h} \in \mathcal{T}(\infty)$ satisfy

$$e^{-\pi \langle \pi_0 x, Q_L \pi_0 x \rangle} \prod_{i=1}^{I} f_i(\pi_i x)^{d_i} \leq e^{-\pi \langle x, Q_R x \rangle} \prod_{j=1}^{J} h_j(T_j x)^{d(j)} \times \tilde{h}(x), \quad x \in E, \tag{A-7}$$

then

$$\mathrm{I}(\mathfrak{D}) \prod_{i=1}^{I} \left(\int_{E_i} f_i \right)^{d_i} \leq \prod_{j=1}^{J} \left(\int_{E(j)} h_j \right)^{d(j)} \int_E \tilde{h}, \tag{A-8}$$

where

$$\mathrm{I}(\mathfrak{D}) := \inf_{\substack{A_i \leq G_i \\ A(j) \leq G(j)}} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}; (A_i)_{i=1}^{I}, (A(j))_{j=1}^{J}).$$

Proof of Claim A.1. In order to apply Theorem 2.1, let us check the nondegenerate conditions (1-7) and (2-5). It is clear from the setup that $\boldsymbol{B}_+ = (\pi_i)_{i=1}^I$ and thus the first condition in (2-7) implies $\ker \boldsymbol{B}_+ \subseteq \ker Q_R$. This means $\mathcal{Q}x = \pi_0^* Q_L \pi_0 x$ whenever $x \in \ker \boldsymbol{B}_+$, which verifies the first condition in (1-7). Also, we note that $s^+(\mathcal{Q}) \le n_0$ and hence $s^+(\mathcal{Q}) + \sum_{i=1}^I n_i \le \dim E$; this ensures the remaining condition in (1-7). Finally, we observe that (2-5) is a direct consequence of the second condition in (2-7).

Now (A-7) clearly implies

$$\int_E \tilde{h} \ge \int_E e^{-\pi\langle x, \mathcal{Q}x\rangle} \prod_{i=1}^I f_i(\pi_i x)^{d_i} \prod_{j=1}^J h_j(T_j x)^{-d(j)}\, dx$$

and an application of Theorem 2.1 with the datum $(\boldsymbol{B}, \boldsymbol{c}, \mathcal{Q}, \boldsymbol{G})$ yields (A-8). □

Returning to the proof that Theorem 2.1 implies Theorem 2.2, we start by fixing a generalized forward-reverse Brascamp–Lieb datum $\mathfrak{D} = (\boldsymbol{T}, \boldsymbol{d}, Q, \boldsymbol{G})$ which is nondegenerate. For each $t > 0$, we set

$$d_{1,t} := 1 + td_1, \dots, d_{I,t} := 1 + td_I, \quad d_t(1) := td(1), \dots, d_t(J) := td(J)$$

and

$$Q_{L,t} := tQ_L, \quad Q_{R,t} := tQ_R,$$

and consider the family of forward-reverse Brascamp–Lieb data $\mathfrak{D}_t = (\boldsymbol{T}, \boldsymbol{d}_t, Q_t, \boldsymbol{G})$. Since the original data $\mathfrak{D}$ is nondegenerate, it is easy to verify that $\mathfrak{D}_t$ is nondegenerate for each $t > 0$.

Now fix $f_i \in \mathcal{T}(G_i)$, $h_j \in \mathcal{T}(G(j))$ satisfying (2-9) and set

$$\tilde{h}_t(x) := e^{-\pi\langle \pi_0 x, Q_{L,t}\pi_0 x\rangle} \prod_{i=1}^I f_i(\pi_i x)^{d_{i,t}} e^{\pi\langle x, Q_{R,t}x\rangle} \prod_{j=1}^J h_j(T_j x)^{-d_t(j)} \quad (x \in E).$$

By Claim A.1,

$$\mathrm{I}(\mathfrak{D}_t) \prod_{i=1}^I \Big(\int_{E_i} f_i\Big)^{d_{i,t}} \le \prod_{j=1}^J \Big(\int_{E(j)} h_j\Big)^{d_t(j)} \int_E \tilde{h}_t,$$

where

$$\mathrm{I}(\mathfrak{D}_t) := \inf_{\substack{A_i \le G_i \\ A(j) \le G(j)}} \mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}_t, \mathcal{Q}_t; ((A_i)_{i=1}^I, (A(j))_{j=1}^J)).$$

Here, $\boldsymbol{B}$ is exactly as in Claim A.1, and

$$\boldsymbol{c}_t := (d_{i,t})_{i=1}^I, (-d_t(j))_{j=1}^J, \quad \mathcal{Q}_t := \pi_0^* Q_{L,t}\pi_0 - Q_{R,t}.$$

Notice that (2-9) and the definition of the exponents $d_{i,t}, d_t(j)$ yield

$$\tilde{h}_t(x) = \prod_{i=1}^I f_i(x_i) \Big(e^{-\pi\langle x_0, Q_L x_0\rangle} \prod_{i=1}^I f_i(x_i)^{d_i} e^{\pi\langle x, Q_R x\rangle} \prod_{j=1}^J h_j(T_j x)^{-d(j)}\Big)^t \le \prod_{i=1}^I f_i(x_i)$$

and hence

$$\mathrm{I}(\mathfrak{D}_t) \prod_{i=1}^I \Big(\int_{E_i} f_i\Big)^{d_{i,t}} \le \prod_{j=1}^J \Big(\int_{E(j)} h_j\Big)^{d_t(j)} \prod_{i=1}^I \int_{E_i} f_i.$$

After rearranging terms and taking a limit, we obtain

$$\prod_{i=1}^{I}\Big(\int_{E_i} f_i\Big)^{d_i} \le \liminf_{t\to\infty} \mathrm{I}(\mathfrak{D}_t)^{-1/t} \prod_{j=1}^{J}\Big(\int_{E(j)} h_j\Big)^{d(j)}$$

and so it suffices to check

$$\liminf_{t\to\infty} \mathrm{I}(\mathfrak{D}_t)^{-1/t} \le \mathrm{FR}(\mathfrak{D}).$$

To investigate $\mathrm{I}(\mathfrak{D}_t)^{-1/t}$, we compute

$$\begin{aligned}
&\int_E e^{-\pi\langle x,\mathcal{Q}_t x\rangle} \prod_{i=1}^{I} \gamma_{A_i}(x_i)^{d_{i,t}} \prod_{j=1}^{J} \gamma_{A(j)}(T_j x)^{-d_t(j)}\,dx \\
&\qquad = \int_E \exp\Big(-\pi\Big\langle x, \big(\pi_0^* Q_{L,t}\pi_0 + \sum_{i=1}^{I} d_{i,t}\pi_i^* A_i \pi_i - Q_{R,t} - \sum_{j=1}^{J} d_t(j) T_j^* A(j) T_j\big)x\Big\rangle\Big)\,dx \\
&\qquad = \det\Big(t\pi_0^* Q_L \pi_0 + \sum_{i=1}^{I}(1+td_i)\pi_i^* A_i\pi_i - tQ_R - \sum_{j=1}^{J} td(j) T_j^* A(j) T_j\Big)^{-1/2}
\end{aligned}$$

if A_i, $A(j)$ satisfy

$$t\pi_0^* Q_L \pi_0 + \sum_{i=1}^{I}(1+td_i)\pi_i^* A_i\pi_i - tQ_R - \sum_{j=1}^{J} td(j) T_j^* A(j) T_j > 0;$$

otherwise the integral coincides with $+\infty$. Moreover, since we take $t\to\infty$, we may restrict attention to A_i, $A(j)$ satisfying

$$\pi_0^* Q_L \pi_0 + \sum_{i=1}^{I} d_i \pi_i^* A_i \pi_i - Q_R - \sum_{j=1}^{J} d(j) T_j^* A(j) T_j \ge 0,$$

which coincides with condition (2-8). With this in mind, and from the lower bound

$$\begin{aligned}
&\mathrm{BL}(\boldsymbol{B}, \boldsymbol{c}_t, \mathcal{Q}_t; ((A_i)_{i=1}^I, (A(j))_{j=1}^J)) \\
&\qquad = \frac{\int_E e^{-\pi\langle x,\mathcal{Q}_t x\rangle} \prod_{i=1}^{I} \gamma_{A_i}(x_i)^{d_{i,t}} \prod_{j=1}^{J} \gamma_{A(j)}(T_j x)^{-d_t(j)}\,dx}{\prod_{i=1}^{I}\big(\int_{E_i}\gamma_{A_i}\big)^{d_{i,t}} \prod_{j=1}^{J}\big(\int_{E(j)}\gamma_{A(j)}\big)^{-d_t(j)}} \\
&\qquad \ge \det\Big(t\pi_0^* Q_L \pi_0 + \sum_{i=1}^{I}(1+td_i)\pi_i^* A_i\pi_i\Big)^{-1/2} \prod_{i=1}^{I}(\det A_i)^{(1+td_i)/2} \prod_{j=1}^{J}(\det A(j))^{-td(j)/2} \\
&\qquad = \Big(t^{n_0}\det Q_L \prod_{i=1}^{I}(1+td_i)^{n_i} \prod_{i=1}^{I}(\det A_i)^{-td_i} \prod_{j=1}^{J}(\det A(j))^{td(j)}\Big)^{-1/2}
\end{aligned}$$

we see that

$$\begin{aligned}
\liminf_{t\to\infty} \mathrm{I}(\mathfrak{D}_t)^{-1/t} &\le \liminf_{t\to\infty}\Big(t^{n_0}\det Q_L \prod_{i=1}^{I}(1+td_i)^{n_i}\Big)^{1/(2t)} \sup_{\substack{A_i\le G_i, A(j)\le G(j)\\ \text{(2-8)}}} \prod_{i=1}^{I}(\det A_i)^{-d_i/2} \prod_{j=1}^{J}(\det A(j))^{d(j)/2} \\
&= \sup_{\substack{A_i\le G_i, A(j)\le G(j)\\ \text{(2-8)}}} \prod_{i=1}^{I}(\det A_i)^{-d_i/2} \prod_{j=1}^{J}(\det A(j))^{d(j)/2}
\end{aligned}$$

which concludes the proof. □

Acknowledgements

Bez would like to thank Jon Bennett for numerous enlightening conversations around the subject of this work. Nakamura would like to thank the organizers of the Harmonic Analysis Seminar held at Shinshu University in 2020 where he learned several ideas used in this work. Finally, both authors would like to express their sincere gratitude to an anonymous referee whose comments led to substantial improvements in the paper. In particular, the referee suggested to us that a result such as Theorem 2.2 should be true, and the manner in which it is equivalent to Theorem 2.1.

References

[Aishwarya and Rotem 2023] G. Aishwarya and L. Rotem, "New Brunn–Minkowski and functional inequalities via convexity of entropy", preprint, 2023. arXiv 2311.05446

[Aoki et al. 2020] Y. Aoki, J. Bennett, N. Bez, S. Machihara, K. Matsuura, and S. Shiraki, "A supersolutions perspective on hypercontractivity", *Ann. Mat. Pura Appl.* (4) **199**:5 (2020), 2105–2116. MR Zbl

[Artstein-Avidan et al. 2004] S. Artstein-Avidan, B. Klartag, and V. Milman, "The Santaló point of a function, and a functional form of the Santaló inequality", *Mathematika* **51**:1-2 (2004), 33–48. MR Zbl

[Bakry et al. 2014] D. Bakry, I. Gentil, and M. Ledoux, *Analysis and geometry of Markov diffusion operators*, Grundl. Math. Wissen. **348**, Springer, 2014. MR Zbl

[Ball 1986] K. Ball, *Isometric problems in ℓ_p and sections of convex sets*, Ph.D. thesis, University of Cambridge, 1986.

[Ball 1991] K. Ball, "Volume ratios and a reverse isoperimetric inequality", *J. London Math. Soc.* (2) **44**:2 (1991), 351–359. MR Zbl

[Ball and Nguyen 2012] K. Ball and V. H. Nguyen, "Entropy jumps for isotropic log-concave random vectors and spectral gap", *Studia Math.* **213**:1 (2012), 81–96. MR Zbl

[Ball et al. 2003] K. Ball, F. Barthe, and A. Naor, "Entropy jumps in the presence of a spectral gap", *Duke Math. J.* **119**:1 (2003), 41–63. MR Zbl

[Barthe 1997] F. Barthe, "Inégalités de Brascamp–Lieb et convexité", *C. R. Acad. Sci. Paris Sér. I Math.* **324**:8 (1997), 885–888. MR Zbl

[Barthe 1998a] F. Barthe, "An extremal property of the mean width of the simplex", *Math. Ann.* **310**:4 (1998), 685–693. MR Zbl

[Barthe 1998b] F. Barthe, "On a reverse form of the Brascamp–Lieb inequality", *Invent. Math.* **134**:2 (1998), 335–361. MR Zbl

[Barthe and Cordero-Erausquin 2004] F. Barthe and D. Cordero-Erausquin, "Inverse Brascamp–Lieb inequalities along the heat equation", pp. 65–71 in *Geometric aspects of functional analysis*, edited by V. D. Milman and G. Schechtman, Lecture Notes in Math. **1850**, Springer, 2004. MR Zbl

[Barthe and Huet 2009] F. Barthe and N. Huet, "On Gaussian Brunn–Minkowski inequalities", *Studia Math.* **191**:3 (2009), 283–304. MR Zbl

[Barthe and Wolff 2014] F. Barthe and P. Wolff, "Positivity improvement and Gaussian kernels", *C. R. Math. Acad. Sci. Paris* **352**:12 (2014), 1017–1021. MR Zbl

[Barthe and Wolff 2022] F. Barthe and P. Wolff, *Positive Gaussian kernels also have Gaussian minimizers*, Mem. Amer. Math. Soc. **1359**, Amer. Math. Soc., Providence, RI, 2022. MR Zbl

[Beckner 1975] W. Beckner, "Inequalities in Fourier analysis", *Ann. of Math.* (2) **102**:1 (1975), 159–182. MR Zbl

[Bennett and Bez 2009] J. Bennett and N. Bez, "Closure properties of solutions to heat inequalities", *J. Geom. Anal.* **19**:3 (2009), 584–600. MR Zbl

[Bennett and Bez 2019] J. Bennett and N. Bez, "Generating monotone quantities for the heat equation", *J. Reine Angew. Math.* **756** (2019), 37–63. MR Zbl

[Bennett and Bez 2021] J. Bennett and N. Bez, "Higher order transversality in harmonic analysis", pp. 75–103 in *Harmonic analysis and nonlinear partial differential equations*, edited by S. Masaki and H. Takaoka, RIMS Kôkyûroku Bessatsu **B88**, Res. Inst. Math. Sci. (RIMS), Kyoto, 2021. MR Zbl

[Bennett et al. 2008] J. Bennett, A. Carbery, M. Christ, and T. Tao, "The Brascamp–Lieb inequalities: finiteness, structure and extremals", *Geom. Funct. Anal.* **17**:5 (2008), 1343–1415. MR Zbl

[Bennett et al. 2010] J. Bennett, A. Carbery, M. Christ, and T. Tao, "Finite bounds for Hölder–Brascamp–Lieb multilinear inequalities", *Math. Res. Lett.* **17**:4 (2010), 647–666. MR Zbl

[Bennett et al. 2017] J. Bennett, N. Bez, M. G. Cowling, and T. C. Flock, "Behaviour of the Brascamp–Lieb constant", *Bull. Lond. Math. Soc.* **49**:3 (2017), 512–518. MR Zbl

[Bennett et al. 2020] J. Bennett, N. Bez, S. Buschenhenke, M. G. Cowling, and T. C. Flock, "On the nonlinear Brascamp–Lieb inequality", *Duke Math. J.* **169**:17 (2020), 3291–3338. MR Zbl

[Bez et al. 2023] N. Bez, S. Nakamura, and H. Tsuji, "Stability of hypercontractivity, the logarithmic Sobolev inequality, and Talagrand's cost inequality", *J. Funct. Anal.* **285**:10 (2023), art. id. 110121. MR Zbl

[Borell 1982] C. Borell, "Positivity improving operators and hypercontractivity", *Math. Z.* **180**:2 (1982), 225–234. MR Zbl

[Borell 1993] C. Borell, "Geometric properties of some familiar diffusions in $\mathbb{R}^n$", *Ann. Probab.* **21**:1 (1993), 482–489. MR Zbl

[Borell 2000] C. Borell, "Diffusion equations and geometric inequalities", *Potential Anal.* **12**:1 (2000), 49–71. MR Zbl

[Borell 2003] C. Borell, "The Ehrhard inequality", *C. R. Math. Acad. Sci. Paris* **337**:10 (2003), 663–666. MR Zbl

[Bourgain 2017] J. Bourgain, "Decoupling, exponential sums and the Riemann zeta function", *J. Amer. Math. Soc.* **30**:1 (2017), 205–224. MR Zbl

[Bourgain and Guth 2011] J. Bourgain and L. Guth, "Bounds on oscillatory integral operators based on multilinear estimates", *Geom. Funct. Anal.* **21**:6 (2011), 1239–1295. MR Zbl

[Bourgain et al. 2016] J. Bourgain, C. Demeter, and L. Guth, "Proof of the main conjecture in Vinogradov's mean value theorem for degrees higher than three", *Ann. of Math.* (2) **184**:2 (2016), 633–682. MR Zbl

[Brascamp and Lieb 1976] H. J. Brascamp and E. H. Lieb, "Best constants in Young's inequality, its converse, and its generalization to more than three functions", *Advances in Math.* **20**:2 (1976), 151–173. MR Zbl

[Carlen and Cordero-Erausquin 2009] E. A. Carlen and D. Cordero-Erausquin, "Subadditivity of the entropy and its relation to Brascamp–Lieb type inequalities", *Geom. Funct. Anal.* **19**:2 (2009), 373–405. MR Zbl

[Carlen et al. 2004] E. A. Carlen, E. H. Lieb, and M. Loss, "A sharp analog of Young's inequality on S^N and related entropy inequalities", *J. Geom. Anal.* **14**:3 (2004), 487–520. MR Zbl

[Chen et al. 2015] W.-K. Chen, N. Dafnis, and G. Paouris, "Improved Hölder and reverse Hölder inequalities for Gaussian random vectors", *Adv. Math.* **280** (2015), 643–689. MR Zbl

[Cordero-Erausquin et al. 2025] D. Cordero-Erausquin, N. Gozlan, S. Nakamura, and H. Tsuji, "Duality and heat flow", *Adv. Math.* **467** (2025), art. id. 110161. MR Zbl

[Courtade and Liu 2021] T. A. Courtade and J. Liu, "Euclidean forward-reverse Brascamp–Lieb inequalities: finiteness, structure, and extremals", *J. Geom. Anal.* **31**:4 (2021), 3300–3350. MR Zbl

[Courtade et al. 2018] T. A. Courtade, M. Fathi, and A. Pananjady, "Quantitative stability of the entropy power inequality", *IEEE Trans. Inform. Theory* **64**:8 (2018), 5691–5703. MR Zbl

[Courtade et al. 2024] T. A. Courtade, M. Fathi, and D. Mikulincer, "Stochastic proof of the sharp symmetrized Talagrand inequality", *C. R. Math. Acad. Sci. Paris* **362** (2024), 1779–1784. MR Zbl

[Durcik and Thiele 2021] P. Durcik and C. Thiele, "Singular Brascamp–Lieb: a survey", pp. 321–349 in *Geometric aspects of harmonic analysis*, edited by P. Ciatti and A. Martini, Springer INdAM Ser. **45**, Springer, 2021. MR Zbl

[Eldan and Lee 2018] R. Eldan and J. R. Lee, "Regularization under diffusion and anticoncentration of the information content", *Duke Math. J.* **167**:5 (2018), 969–993. MR Zbl

[Eldan and Mikulincer 2020] R. Eldan and D. Mikulincer, "Stability of the Shannon–Stam inequality via the Föllmer process", *Probab. Theory Related Fields* **177**:3-4 (2020), 891–922. MR Zbl

[Eldan et al. 2020] R. Eldan, J. Lehec, and Y. Shenfeld, "Stability of the logarithmic Sobolev inequality via the Föllmer process", *Ann. Inst. Henri Poincaré Probab. Stat.* **56**:3 (2020), 2253–2269. MR Zbl

[Fathi 2018] M. Fathi, "A sharp symmetrized form of Talagrand's transport-entropy inequality for the Gaussian measure", *Electron. Commun. Probab.* **23** (2018), art. id. 81. MR Zbl

[Fathi et al. 2016] M. Fathi, E. Indrei, and M. Ledoux, "Quantitative logarithmic Sobolev inequalities and stability estimates", *Discrete Contin. Dyn. Syst.* **36**:12 (2016), 6835–6853. MR Zbl

[Fradelizi and Meyer 2007] M. Fradelizi and M. Meyer, "Some functional forms of Blaschke–Santaló inequality", *Math. Z.* **256**:2 (2007), 379–395. MR Zbl

[Fradelizi et al. 2023] M. Fradelizi, M. Meyer, and A. Zvavitch, "Volume product", pp. 163–222 in *Harmonic analysis and convexity*, edited by A. Koldobsky and A. Volberg, Adv. Anal. Geom. **9**, De Gruyter, Berlin, 2023. MR Zbl

[Gardner 2002] R. J. Gardner, "The Brunn–Minkowski inequality", *Bull. Amer. Math. Soc.* (*N.S.*) **39**:3 (2002), 355–405. MR Zbl

[Geng and Nair 2014] Y. Geng and C. Nair, "The capacity region of the two-receiver Gaussian vector broadcast channel with private and common messages", *IEEE Trans. Inform. Theory* **60**:4 (2014), 2087–2104. MR Zbl

[Gressman 2025] P. T. Gressman, "Testing conditions for multilinear Radon–Brascamp–Lieb inequalities", *Trans. Amer. Math. Soc.* **378**:2 (2025), 751–804. MR Zbl

[Guo and Zhang 2019] S. Guo and R. Zhang, "On integer solutions of Parsell–Vinogradov systems", *Invent. Math.* **218**:1 (2019), 1–81. MR Zbl

[Guo and Zorin-Kranich 2020] S. Guo and P. Zorin-Kranich, "Decoupling for moment manifolds associated to Arkhipov–Chubarikov–Karatsuba systems", *Adv. Math.* **360** (2020), art. id. 106889. MR Zbl

[Guo et al. 2023] S. Guo, C. Oh, R. Zhang, and P. Zorin-Kranich, "Decoupling inequalities for quadratic forms", *Duke Math. J.* **172**:2 (2023), 387–445. MR Zbl

[Hiai 2010] F. Hiai, "Matrix analysis: matrix monotone functions, matrix means, and majorization", *Interdiscip. Inform. Sci.* **16**:2 (2010), 139–248. MR Zbl

[Klartag and Milman 2008] B. Klartag and E. Milman, "On volume distribution in 2-convex bodies", *Israel J. Math.* **164** (2008), 221–249. MR Zbl

[Kolesnikov and Werner 2022] A. V. Kolesnikov and E. M. Werner, "Blaschke–Santaló inequality for many functions and geodesic barycenters of measures", *Adv. Math.* **396** (2022), art. id. 108110. MR Zbl

[Lehec 2009a] J. Lehec, "A direct proof of the functional Santaló inequality", *C. R. Math. Acad. Sci. Paris* **347**:1-2 (2009), 55–58. MR Zbl

[Lehec 2009b] J. Lehec, "Partitions and functional Santaló inequalities", *Arch. Math.* (*Basel*) **92**:1 (2009), 89–94. MR Zbl

[Lehec 2014] J. Lehec, "Short probabilistic proof of the Brascamp–Lieb and Barthe theorems", *Canad. Math. Bull.* **57**:3 (2014), 585–597. MR Zbl

[Li and Yau 1986] P. Li and S.-T. Yau, "On the parabolic kernel of the Schrödinger operator", *Acta Math.* **156**:3-4 (1986), 153–201. MR Zbl

[Lieb 1990] E. H. Lieb, "Gaussian kernels have only Gaussian maximizers", *Invent. Math.* **102**:1 (1990), 179–208. MR Zbl

[Liu et al. 2018] J. Liu, T. A. Courtade, P. W. Cuff, and S. Verdú, "A forward-reverse Brascamp–Lieb inequality: entropic duality and Gaussian optimality", *Entropy* **20**:6 (2018), art. id. 418. MR

[Maldague 2022] D. Maldague, "Regularized Brascamp–Lieb inequalities and an application", *Q. J. Math.* **73**:1 (2022), 311–331. MR Zbl

[Nakamura and Tsuji 2022] S. Nakamura and H. Tsuji, "Hypercontractivity beyond Nelson's time and its applications to Blaschke–Santaló inequality and inverse Santaló inequality", preprint, 2022. arXiv 2212.02866

[Nakamura and Tsuji 2024] S. Nakamura and H. Tsuji, "The functional volume product under heat flow", preprint, 2024. arXiv 2401.00427

[Nelson 1973] E. Nelson, "The free Markoff field", *J. Functional Analysis* **12** (1973), 211–227. MR Zbl

[Schmuckenschläger 1995] M. Schmuckenschläger, "A concentration of measure phenomenon on uniformly convex bodies", pp. 275–287 in *Geometric aspects of functional analysis* (Israel, 1992–1994), edited by J. Lindenstrauss and V. Milman, Oper. Theory Adv. Appl. **77**, Birkhäuser, Basel, 1995. MR Zbl

[Valdimarsson 2007] S. I. Valdimarsson, "On the Hessian of the optimal transport potential", *Ann. Sc. Norm. Super. Pisa Cl. Sci.* (5) **6**:3 (2007), 441–456. MR Zbl

[Valdimarsson 2008] S. I. Valdimarsson, "Optimisers for the Brascamp–Lieb inequality", *Israel J. Math.* **168** (2008), 253–274. MR Zbl

[Zhang 2022] R. Zhang, "The Brascamp–Lieb inequality and its influence on Fourier analysis", pp. 585–628 in *The physics and mathematics of Elliott Lieb—the* 90*th anniversary*, *Vol. II*, edited by R. L. Frank et al., EMS Press, Berlin, 2022. MR Zbl

[Zorin-Kranich 2020] P. Zorin-Kranich, "Kakeya–Brascamp–Lieb inequalities", *Collect. Math.* **71**:3 (2020), 471–492. MR Zbl

Received 8 Nov 2021. Revised 22 Aug 2024. Accepted 20 Sep 2024.

NEAL BEZ: nealbez@mail.saitama-u.ac.jp
bez@math.nagoya-u.ac.jp
Department of Mathematics, Graduate School of Science and Engineering, Saitama University, Saitama, Japan

and

Graduate School of Mathematics, Nagoya University, Nagoya, Japan

SHOHEI NAKAMURA: srmkn@math.sci.osaka-u.ac.jp
s.nakamura@bham.ac.uk
Department of Mathematics, Graduate School of Science, Osaka University, Osaka, Japan

and

School of Mathematics, University of Birmingham, Birmingham, United Kingdom

ANALYSIS AND PDE
Vol. 18 (2025), No. 7, pp. 1615–1713
DOI: 10.2140/apde.2025.18.1615

msp

COSMIC CENSORSHIP NEAR FLRW SPACETIMES WITH NEGATIVE SPATIAL CURVATURE

DAVID FAJMAN AND LIAM URBAN

We consider general initial data for the Einstein scalar-field system on a closed 3-manifold (M, γ) which is close to data for a Friedman–Lemaître–Robertson–Walker solution with homogeneous scalar field matter and a negative Einstein metric γ as spatial geometry. We prove that the maximal globally hyperbolic development of such initial data in the Einstein scalar-field system is past incomplete in the contracting direction and exhibits stable collapse into a big bang curvature singularity. Under an additional condition on the first positive eigenvalue of $-\Delta_\gamma$ satisfied, for example, by closed hyperbolic 3-manifolds of small diameter, we prove that the data evolves to a future complete spacetime in the expanding direction which asymptotes to a vacuum Friedman solution with (M, γ) as the expansion normalized spatial geometry. In particular, the strong cosmic censorship conjecture holds for this class of solutions in the C^2-sense.

1. Introduction

1.1. *Setting and main results.*

We consider the Einstein scalar-field system

$$\mathrm{Ric}[\bar{g}]_{\mu\nu} - \tfrac{1}{2}R[\bar{g}]\bar{g}_{\mu\nu} = 8\pi T_{\mu\nu}[\bar{g}, \phi], \tag{1-1a}$$

$$T_{\mu\nu} = \bar{\nabla}_\mu\phi\bar{\nabla}_\nu\phi - \tfrac{1}{2}\bar{g}_{\mu\nu}\bar{\nabla}^\alpha\phi\bar{\nabla}_\alpha\phi, \tag{1-1b}$$

$$\Box_{\bar{g}}\phi = 0, \tag{1-1c}$$

MSC2020: primary 83C75; secondary 35B35, 35Q76, 83C05, 83F05.
Keywords: Einstein scalar-field system, stability, blow-up profile, cosmic censorship, big bang singularity.

with initial data $(g_0, k_0, \pi_0, \psi_0)$ on a closed 3-manifold M that admits a negative Einstein metric γ.[1] In this paper, we determine the maximal globally hyperbolic development emanating from such initial data given that it is sufficiently close to the initial data of a homogeneous solution with a nontrivial scalar field.

In the collapsing direction, we prove a stable big bang formation and curvature blow-up result, which requires the presence of a nontrivial scalar field. The results complement those in [Rodnianski and Speck 2018b; Speck 2018], which cover flat and spherical spatial geometry. In the expanding direction, we prove a nonlinear future stability result of the corresponding vacuum background solution, which is the Milne model, under a mild condition for the first positive eigenvalue of $-\Delta_\gamma$ (see Definition 9.2). As discussed in more detail in Remark 9.3, numerical studies (see [Cornish and Spergel 1999; Inoue 2001]) show that this condition holds for an analogue of Weeks space, and suggest that this may hold for all closed hyperbolic 3-manifolds with sectional curvature $-\frac{1}{9}$.

Connecting the two regions, we prove the global stability (i.e., past and future stability) of the spacetime

$$([0, \infty) \times M, -dt^2 + a(t)^2\gamma), \tag{1-2a}$$

given a negative Einstein manifold (M, γ) obeying the aforementioned spectral condition, with

$$a(0) = 0, \quad \dot{a} = \sqrt{\tfrac{1}{9} + \tfrac{4\pi}{3}C^2 a^{-4}} \tag{1-2b}$$

for some given constant $C > 0$, and the scalar field given by

$$\partial_t \phi = Ca^{-3}, \quad \nabla\phi = 0. \tag{1-2c}$$

The scale factor consequently exhibits the following asymptotic behaviour:

$$a(t) \simeq t^{1/3} \text{ as } t \searrow 0 \quad \text{and} \quad a(t) \simeq t \text{ as } t \nearrow \infty. \tag{1-2d}$$

The main result can be split into two parts:

Theorem 1.1 (big bang stability: rough version). *Let $(M, g_0, k_0, \pi_0, \psi_0)$ be initial data for the Einstein scalar-field system that is sufficiently close to $(M, a(t_0)^2\gamma, -\dot{a}(t_0)a(t_0)\gamma, 0, Ca(t_0)^{-3})$, where $C > 0$ and (M, γ) is a closed Riemannian 3-manifold with $\mathrm{Ric}[\gamma] = -\frac{2}{9}\gamma$ $\big($i.e., a closed negative Einstein manifold with scalar curvature $-\frac{2}{3}\big)$.*

Then, the past maximal globally hyperbolic development $((0, t_0] \times M, \bar{g}, \phi)$ of the initial data within the Einstein scalar-field system (1-1a)–(1-1c) *admits a foliation by CMC hypersurfaces $\Sigma_s = t^{-1}(\{s\})$ with zero shift. This development remains close to the FLRW solution described in* (1-2a)–(1-2c) *in the past of the initial data slice Σ_{t_0}. In particular, the solution exhibits curvature blow-up of order t^{-4} and every causal geodesic becomes incomplete as t approaches 0.*

Theorem 1.2 (global stability). *Let $(M, g_0, k_0, \pi_0, \psi_0)$ be initial data as in Theorem 1.1. In addition, we suppose that the smallest positive eigenvalue of $-\Delta_\gamma$ acting on scalar functions is strictly greater than $\frac{1}{9}$.*

Then, the initial data admits a maximal globally hyperbolic development $((0, \infty) \times M, \bar{g}, \phi)$ solving the Einstein scalar-field system that, in addition to the results of Theorem 1.1, is future (causally) complete. As

[1] Here and throughout, π_0 and ψ_0 prescribe data for $\nabla\phi|_{\Sigma_{t_0}}$ and $\partial_0\phi|_{\Sigma_{t_0}}$ respectively.

$t \nearrow \infty$, the solution is attracted by Milne spacetime in the sense that the expansion normalized variables $(\boldsymbol{g}, \boldsymbol{k}, \nabla\phi, \phi')$ (see Definition 9.4) converge toward $\big(\gamma, \frac{1}{3}\gamma, 0, 0\big)$.

A more detailed statement of Theorem 1.1 is provided in Theorem 8.2. The additional spectral condition in Theorem 1.2 is discussed at the end of Section 1.3, and the statement itself is proven in Section 10 to be an extension of the Milne future stability result in Theorem 9.1.

1.2. ***Background material.*** We now provide context for the previously discussed setting and the results in Theorems 1.1–1.2:

1.2.1. *Initial data to the Einstein scalar-field equations.* It is well known that the Einstein equations can, via the 3+1 decomposition, be viewed as an elliptic-hyperbolic system of PDEs (see, for example, [Andersson and Moncrief 2003]). This reduces solving the Einstein equations to two problems: finding admissible Einstein initial data in physical space, and then solving the corresponding initial value problem. Regarding the former, initial data to the Einstein scalar-field system takes the form

$$(M, \mathring{g}, \mathring{k}, \mathring{\pi}, \mathring{\psi}),$$

where $\mathring{g}$ and $\mathring{k}$ are symmetric (0, 2)-tensors on M, $\mathring{\pi}$ is a (0, 1)-tensor (corresponding to $\nabla\phi$) and $\mathring{\psi}$ is a scalar function (corresponding to the future-directed normal derivative $\partial_0\phi$ of the scalar field). The initial data must satisfy the Hamiltonian and momentum constraints

$$R[\mathring{g}] + (\mathring{k}^a{}_a)^2 - (\mathring{k}^a{}_b\mathring{k}^b{}_a) = 8\pi[|\mathring{\psi}|^2 + |\mathring{\pi}|^2_{\mathring{g}}], \tag{1-3a}$$

$$\operatorname{div}_{\mathring{g}} \mathring{k} = -8\pi \cdot \mathring{\pi} \cdot \mathring{\psi} \tag{1-3b}$$

(see (2-16a) and (2-16b)), where the indices of $\mathring{k}$ in the first line are raised with respect to $\mathring{g}$.

We note that, in our argument, we will additionally assume that our initial data has constant mean curvature so that our gauges can be satisfied initially — this is enforced on the level of initial data by requiring

$$\operatorname{tr}_{\mathring{g}} \mathring{k} = -3\frac{\dot{a}(t_0)}{a(t_0)}$$

(see (2-10)). We will argue in Remark 8.1 why the initial data being near-FLRW allows us to assume the initial hypersurface to be CMC without loss of generality.

The results of [Fourès-Bruhat 1952; Choquet-Bruhat and Geroch 1969] show that there exists an embedding[2] $\iota : M \hookrightarrow \iota(M) \subset \overline{M}$ and a maximal solution $(\overline{M}, \bar{g}, \nabla\phi, \partial_0\phi)$ to the Einstein scalar-field equations such that $\iota(M) = \Sigma_{t_0}$ is a Cauchy hypersurface and such that

$$\iota^*\bar{g} = \mathring{g}, \quad \iota^*k = \mathring{k}, \quad \iota^*\pi = \mathring{\pi} \quad \text{and} \quad \iota^*\partial_0\phi = \mathring{\psi}_0.$$

We will perturb around initial data corresponding to data for an FLRW spacetime at time $t = t_0$, i.e.,

$$(M \cong \Sigma_{t_0}, a(t_0)^2\gamma, -\dot{a}(t_0)a(t_0)\gamma, 0, Ca(t_0)^{-3}).$$

[2]We usually ignore the embedding in notation.

Furthermore, the maximal globally hyperbolic development (MGHD) is unique (up to diffeomorphism), and thus we can assume $(\overline{M}, \bar{g}, \nabla\phi, \partial_0\phi)$ to be globally hyperbolic. However, these statements provide little information on the properties of the MGHD in the future and past of the initial data slice.

1.2.2. *Strong cosmic censorship.* In their groundbreaking papers on singularity theorems, Hawking [1967] and Penrose [1965] established very general criteria for the MGHD of spacetimes to become causally geodesically incomplete. Many spacetimes of physical relevance satisfy these criteria, including the spacetimes considered in this article. While giving us more information on the MGHD than the existence and uniqueness results mentioned above, a key issue in the application of this mathematical result to general relativity is that no statement is made on how precisely the singularity comes about: In particular, such incompleteness (within a given regularity class) could either mean that the geodesic is inextendible — which must be caused by the blow-up of some geometric quantity — or that there exist multiple inequivalent extensions. While the latter behaviour is exhibited even for some cosmological spacetimes (see, for example, the Taub solutions discussed in [Chruściel and Isenberg 1993]), such behaviour is usually considered to be unphysical since it would imply a breakdown of determinism. The *strong cosmic censorship conjecture* (SCCC) posits in its most general form that, for generic solutions to the Einstein equations, this incompleteness instead manifests as inextendibility at a given level of regularity (e.g., C^0, C^2, C^∞, ...).

In certain frameworks in the homogeneous cosmological setting — i.e., for homogeneous initial data on a closed spatial hypersurface — it was shown in fundamental works [Chruściel and Rendall 1995; Ringström 2009] that the so-called Kretschmann scalar $R_{\alpha\beta\gamma\delta}R^{\alpha\beta\gamma\delta}$ is unbounded where incompleteness manifests. Thus, it is the driving force behind geodesic incompleteness in these cases, forcing C^2-inextendibility of the MGHD. For the purposes of analyzing cosmologically relevant spacetimes, the SCCC is hence often rephrased as follows:

Conjecture 1.3 (cosmological SCCC; see, e.g., [Ringström 2009, Chapter 17]). *For generic initial data, the Kretschmann scalar is unbounded where causal geodesics become incomplete.*

Theorem 1.1, in short, shows that this conjecture is rigorously supported in the case of FLRW spacetimes with negative spatial curvature. More precisely, the past asymptotics of such spacetimes, determined by initial data on Σ_{t_0} as discussed above, are generic in the following sense: There exists an open neighbourhood of said FLRW data within the set of Einstein scalar-field initial data such that the solutions past-directed causal geodesics become incomplete, and the incompleteness is driven by blow-up of Kretschmann scalar with the same asymptotics as the FLRW solution. The global result in Theorem 1.2 portrays the other side of cosmic censorship — as with the past evolution, near-FLRW data fully determines the future of the spacetime in the sense that the MGHD is future complete, again showing that this feature of FLRW spacetimes with negative spatial sectional curvature is generic.

1.2.3. *FLRW and generalized Kasner spacetimes with scalar fields.* On a large scale, the universe is often viewed as spatially homogeneous and isotropic, i.e., no point in space and no direction are distinguishable from any other point and direction (referred to as the "cosmological principle"). In 1935, it was shown by Robertson and Walker that, under a few very natural additional assumptions, this restricts the class of

potential spacetimes to the FLRW class

$$(I \times \widetilde{M},\ \tilde{g}_{\mathrm{FLRW}} = -dt^2 + a(t)^2\tilde{\gamma}),$$

where $(\widetilde{M}, \tilde{\gamma})$ is a manifold of constant sectional curvature κ and where the scale factor a depends smoothly on t. This holds before taking the Einstein equations into consideration — when doing so, the matter model determines how space expands within the cosmological model via a. We refer to Lemma 2.3 for the scalar-field solution for $\kappa = -\frac{1}{9}$, but note that the scale factor behaves like $t^{1/3}$ for scalar-field matter, regardless of spatial geometry, and that the Kretschmann scalar blows up at order $\mathcal{O}(t^{-4})$ toward the big bang ($t \downarrow 0$).

Spatially flat FLRW spacetimes are a subclass of the closely related *generalized Kasner spacetimes*, which are still spatially homogeneous but anisotropic in general. For scalar field matter, the spacetime metric is given by

$$\bar{g}_{\mathrm{Kasner}} = -dt^2 + \sum_{i=1}^{D} t^{2p_i}\, dx^i \otimes dx^i, \quad \sum_{i=1}^{D} p_i = 1,\ \sum_{i=1}^{D} p_i^2 = 1 - 8\pi A^2, \quad \bar{\phi}_{\mathrm{Kasner}}(t) = A\log(t).$$

The standard Kasner family is obtained by considering the vacuum case ($A = 0$), and the spatially flat FLRW spacetime by setting $D = 3,\ p_i = \frac{1}{3},\ A = \sqrt{\frac{1}{12\pi}}$. If more than one of the Kasner exponents is nonzero, the generalized Kasner family satisfies the SCCC, also by exhibiting Kretschmann scalar blow-up of order t^{-4}as $t \downarrow 0$ (see [Rodnianski and Speck 2018b, (1.8)]).

Kasner spacetimes are of particular relevance to cosmology due to their relationship with the *BKL conjecture*: Heuristically, this conjecture states that the dynamics of cosmological spacetimes near a spacelike singularity generically exhibit chaotic and highly oscillatory behaviour, often referred to as "mixmaster" behaviour. This behaviour is driven by velocity terms within the Einstein equations and is locally comparable to that of (vacuum) Kasner solutions. However, even if the BKL picture is to be believed in general, scalar-field (or, more generally, stiff-fluid) solutions seem to form an exception to it: They have a dampening effect on said oscillations, thus generating big bang stability as shown rigorously in [Rodnianski and Speck 2018b; Fournodavlos et al. 2023] for Kasner spacetimes (for more details, see Section 1.3). This scenario, often referred to as *quiescent cosmology*, was studied in, for example, [Belinskiĭ and Khalatnikov 1973; Barrow 1978; Andersson and Rendall 2001]. With this in mind, both the aforementioned Kasner results and the results within this article, along with the prior FLRW results [Rodnianski and Speck 2018b; Speck 2018], confirm this quiescent effect of scalar fields in cosmology.

We note that one can view this as a scalar field ensuring a specific scenario in the very early universe given a class of initial data, namely matching the asymptotic behaviour of the big bang singularity. This fits into the recent use of nonlinear scalar fields in string cosmology, where specific choices of field are made to specific behaviours (e.g., inflation) in the early universe. For a recent review, we refer to [Cicoli et al. 2024].

1.3. ***Relation to previous work.*** Theorem 1.2 is the first theorem about the full global structure of FLRW spacetimes with negatively curved spatial geometry. For such solutions, prior results exclusively concern future stability, which we further discuss below. Besides [Speck 2018] covering the $\mathbb{S}^3$-case, it is the only open set of initial data for cosmological spacetimes (i.e., without symmetry assumptions) with $\Lambda = 0$

and in absence of accelerated expansion for which the global (future and past) dynamics are now fully understood.[3]

Scalar field matter (and, more generally, matter obeying semilinear wave equations or fluid matter) and their asymptotic behaviour on fixed cosmological backgrounds have been studied extensively, for example in [Allen and Rendall 2010; Alho et al. 2019; Bachelot 2019; Beyer and Oliynyk 2024b; Ringström 2019; 2020; 2021; Wang 2021]. While many of the results, in particular [Ringström 2020], manage to analyze very general classes of equations and spacetime geometries, including the wave equation on the FLRW backgrounds studied in [Alho et al. 2019; Fajman and Urban 2022], the methods used are often difficult to apply to the full Einstein scalar-field system. In [Fajman and Urban 2022], we extended the approach of [Alho et al. 2019] to be able to deal with various warped product spacetimes, and in particular FLRW spacetimes with negatively curved spatial geometry, by using the spatial Laplace operator to control high-order derivatives. The perturbation-adapted analogue of this strategy is at the basis of the energy method in this paper.

We also note that, by the results of [Girão et al. 2019], there are nontrivial waves on fixed FLRW backgrounds that converge toward the big bang singularity, even if, as demonstrated in [Alho et al. 2019; Fajman and Urban 2022], this behaviour is nongeneric. Such waves can give rise to convergent asymptotics on cosmological backgrounds, as studied in [Ringström 2020]. Thus, it will likely be difficult to replace (1-2c) with an arbitrary nontrivial reference wave, while keeping past stability intact. However, by restricting to an open neighbourhood near the solution described in (1-2a)–(1-2c), potential nongeneric solutions of this type are excluded. For the more general conditions on initial data that lead to quiescent asymptotics, we refer to [Oude Groeniger et al. 2023], which will be discussed further below.

Theorem 1.1 forms the counterpart to the pioneering works [Rodnianski and Speck 2018a; 2018b; Speck 2018], which cover nonlinear big bang stability for FLRW spacetimes with spatial geometry $\mathbb{T}^3$ and $\mathbb{S}^3$ respectively. These results were extended to Kasner spacetimes in [Rodnianski and Speck 2022] with $|q_i| < \frac{1}{6}$, and to the full subcritical regime in [Fournodavlos et al. 2023], i.e., (generalized) Kasner spacetimes as discussed in Section 1.2.3 with $\max_{i,j,k=1,\dots,D}(p_i + p_j - p_k) < 1$. The former necessitates considering $(1+D)$-dimensional Kasner spacetimes with $D \geq 38$, while the latter result also can be satisfied in $D = 3$ for generalized Kasner spacetimes. Recall that this means, in contrast to our setting, that the reference spacetime can be anisotropic, even if the conditions on Kasner exponents rule out extremely anisotropic regimes. As a result, the analysis therein becomes significantly more involved, especially at top order, since approximately monotonic energy identities as used in our work, as well as in [Rodnianski and Speck 2018b; Speck 2018], have not been found in these anisotropic settings.

We note that the argument in [Fournodavlos et al. 2023] relies on identifying an almost-diagonal structure for the asymptotics of (combined) connection coefficients for an adapted frame that is carried along by Fermi–Walker transport; this is precisely where subcriticality enters. Given that these no longer can vanish in a reference frame adapted to near-hyperbolic spatial geometry, it is a priori unclear whether this structure is sufficiently maintained.

[3]For a related future stability result in accelerated expansion, see [Ringström 2008], which considers scalar fields with a nontrivial potential.

The impressive recent preprint of Oude Groeniger, Petersen and Ringström [Oude Groeniger et al. 2023] circumvents this issue and uses the equations considered in [Fournodavlos et al. 2023] to establish general conditions for initial data to the Einstein (nonlinear) scalar-field equations to give rise to quiescent singularities (see [Oude Groeniger et al. 2023, Theorem 12]). Additionally, they show that a large class of cosmological model solutions exhibit stable big bang formation (see [Oude Groeniger et al. 2023, Theorem 49]). In particular, by only requiring that the mean curvature is sufficiently large compared to the expansion-normalized data, the rescaled connection coefficients can be made to be sufficiently small even if they are nontrivial in the reference. However, this high level of generality comes at the cost of no longer being able to ensure that the expansion-normalized solution variables themselves, in particular the generalized Kasner exponents, remain close to the reference solution, in contrast to our asymptotic results in Theorem 8.2.

Furthermore, Beyer and Oliynyk [2024a] have recently shown that, over $\mathbb{T}^3$, the big bang formation can be localized in the sense that data given solely on a ball within the initial hypersurface must also cause stable blow-up on a (smaller) ball on the big bang hypersurface. While this result further indicates that blow-up behaviour of near-FLRW spacetimes might be, at least, independent of global geometric properties as it seems to be a localizable, we note that proof of localized stability crucially relies on the flatness of the conformal reference spacetime. To be more precise, the proof relies on extending the local initial data to global data for a Fuchsian system of metric and matter quantities as well as, again, connection coefficients for an adapted, Fermi–Walker transported frame. However, the derivation of the system for the former explicitly seems to use flat spatial geometry to obtain the necessary Fuchsian form. This form seems to similarly be broken as soon as the connection coefficients are not perturbed around 0, since this would lead to inhomogeneous error terms of order t^{-1} for the rescaled variables which are stronger than what the method, so far, accounts for.

By contrast, in [Rodnianski and Speck 2018b; Speck 2018], the reference frame itself is used in the commutator method to obtain the necessary energy identities at high orders. In all of these works, it hence is a priori unclear how one could extend these methods to the negative spatial Einstein geometry of (M, γ). We provide an alternative approach that, besides establishing the complementary stability result to [Rodnianski and Speck 2018b; Speck 2018], does not rely on any information on the spatial geometry of the reference manifold in its methodology (although it is of course relevant in determining the FLRW reference solution that we are studying). Instead, we rely on differential operators adapted to the evolved spatial metric. Hence, we believe that our approach may also prove useful for stability problems in spatially inhomogeneous (and hence also anisotropic) settings. In light of [Rodnianski and Speck 2022; Fournodavlos et al. 2023] in particular, the main challenge in achieving this would either be to find approximately monotonic energy identities with our Bel–Robinson approach that have not been observed previously, or to also find ways to circumvent the lack thereof.

To obtain Theorem 1.1, we use the Laplace–Beltrami operator (acting, respectively, on scalar functions and tensor fields) with respect to the (rescaled) evolved metric as our commutating operator instead of a fixed reference frame. This, in turn, leads us to replacing the wave-like system for metric and second fundamental form exploited in [Rodnianski and Speck 2018b; Speck 2018] by an evolutionary system

in the second fundamental form and Bel–Robinson variables. The latter technique dates back to the fundamental works [Christodoulou and Klainerman 1990; 1993], where it was used to analyse field equations on Minkowski space and then to show global stability of Minkowski space itself. It has also been applied to the future stability of Milne spacetimes in the vacuum Einstein equations in [Andersson and Moncrief 2004] and, more recently, within the massive Einstein–Klein–Gordon system in [Wang 2019]. As far as we are aware, this method has not yet been applied to solutions that are not near-vacuum or in the context of big bang singularity formation.

Toward the big bang, the solutions exhibit asymptotically velocity dominated (AVTD) behaviour in the sense that they behave, to leading order, like solutions to the Einstein scalar-field equations in CMC gauge with zero shift with all terms involving spatial derivatives set to zero (the "velocity term dominated" (VTD) equations). This behaviour also matches results obtained by studying high-regularity solutions (e.g., [Andersson and Rendall 2001]), or related works using Fuchsian methods that prescribe a behaviour at the singularity and then develop it locally, often under additional symmetry assumptions (e.g., [Damour et al. 2002; Choquet-Bruhat et al. 2004; Isenberg and Moncrief 2002; Fournodavlos and Luk 2023]). In particular, this asymptotic behaviour leads to the same types of "Kasner footprint states" as in [Rodnianski and Speck 2018a; 2018b]: As one approaches the big bang, the rescaled variables converge toward tensor fields on the big bang hypersurface that precisely solve the truncated VTD equations. Further, the distance between the footprints of the FLRW and the perturbed solution are controlled by the initial data. For example, the rescaled Weingarten map $a^3 k^a{}_b$ converges to $(K_{\mathrm{Bang}})^a{}_b$ on the big bang hypersurface, which is close to $\frac{\sqrt{4\pi}}{3} C\!\!\!/{}^a_b$, the rescaled FLRW footprint (see (8-3e) and (8-5c)).

What remains to be considered to obtain Theorem 1.2 is future stability, which we can reduce to future stability of the vacuum solution in the Einstein scalar-field system. This solution, called the Milne spacetime, has been shown to be stable within the set of vacuum solutions — see [Andersson and Moncrief 2011] — and a range of other Einstein systems — see, for example, [Wang 2019; Andersson and Fajman 2020; Fajman and Wyatt 2021; Fajman et al. 2024; Barzegar and Fajman 2022; Branding et al. 2019] and related work in lower dimensions, e.g., [Andersson et al. 1997; Moncrief 2008; Fajman 2017; 2020; Mondal 2023]. As such, our contribution to the study of future stability of Milne spacetimes is that we deal with the massless scalar field matter via corrected energy estimates which are inspired by work in [Choquet-Bruhat and Moncrief 2001] for vacuum Einstein equations with $U(1)$-symmetry. Out of the works listed above, only [Wang 2019; Fajman and Wyatt 2021] deal with scalar field matter at all, namely the massive case. These fields exhibit stronger decay toward the future, making the matter components easier to deal with than in our analysis.

The additional spectral condition is needed to ensure coercivity of the corrected scalar field energy. Numerical work, e.g., [Cornish and Spergel 1999; Inoue 2001], does not suggest that this condition is violated by any closed 3-manifold with constant sectional curvature $\kappa = -\frac{1}{9}$, and verifies that it is satisfied, for example, by an analogue of Weeks space in which the metric is appropriately scaled to have the required sectional curvature. The latter is also verified by the recent result [Bonifacio et al. 2025] that, amongst considering more general related settings, sufficiently constrains the spectrum of the Laplacian on Weeks space. We refer to Remark 9.3 where this is discussed in more detail.

1.4. ***Challenges in the proof.*** The contracting and expanding regimes of near-FLRW spacetime are analyzed in two separate and methodologically independent parts. Before providing an overview of both arguments, we summarize the challenges that arise:

1.4.1. *Big bang stability.* The main difficulties in establishing big bang stability are three-fold:

Firstly, we have to expect that the solutions are asymptotically velocity term dominated (as argued in Remark 8.3, we end up proving that this is the case), and thus that rescaled variables at best exhibit the same asymptotic behaviour as their counterparts in FLRW spacetime, up to a small perturbation in the asymptotic footprint. For example, note that, in the reference FLRW spacetime, one has

$$(k_{\mathrm{FLRW}})^i{}_j = -3\frac{\dot{a}}{a}\mathbb{I}^i_j \approx -\frac{1}{t}\mathbb{I}^i_j.$$

At best, the shear $\hat{k}^i_j$ of the perturbed solution then behaves like ε/t. In fact, we show that this is the case in (4-2b). This implies that the contraction rescaled metric $G_{ij} = a^{-2}g_{ij}$ can only be controlled up to $\mathcal{O}(t^{-c\sqrt{\varepsilon}})$ (see (4-4c)), since one has $\partial_t g_{ij} \approx -2g_{il}k^l{}_j$ and thus

$$\partial_t G_{ij} \approx G_{il}\hat{k}^i{}_j \approx \frac{\varepsilon}{t} * G.$$

However, to be able to use the structure of the evolution equations to cancel terms in our energy arguments, we have to work with adapted quantities. For example, we need to use integration by parts with respect to (Σ_t, G_t) to cancel high-order scalar field terms with help of the (rescaled) wave equation that contains Δ_G, or to obtain elliptic estimates from the lapse equation via the operator Δ_G or from the adapted div-curl-system for Σ arising from the constraint equations.

As a result, even the rescaled solution variables will diverge at order $\mathcal{O}(t^{-c\sqrt{\varepsilon}})$ toward the singularity, so we need to track and control their rate of divergence within the bootstrap argument. This significantly complicates dealing with nonlinear terms, where the bootstrap assumptions often cannot be inserted naively. This in turn makes coercivity of the energies more involved to establish (see Lemma 4.5 and Remark 4.6), since this only holds up to curvature errors that also diverge and thus need to be carefully tracked.

Secondly, and in contrast to [Rodnianski and Speck 2018b; Speck 2018], replacing the wave structure of the geometric evolution in the Einstein equations with our less geometry dependent Bel–Robinson framework seems to lose regularity at first glance: The energy estimates for the evolution system for the scalar field energy and the geometric energies can be caricatured as follows:

$$-\frac{d}{dt}\mathcal{E}^{(L)}(\phi,\cdot) \lesssim \frac{\varepsilon^{1/8}}{t}[\mathcal{E}^{(L)}(\phi,\cdot)+\mathcal{E}^{(L)}(\Sigma,\cdot)]+\cdots,$$

$$-\frac{d}{dt}[\mathcal{E}^{(L)}(\Sigma,\cdot)+\mathcal{E}^{(L)}(W,\cdot)]+\cdots \lesssim \frac{\varepsilon^{1/8}}{t}[\mathcal{E}^{(L)}(\Sigma,\cdot)+\mathcal{E}^{(L)}(W,\cdot)]+\frac{\varepsilon^{-1/8}}{t}\cdot a^4\mathcal{E}^{(L+1)}(\phi,\cdot)+\cdots.$$

Herein, the superscript refers to the order of derivatives, while $\mathcal{E}^{(L)}(\phi,\cdot)$, $\mathcal{E}^{(L)}(\Sigma,\cdot)$ and $\mathcal{E}^{(L)}(W,\cdot)$ refer to energies for the scalar field, the rescaled tracefree part Σ of the second fundamental form and the Bel–Robinson variables respectively. Thus, it seems that we lose derivatives in the scalar field and are not able to close the argument. This is remedied using the div-curl-system in Σ, see (2-36a) and (2-36b),

which yields a weak estimate of the form

$$a^4\mathcal{E}^{(L+1)}(\Sigma,\cdot)\lesssim\mathcal{E}^{(L)}(\phi,\cdot)+\mathcal{E}^{(L)}(W,\cdot)+\mathcal{E}^{(L)}(\Sigma,\cdot)+\cdots.$$

Combining these estimates to improve the bootstrap assumptions then necessitates an intricately constructed total energy to balance these different types of estimates against one another.

Finally, given (1-2c), the rescaled time derivative of the scalar field is not small and does not become so toward the big bang. This leads to various terms within the core linearized evolutionary system of both matter and geometry that, if estimated naively, could lead to exponential blow-up toward the singularity. When such terms occur in the scalar field energy evolution, this can be dealt with along similar lines as in [Rodnianski and Speck 2018b; Speck 2018], but we incur additional large terms in our geometric evolution that only cancel using the explicit form of the Friedman equations, which we highlight in Lemma 7.1 and its proof.

1.4.2. *Future and global stability.* For Milne stability, the canonical Sobolev energies for the scalar field variables, i.e.,

$$\int_M |\phi'|_g^2+|\nabla\phi|_g^2\,\mathrm{vol}_g$$

and higher-order analogues, do not obey useful energy estimates. This can be overcome by adding an indefinite correction term of the type

$$\int_M \phi'(\phi-\bar{\phi})\mathrm{vol}_g$$

to the canonical energy; see Definition 9.6. This is similar to what was done in [Choquet-Bruhat and Moncrief 2001] in a 2+1-dimensional setting, as well as similar to the indefinite terms we introduce in our geometric energy to control the wave system in the metric variables, as in previous work on Milne stability in different matter models, including [Andersson and Fajman 2020; Fajman and Wyatt 2021]. That this corrected energy controls Sobolev norms relies on the aforementioned spectral condition. As a result, and unlike for past stability, the specific spatial geometry is crucial in generating decay from energy estimates, even before considering the geometric evolution.

Moreover, we need to transition from the near-FLRW data used to analyze the contracting regime to data in the expanding regime on a distant enough future hypersurface such that it is near-Milne and the future stability result applies. This requires a gauge switch from CMC gauge with zero shift to CMCSH gauge, as well as careful control of the solution variables over a finite time interval using continuous dependence on initial data. For the former, close inspection of [Fajman and Kröncke 2020] gives us a diffeomorphism close to the identity that maps the initial data for the metric to new data satisfying the spatially harmonic gauge condition, thus allowing us to switch gauges without losing proximity to the reference solution. This is discussed in detail in Section 10.

1.5. *Proof outline.*

1.5.1. *Big bang stability.*

***The big picture*:** The key argument in our big bang stability proof is a hierarchized series of energy estimates that establishes the asymptotic behaviour of solution variables toward the singularity. We rely

on a bootstrap argument which establishes that energies $\mathcal{E}^{(L)}$ (see Definition 3.9) for the scalar field, the rescaled shear, the Bel–Robinson variables, the lapse and the curvature at worst only diverge slightly. Here, $0 \leq L \leq 18$ denotes the order of derivatives considered. To this end, we make a bootstrap assumption on the solution norm $\mathcal{C}$ (see Definition 3.6) which controls the distance of these rescaled variables, as well as the metric itself, to their FLRW counterparts in terms of supremum norms with respect to G, where $G = a^{-2}g$ is the rescaled *adapted* spatial metric (see Definition 2.9). We refer to Assumption 3.16 and Remark 3.19 for the detailed bootstrap assumptions and improvements, as well as to Lemma 3.14 for the underpinning local well-posedness result. That this bootstrap argument implies Theorem 1.1 follows from a straightforward adaptation of the arguments in [Rodnianski and Speck 2018b, Theorem 15.1].

We work with evolution-adapted norms even though $G(t, x)$ degenerates toward the big bang singularity. Indeed, since we need to exploit the structure of the evolutionary equations, it is more convenient to have these adapted quantities controlled by the solution norms $\mathcal{H}$ and $\mathcal{C}$ directly instead of having to perform changes of metric at that point. Once the improved energy estimates are shown, a (time-scaled) coercivity notion (see Lemma 4.5 and the proof of Corollary 7.3) and Sobolev embeddings with respect to the reference metric γ then ensure that these improved estimates translate to $\mathcal{H}$ and $\mathcal{C}$. This then closes the bootstrap. To actually achieve this improved energy behaviour, we derive elliptic energy estimates or integral-type estimates that, once suitably combined and scaled, yield the desired improvements by straightforwardly applying the Gronwall lemma. Additionally, note that we assume that the initial data is close to FLRW data not just in $\mathcal{H}$, which contains precisely the norms needed to control $\mathcal{C}$ by Sobolev embedding, but also scaled smallness assumptions at one order higher, contained in the top-order seminorm $\mathcal{H}_{\text{top}}$ (see Assumption 3.10). This is needed to ensure that the top-order energy is small initially, and thus to close the bootstrap.

Scale factor $a(t)$: The precise structure of the Friedman equations (2-3)–(2-4) is crucial not only to control time integral quantities up to the big bang hypersurface (see Lemma 2.4), but also to ensure that certain terms in the evolution that would otherwise cause large divergences contribute with favourable sign (see the arguments in Lemma 6.2, as well as Lemma 7.1). It turns out that the sectional curvature entering the Friedman equations actually is not of key importance to large parts of the big bang stability analysis: The leading-order behaviour of the scale factor toward the big bang singularity is determined via the Friedman equation (1-2b) by the matter term, not the sectional curvature. This indicates that our method might extend to different settings.

***Gauge choice, commutation method and Bel–Robinson variables*:** We commute the resulting elliptic-hyperbolic Einstein system with the Laplace–Beltrami operator Δ_G with respect to the rescaled evolved spatial metric $G(t, x)$ to obtain higher-order energy control. Commuting with this operator has the advantage of leaving many integration-by-parts identities intact. These are needed to provide specific cancellations, e.g., to cancel $\Delta^{L/2+1}\phi$-terms arising from the wave equation when computing $\partial_t \mathcal{E}^{(L)}(\phi, \cdot)$. We also note that the only feature of the adapted metric we use is that it is close to γ, and do not use any further information on the geometry, e.g., by choosing a specific reference frame in our commutation method. Further, we employ CMC gauge with zero shift to avoid badly behaved shift terms (see Remark 1.4).

We still, however, need to deal with the Ricci term in the evolution equation for the second fundamental form. To this end, we consider the Bel–Robinson variables E and B, which are Σ_t-tangent symmetric tracefree $(0, 2)$-tensors and contain all information of the spacetime Weyl tensor $W[\bar{g}]$ (see Section 2.4). Suitably projecting the Gauss–Codazzi equations admits additional constraint equations in terms of E and B that allow us to replace the Ricci tensor at the "cost" of introducing Bel–Robinson energies into the formalism; see (2-24a) and the rescaled version (2-29c). Further, E and B satisfy a Maxwell-type system (see Lemma 2.7) that can be exploited to obtain energy estimates and, as with the other evolution equations, is well adapted to commutation with Δ_G.

***A priori low-order C_G-control*:** By applying the bootstrap assumptions on $\mathcal{C}$ to the evolution equations, we can immediately deduce improved low-order estimates in C^l_G for $l \geq 10$ for the solution variables by inserting them into the respective evolution equations (see Lemma 4.3), as well as via the maximum principle for the lapse (see Lemma 4.1). These usually still diverge slightly, mostly due to the asymptotic behaviour of G. However and crucially to our argument, at order 0, the renormalized time derivative Ψ of the wave, the rescaled tracefree part Σ of the second fundamental form and the rescaled Bel–Robinson variable $\boldsymbol{E}$ are in fact $K\varepsilon$-small in C^0_G on the bootstrap interval (see Lemma 4.2). If these estimates did not hold, it would lead to terms that diverge at order $\mathcal{O}(a^{-3-c\sqrt{\varepsilon}})$ in the differential inequalities, and thus cause exponential energy blow-up of order $\mathcal{O}(e^{a^{-c\sqrt{\varepsilon}}})$ that we could no longer control. This behaviour is closely related to the fact that Ψ and Σ converge toward footprint states on the big bang hypersurface that remain $K\varepsilon$-small (see (8-3c) and (8-3e)), and then pass this convergence on to $|\boldsymbol{E}|_G$ (see (8-8a)).

***Energy estimates and hierarchy*:** The main part of the analysis is establishing various energy estimates.

• For the *lapse* (see Section 5), the relevant estimates are direct results of the elliptic lapse equations (2-30a)–(2-30b). The nonlapse terms on the right-hand side of (2-30a) only diverge slightly toward the big bang, in contrast to the divergence at order a^{-4} in (2-30a), and thus allows one to show that, at lower derivative order, the lapse converges to 1. However, since the right-hand side of (2-30b) contains the scalar curvature of G, this estimate loses derivatives. On the other hand, (2-30a) does not lose derivatives, and the elliptic nature in fact allows one to estimate lapse energies of order $L + 2$ by energies in Σ and the scalar field of order L. This makes it possible to control the higher-order lapse term occurring, for example, in (2-28c), without losing regularity. Conversely, both of these gains in regularity are at the cost of losing powers of a. In short, (2-30b) is needed to establish the asymptotic behaviour of the lapse, and (2-30a) to obtain improved energy bounds as a whole.

• The core *matter* energy estimate (see Lemma 6.2) relies on delicate cancellations when computing the time derivative of $\mathcal{E}^{(L)}(\phi, \cdot)$. While we derive this in a fashion that differs from the energy flux method used in [Speck 2018], the necessary cancellations to arrive at Lemma 6.2 are similar.

• The (rescaled) tracefree component of the *second fundamental form* Σ (see Lemma 6.8) and the (rescaled) *Bel–Robinson variables* $\boldsymbol{E}$ and $\boldsymbol{B}$ (see Lemma 6.6) need to be treated simultaneously to deal with the leading curvature term in the evolution of the former by inserting a constraint equation in which $\boldsymbol{E}$ occurs as the leading term (see (2-36d)). However, the matter terms within the evolution of $\boldsymbol{E}$ and $\boldsymbol{B}$

contain, firstly, terms where we again need very precise estimates to show that they do not contribute large a^{-3}-divergences, and, secondly, matter terms that lose one order of derivative.

This order of regularity can be regained using the momentum constraint equation (2-36a) and its Bel–Robinson counterpart (2-36b) containing $\boldsymbol{B}$, which leads to a div-curl-system for Σ (see Lemma 6.10). This is, again, at the cost of losing powers of a.

• As a result, the *core Gronwall argument* performed in Proposition 7.2 combines energies for the matter variables, Σ and the Bel–Robinson variables, as well as energies for $\mathrm{Ric}[G]$. In particular, the curvature energies are necessary to handle commutation errors within the energy estimates, and improved bounds on them need to be obtained to apply the coercivity results in Lemma 4.5 — else, none of energy improvements would extend to improved Sobolev norm bounds and the bootstrap argument would not close.

As many of the a priori C_G-norm estimates add small additional divergences, it is necessary to perform an induction over derivative orders within this mechanism to deal with lower-order error terms. Since Δ_G is elliptic, it is sufficient to perform this for even orders. Along with energies at order $L \in 2\mathbb{N}_0$, the total energy also includes the energy controlling Σ, as well as the scalar field and curvature energies at order $L+1$, appropriately scaled to account for the degenerate elliptic estimate for Σ from Lemma 6.10. This remedies the derivative loss in the Bel–Robinson energy and allows one to improve the total energy at each order until reaching $L = 18$, at which point the bootstrap argument can be closed.

• Note that *the **metric** itself does not enter the core energy mechanism*. In fact, trying to replace control of the Ricci tensor by control of G is likely too imprecise in dealing with high-order curvature errors. Instead, control of $G - \gamma$ and $\Gamma[G] - \widehat{\Gamma}[\gamma]$ is a consequence of a simple integral energy inequality and the improvements achieved for Σ and matter variables (see Lemma 6.14 and Corollary 7.3). Since we cannot utilize any additional structure in dealing with the metric, we have to construct our argument carefully to allow for the metric control to be weaker than what one gets for the core variables, while still being sufficiently strong to constitute an improvement and allowing to switch between H_G and H_γ (and, respectively, C_G and C_γ) norms.

We also point to Remark 6.1 for a more detailed sketch of how the integral inequalities for the core Gronwall argument are structured and how this leads to the bootstrap improvement for the energies.

1.5.2. *Future stability and connecting the regions.* We follow similar lines as in [Andersson and Fajman 2020; Fajman and Wyatt 2021] to prove that near-FLRW spacetimes in negative spatial geometry are future stable. Since $\partial_t \phi$ decays like $a^{-3} \simeq t^{-3}$ in the reference spacetime, the sectional curvature becomes dominant in the Friedman equations and the scale factor approaches that of Milne spacetime as t approaches ∞. Hence, if one moves sufficiently far to the future, choosing near-FLRW data with a homogeneous scalar field is equivalent to choosing near-vacuum data. Thus, what we prove first in Section 9 is future stability of near-Milne spacetimes under the Einstein scalar-field system. Once this is established, we argue in Section 10 how early near-FLRW initial data evolves to data that is sufficiently close to Milne for large enough times, which is essentially a consequence of the scale factor and the (physical) mean curvature approaching that of Milne, up to a multiplicative constant.

In terms of dealing with geometric and elliptic estimates, we can essentially carry over the results of [Andersson and Fajman 2020], as was also done in [Fajman and Wyatt 2021], by working in CMCSH gauge and verifying that the matter components are indeed only perturbative terms within the geometric evolution.

This leaves only the scalar field to be examined. Here, we introduce corrective terms to the energies (see Definition 9.6) which yield decay estimates for the corrected scalar field energy (see Lemmas 9.16 and 9.17). That these energies are coercive (see Lemmas 9.12 and 9.13) requires the aforementioned lower bound for the first positive eigenvalue of $-\Delta_\gamma$.

Remark 1.4 (Why not use CMCSH gauge to prove big bang stability?). One might consider applying this gauge to big bang stability as well since this is precisely the choice of gauge turning the geometric evolution into a wave-like system in (g, k), which seems simpler than our chosen approach in CMC gauge with zero shift. In particular, this would also not rely on any choice of reference frame, and keep the wave structure of the geometric evolution intact, unlike when using Bel–Robinson variables. However, the issue with this approach lies in the shift equation, which would take the following form for the rescaled shift vector $X = a^3\widetilde{X}$:

$$\Delta_G X^l + \mathrm{Ric}[G]^l_m X^m = -2(N+1)(G^{-1})^{im}(G^{-1})^{jn}\Sigma_{ij}(\Gamma^l_{mn} - \widehat{\Gamma}^l_{mn}) + 2(G^{-1})^{im}\nabla_i X^n(\Gamma^l_{mn} - \widehat{\Gamma}^l_{mn}) + \langle \text{error terms in lapse and matter} \rangle. \quad (1\text{-}4)$$

As a result, the first term has to be expected to diverge at the same rate as the metric, i.e., we expect even low-order norms of $\widetilde{X}$ to behave like $a^{-3-c\sqrt{\varepsilon}}$ at best up to small prefactors. However, computing the time derivative of an integral over $|G-\gamma|^2_G$ (or derivatives thereof) becomes the integral over the $(\partial_t - \mathcal{L}_{\widetilde{X}})$-derivative of this quantity, and hence we get explicit terms of the form $\mathcal{L}_{\widetilde{X}}\gamma$, which always exist at highest order and diverge worse than t^{-1}. In short, the fact that the metric cannot be expected to converge to a footprint state leads to leading-order terms in the differential energy estimates to carry strongly divergent prefactors in CMCSH gauge. This obstructs improvements in a tentative bootstrap argument.

1.6. *Paper outline.*

- Sections 2–8 cover the proof of big bang stability:
  - In Section 2, we introduce notation and provide the necessary information on the FLRW background solution, as well as the equations relevant to the subsequent analysis.
  - Then, in Section 3, we discuss the solution norms and energies and state the initial data and bootstrap assumptions.
  - In Section 4, improved low-order C_G-norm estimates that follow directly from the bootstrap assumptions are established, along with additional formulas and a priori estimates.
  - Section 5 concerns the elliptic estimates for the lapse.
  - In Section 6, we discuss the energy and Sobolev norm estimates for all other variables, all of which are integral estimates except for the aforementioned elliptic estimate for Σ, as well as a norm bound for $\nabla\phi$ that is not needed for the energy improvement.

- These are all combined in Section 7 to improve the bootstrap assumptions — first for the energies, then for $\mathcal{H}$ and finally $\mathcal{C}$.
- In Section 8, we show how this bootstrap argument implies the main big bang stability result (see Theorem 8.2, which is the formal version of Theorem 1.1).

- Section 9 contains the proof of near-Milne future stability.
- In Section 10, we show that this is sufficient for future stability of near-FLRW spacetimes, proving Theorem 1.2.
- Appendices A and B collect various basic formulas and commutator expressions, as well as error terms and how these can be estimated.

2. Big bang stability: preliminaries

2.1. *Notation.*

2.1.1. *Foliations.* On a spacetime manifold $(\overline{M}, \bar{g})$, we assume the existence of a spacelike Cauchy hypersurface Σ_{t_0} that is diffeomorphic to M. As we argue in Remark 8.1, we can assume without loss of generality that it has constant mean curvature. We will ultimately show that there exists a time function t such that the past of $\Sigma_{t_0} = t^{-1}(t_0)$ can be foliated by $\Sigma_s = t^{-1}(s)$ for $s \in (0, t_0)$, and that where the solution exists, this is at least possible up to some $T \in (0, t_0)$. These constant time surfaces are then also spacelike Cauchy hypersurfaces diffeomorphic to M and CMC. We will use this notation throughout with little comment and often simply view Σ_s as $\{s\} \times M$.

2.1.2. *Metrics.* The spacetime metric $\bar{g}$ on $\overline{M}$ takes the general form

$$\bar{g} = -n^2\,dt^2 + g_{ab}\,dx^a\,dx^b,$$

where $n \equiv n(t, x)$ is the lapse function and $g|_{\Sigma_t} \equiv g|_{\Sigma_t}(t, x)$ is a Riemannian metric on Σ_t. We will often simply denote the spatial metric by g. Furthermore, we denote the rescaled spatial metric by $G_{ij} = a^{-2}g_{ij}$ (see Definition 2.9) and the tensor field induced by the matrix inverse of (G_{ij}) by G^{-1}. Similarly, $\det g$ and $\det G$ are also meant as the determinants in the matrix sense. Finally, we define vol_g and μ_g as the volume form and volume element with regard to g, and the same for γ and G.

2.1.3. *Indices and coordinates.* Greek indices $\alpha, \beta, \ldots, \mu, \nu, \ldots$ run from 0 to 3, lowercase Latin indices $a, b, \ldots, i, j, \ldots$ from 1 to 3. The spatial indices on some coordinate neighbourhood $V \subseteq \overline{M}$ are always with regard to the local frame induced by coordinates (x^1, x^2, x^3) on M, applied to each $V \cap \Sigma_t$ by the standard embedding where this intersection is nonempty. The index 0 always denotes components relative to $\partial_0 = n^{-1}\partial_t$, where ∂_t is the derivative associate to the time function t. The Levi-Civita connections associated to $\bar{g}$, respectively g and G, are denoted by $\bar{\nabla}$, respectively ∇.[4] Additionally, for the hyperbolic spatial reference metric γ on M (see Definition 2.1), we write the Levi-Civita connection as $\hat{\nabla}$.

[4]Note that g and G have the same Levi-Civita connection since, on every hypersurface Σ_t, they are related by a scalar multiple.

Whenever we raise or lower Greek (resp. Latin) indices without additional notation, it is with regard to $\bar{g}$ (resp. g). When we raise indices of a tensor $\mathfrak{T}$ with regard to the rescaled spatial metric G, we flag this by writing $\mathfrak{T}^\sharp$. We never raise or lower with respect to γ. Refer to Section 2.1.9 as to how we distinguish taking multiple covariant derivatives from index raising.

2.1.4. *Σ_t-tangent tensors.* For any Σ_t-tangent tensor $\xi^{\alpha_1\cdots\alpha_r}{}_{\beta_1\cdots\beta_s}$, we write $\xi(t)^{a_1\cdots a_r}{}_{b_1\cdots b_r}$ for the $\bar{g}$-orthogonal projection of ξ onto the hypersurface Σ_t. When clear from context, we will drop the time-dependency in notation.

2.1.5. *Sign conventions.* Within this paper, the second fundamental form with regard to Σ_t is defined as the $(0, 2)$-tensor k given by

$$k(X, Y) = -\bar{g}(\bar{\nabla}_X \partial_0, Y),$$

where X and Y are Σ_t-tangent vectors. The Riemann curvature tensor of $\bar{g}$ is taken to be

$$\bar{\nabla}_\alpha \bar{\nabla}_\beta Z_\gamma - \bar{\nabla}_\beta \bar{\nabla}_\alpha Z_\gamma = \mathrm{Riem}[\bar{g}]_{\alpha\beta\gamma}{}^\delta Z_\delta$$

for the covariant vector field (Z_μ), and the analogous convention holds for all other Riemann curvature tensors that appear.

2.1.6. *Constants.* For two nonnegative scalar functions ζ_1, ζ_2, we write $\zeta_1 \lesssim \zeta_2$ if and only if there exists a constant $K > 0$ such that $\zeta_1 \leq K\zeta_2$. This implicit constant may depend on information from the FLRW reference solution at the starting point of the evolution (in particular on γ and $a(t_0)$, see Definition 2.1) and combinatorial quantities. We extend this notation to a real function ζ_1' by

$$\zeta_1' \lesssim \zeta_2 \quad :\Longleftrightarrow \quad \max(\zeta_1', 0) \lesssim \zeta_2.$$

Additionally, we write $\zeta_1 \simeq \zeta_2$ if and only if $\zeta_1 \lesssim \zeta_2 \lesssim \zeta_1$ is satisfied.

2.1.7. *Tensor contractions.* We denote by $\boldsymbol{\varepsilon}_{\alpha\beta\gamma\delta}$ the Levi-Civita tensor with regard to $\bar{g}$ and define Levi-Civita tensor on spatial hypersurfaces Σ_t by $\boldsymbol{\varepsilon}[g]_{ijk} = \boldsymbol{\varepsilon}_{0ijk}$. Notice that this corresponds to the Levi-Civita tensor associated to g. Further, $\boldsymbol{\varepsilon}[G]_{ijk} = a^{-3}\boldsymbol{\varepsilon}[g]_{ijk}$ is the Levi-Civita tensor with respect to the rescaled metric G (see (2-27a)).

For Σ_t-tangent $(0, 2)$-tensors A, $\tilde{A}$ and vector field v, we define the following objects as in [Andersson and Moncrief 2004, Section A.2]:

$$\begin{aligned}
A \cdot \tilde{A} &= A_{ab}\tilde{A}^{ab} = \langle A, \tilde{A}\rangle_g,\\
(A \odot_g \tilde{A})_{ij} &= A_{ik}\tilde{A}^k{}_j,\\
(A \wedge \tilde{A})_i &= \boldsymbol{\varepsilon}_i{}^{jp} A_j{}^q \tilde{A}_{qp},\\
(v \wedge A)_{ab} &= \boldsymbol{\varepsilon}_a{}^{cd} v_c A_{db} + \boldsymbol{\varepsilon}_b{}^{cd} v_c A_{ad},\\
(A \times \tilde{A})_{ij} &= \boldsymbol{\varepsilon}_i{}^{ab}\boldsymbol{\varepsilon}_j{}^{pq} A_{ap}\tilde{A}_{bq} + \tfrac{1}{3}(A \cdot \tilde{A})g_{ij} - \tfrac{1}{3}(\mathrm{tr}_g A \cdot \mathrm{tr}_g \tilde{A})g_{ij},\\
(\mathrm{curl}\, A)_{ij} = (\mathrm{curl}_g A)_{ij} &= \tfrac{1}{2}[\boldsymbol{\varepsilon}_i{}^{cd}\nabla_d A_{cj} + \boldsymbol{\varepsilon}_j{}^{cd}\nabla_d A_{ci}],\\
(\mathrm{div}_g A)_i &= \nabla^b A_{ib}.
\end{aligned}$$

The operations $\odot_G$, $\langle\cdot,\cdot\rangle_G$ and div_G are defined analogously, with all indices raised and lowered by G instead of g. Finally, for two (0, 1)-tensors $\pi, \tilde{\pi}$, we denote their symmetrized product by

$$(\pi \otimes \tilde{\pi})_{ij} = \tfrac{1}{2}(\pi_i \tilde{\pi}_j + \pi_j \tilde{\pi}_i).$$

For pointwise estimates of these quantities, refer to Lemma A.3.

2.1.8. *Schematic term notation.* We will denote as $\mathfrak{T}_1 * \cdots * \mathfrak{T}_l$, where $\mathfrak{T}_i$ are Σ_t-tangent tensors, any multiple of $(\mathfrak{T}_i)$, with regard to the rescaled adapted spatial metric G or as standard multiplication if no summation over indices occurs between factors. Constant prefactors and contractions with regard to G are also suppressed in this notation.

When working with terms where such notation is used, we will estimate these inner products by $\lesssim \prod_{i=1}^{l} |\mathfrak{T}_i|_G$, making any constant in front irrelevant, and further we can view any contraction with regard to G as a product of the noncontracted tensor $\mathfrak{T}$ with G or G^{-1}, and estimate that up to constant by $|G|_G |\mathfrak{T}|_G$, where the first factor is simply $\sqrt{3}$.

For similar products with respect to γ, we denote them by $*_\gamma$.

2.1.9. *On multiple derivatives of variables.* For a scalar function ζ, an (r, s)-tensor field $\mathfrak{T}$ and *capitalized* integers $I, J, \ldots \in \mathbb{N}_0$, we denote by $\nabla^I \zeta$ and $\nabla^I \mathfrak{T}$ the tensors $\nabla_{l_1} \cdots \nabla_{l_I} \zeta$ and $\nabla_{l_1} \cdots \nabla_{l_I} \mathfrak{T}^{i_1 \cdots i_r}{}_{j_1 \cdots j_s}$. We extend this notation to other covariant derivatives analogously. To avoid potential ambiguity with an index raised by g, we will apply the following convention:

- If a covariant derivative carries an uppercase letter, a formula with more than one symbol or a positive integer in its superscript, this refers to taking a derivative of that order.
- If a covariant derivative carries a lowercase letter or 0 in its superscript, this refers to an index.

Further, we will only apply this notation where the precise distribution of indices is not important (e.g., in schematic notation, see Section 2.1.8).

2.2. ***FLRW spacetimes and the Friedman equations.*** Herein, we collect the properties of the reference FLRW solution to the Einstein scalar-field system in CMC-transported coordinates. Our main focus will lie on the behaviour of the scale factor as determined by the Friedman equations. Before moving on to that, we collect the information on the spatial geometry we will need:

Definition 2.1 (hyperbolic reference geometry). (M, γ) is a three-dimensional, connected, closed, orientable Riemannian manifold with constant sectional curvature $-\frac{1}{9}$, and hence Ricci tensor $\mathrm{Ric}[\gamma] = -\frac{2}{9}\gamma_{ij}$ and scalar curvature $R[\gamma] = -\frac{2}{3}$.

Remark 2.2 (Orientability is not a restriction.). We assume that M is orientable for the sake of simplicity. If M should be nonorientable, we may pass the initial data to the oriented double cover and solve the problem there. Since the result is equivariant with respect to the double covering map, this then solves the original problem.

With this in hand, we can express our classical family of solutions to the Einstein scalar-field system as follows:

Lemma 2.3 (FLRW solutions and Friedman equations). *Consider FLRW spacetimes* $(\overline{M}, \bar{g}_{\mathrm{FLRW}})$ *with* $\overline{M} = (0,\infty)\times M$, *where* (M,γ) *is as in Definition 2.1 and where*

$$\bar{g}_{\mathrm{FLRW}} = -dt^2 + a(t)^2\gamma \tag{2-1}$$

holds for some $a \in C^\infty((0,\infty))$, *with the conventions* $a(0) = 0$ *and* $\dot{a}(T) > 0$ *for some arbitrary* $T > 0$. *Further, choose a (smooth) scalar function* ϕ_{FLRW} *such that*

$$\partial_t\phi_{\mathrm{FLRW}} = C\cdot a(t)^{-3}, \quad \nabla\phi_{\mathrm{FLRW}} = 0, \quad \Box_{\bar{g}_{\mathrm{FLRW}}}\phi_{\mathrm{FLRW}} = 0. \tag{2-2}$$

Such a pair $(\bar{g}_{\mathrm{FLRW}}, \phi_{\mathrm{FLRW}})$ *solves the Einstein scalar-field system* (1-1a)–(1-1b) *on* $\overline{M}$ *if and only if a satisfies the Friedman equation*

$$\dot{a} = \sqrt{\tfrac{1}{9} + \tfrac{4\pi}{3}C^2a^{-4}}. \tag{2-3}$$

In particular, one has

$$\ddot{a} = -\tfrac{8\pi}{3}C^2a^{-5}. \tag{2-4}$$

Proof. The first statement follows from explicitly computing $\mathrm{Ric}[\bar{g}]$ as in [O'Neill 1983, p. 345]. That (2-3) implies (2-4) follows simply by computing the derivative of $\dot{a}^2$. □

In the subsequent analysis, the following properties of a that follow from (2-3) will be crucial for our analysis:

Lemma 2.4 (scale factor analysis). *Let a solve* (2-3) *with* $a(0) = 0$. *Then a is analytic on* $(0,\infty)$ *and extends to a continuous function on* $[0,\infty)$ *with* $a(t) \simeq t^{1/3}$ *being satisfied near* $t = 0$. *Further, for any* $p > 0$, *there exist constants* $c > 0$ *and* $K_p > 0$, *where c is independent of p and* K_p *depends analytically on p, such that, for any* $t \in (t, t_0]$, *one has*

$$\exp\Big(p\int_t^{t_0} a(s)^{-3}\,ds\Big) \le K_p a(t)^{-cp} \tag{2-5}$$

and

$$\int_t^{t_0} a(s)^{-3-p}\,ds \lesssim \frac{1}{p}a(t)^{-p}, \quad \int_t^{t_0} a(s)^{-3+p}\,ds \lesssim \frac{1}{p}. \tag{2-6}$$

Moreover, for any $t \in (0, t_0]$ *and any* $q > 0$, *there exist constants* $c > 0$ *and* $K > 0$ *which both are independent of q such that one has*

$$\int_t^{t_0} a(s)^{-3}\,ds \le \frac{K}{q}a(t)^{-cq}. \tag{2-7}$$

Finally, (2-3) *also implies*

$$\sqrt{\tfrac{4\pi}{3}}Ca^{-2} \le \dot{a}. \tag{2-8}$$

Remark 2.5. We will use the estimates in Lemma 2.4 where p is a positive power of ε up to algebraic constants. Then, we can simply replace K_p in (2-5) by a uniform constant.

Proof. For the first statement, we refer to [Fajman and Urban 2022, Lemma 2.1] with $\gamma = 2$. We also collect from there[5] that, for $t < t_0$,

$$\int_t^{t_0} a(s)^{-3}\, ds \lesssim 1 + \left|\log\Big(\frac{t}{t_0}\Big)\right|$$

is satisfied. Hence, there exists some $c' > 0$ such that

$$\exp\Big(p\int_t^{t_0} a(s)^{-3}\, ds\Big) \leq K_p \exp(c' \cdot p \log(t_0)) \cdot \exp(-c' \cdot p) \leq K'_p t^{-c'p}.$$

Then (2-5) follows by applying $a(t) \simeq t^{1/3}$. Noting that $a^{-3} \simeq \dot{a}/a$ holds, one further has

$$\int_t^{t_0} a(s)^{-3-p}\, ds \lesssim \int_{a(t)}^{a(t_0)} y^{-1-p} dy = \frac{1}{p}(a(t)^{-p} - a(t_0)^{-p}) \leq \frac{1}{p} a(t)^{-p}, \tag{2-9}$$

and the other inequality in (2-6) follows analogously. Finally, (2-7) follows directly from (2-6) when assuming without loss of generality that $a|_{(t,t_0)}$ only takes values in $(0, 1)$. □

2.3. ***Solutions to the Einstein scalar-field equations in CMC gauge.*** From here on out, we impose the CMC condition[6]

$$k^l{}_l(t, \cdot) = \tau(t) = -3\frac{\dot{a}(t)}{a(t)}. \tag{2-10}$$

We use (2-3) and (2-4) to collect the following formulas for the mean curvature:

$$\partial_t \tau = 12\pi C^2 a^{-6} + \tfrac{1}{3} a^{-2}, \tag{2-11}$$

$$\tau^2 = 9\frac{\dot{a}^2}{a^2} = 12\pi C^2 a^{-6} + a^{-2}. \tag{2-12}$$

We consequently define the trace-free component $\hat{k}$ of k as

$$\hat{k}_{ij} = k_{ij} - \tfrac{1}{3}\tau g_{ij}, \tag{2-13}$$

and recall that the future-directed unit normal to our foliation is written as

$$\partial_0 = n^{-1}\partial_t. \tag{2-14}$$

With this, we can express the Einstein scalar-field equations in our gauge as follows:

Proposition 2.6 (the Einstein scalar-field system in CMC gauge). *A pair* $(\bar{g}, \phi)$ *solves the Einstein scalar-field equations* (1-1a)–(1-1c) *on* $I \times M$ *in CMC gauge* (2-10) *for some interval* $I \subseteq (0, t_0]$, *where the scale factor satisfies* (2-3), *if and only if the following equations are satisfied on* $I \times M$:

[5] In [Fajman and Urban 2022, Lemma 2.1], one at first only has $\int_t^{t_0} a(s)^{-3}\, ds \lesssim \log(t_0) - \log(t)$ for t_0 small enough that we can estimate $a(t)$ by $t^{1/3}$ up to constant. However, assuming this inequality were to hold up to $\bar{t} > 0$ and one had $t_0 > \bar{t}$, the contribution $\int_{\bar{t}}^{t_0} a(s)^{-3}\, ds$ only adds a constant that we can absorb into our notation. Similarly, $a(t) \simeq t^{1/3}$ can be assumed to hold on $(0, t_0]$ for any fixed $t_0 > 0$, and we can ignore this technicality in proving the integral formulas in Lemma 2.4.

[6] Recall that k is negative in our sign convention; see Section 2.1.5.

*The **metric evolution equations***

$$\partial_t g_{ij} = -2nk_{ij} = -2n\hat{k}_{ij} + 2n\frac{\dot{a}}{a}g_{ij}, \tag{2-15a}$$

$$\partial_t \hat{k}_{ij} = -\nabla_i\nabla_j n + n\Big[\mathrm{Ric}[g]_{ij} - \frac{\dot{a}}{a}\hat{k}_{ij} - 2\hat{k}_{il}\hat{k}^l_j - 8\pi\nabla_i\phi\nabla_j\phi\Big] + 4\pi C^2 a^{-6}(n-1)g_{ij} + \tfrac{1}{9}(3n-1)a^{-2}g_{ij}, \tag{2-15b}$$

*the **Hamiltonian** and **momentum constraint equations***

$$R[g] + \tfrac{2}{3}\tau^2 - \langle\hat{k},\hat{k}\rangle_g = 8\pi[|\partial_0\phi|^2 + |\nabla\phi|^2_g], \tag{2-16a}$$

$$\nabla^l\hat{k}_{lj} = -8\pi\nabla_j\phi\cdot\partial_0\phi, \tag{2-16b}$$

*the **lapse equation***

$$\Delta_g n = -12\pi C^2 a^{-6} - \tfrac{1}{3}a^{-2} + n\big[\tfrac{1}{3}a^{-2} + 4\pi C^2 a^{-6} + \langle\hat{k},\hat{k}\rangle_g + 8\pi|\partial_0\phi|^2\big], \tag{2-17a}$$

or equivalently by (2-16a)

$$\Delta_g n = -12\pi C^2 a^{-6} - \tfrac{1}{3}a^{-2} + n\big[R[g] - 8\pi|\nabla\phi|^2_g + 12\pi C^2 a^{-6} + a^{-2}\big], \tag{2-17b}$$

*and the **wave equation***

$$\Box_{\bar{g}}\phi = -\partial_0^2\phi + n^{-1}g^{ij}\nabla_i n\nabla_j\phi + \Delta_g\phi + \tau\partial_0\phi = 0. \tag{2-18}$$

Proof. These are standard equations that follow from [Rendall 2008, Chapter 2.3] and applying (2-10)–(2-12). □

2.4. ***Bel–Robinson variables.*** In this subsection, we briefly (re-)establish Bel–Robinson variables and how they behave within the Einstein scalar-field system.

Recall that the Weyl tensor $W \equiv W[\bar{g}]$ is the trace-free component of the spacetime curvature and, in the Einstein scalar-field system, takes the form

$$\begin{aligned} W_{\alpha\beta\gamma\delta} &= \mathrm{Riem}[\bar{g}]_{\alpha\beta\gamma\delta} - P[\bar{g}]_{\alpha\beta\gamma\delta},\\ P[\bar{g}]_{\alpha\beta\gamma\delta} &= \tfrac{1}{2}\big(\bar{g}_{\alpha\gamma}\,\mathrm{Ric}[\bar{g}]_{\beta\delta} - \bar{g}_{\beta\gamma}\,\mathrm{Ric}[\bar{g}]_{\alpha\delta} - \bar{g}_{\alpha\delta}\,\mathrm{Ric}[\bar{g}]_{\gamma\beta} + \bar{g}_{\beta\delta}\,\mathrm{Ric}[\bar{g}]_{\alpha\gamma}\big) - \tfrac{1}{6}R[\bar{g}](\bar{g}_{\alpha\gamma}\bar{g}_{\beta\delta} - \bar{g}_{\alpha\delta}\bar{g}_{\beta\gamma})\\ &= 4\pi\big(\bar{g}_{\alpha\gamma}\bar{\nabla}_\beta\phi\bar{\nabla}_\delta\phi - \bar{g}_{\beta\gamma}\bar{\nabla}_\alpha\phi\bar{\nabla}_\delta\phi - \bar{g}_{\alpha\delta}\bar{\nabla}_\beta\phi\bar{\nabla}_\gamma\phi + \bar{g}_{\beta\delta}\bar{\nabla}_\alpha\phi\bar{\nabla}_\gamma\phi\big)\\ &\quad - \tfrac{4\pi}{3}(\bar{\nabla}^\rho\phi\bar{\nabla}_\rho\phi)(\bar{g}_{\alpha\gamma}\bar{g}_{\beta\delta} - \bar{g}_{\alpha\delta}\bar{g}_{\beta\gamma}). \end{aligned}$$

We define the dual W^* of the Weyl tensor as

$$W^*_{\alpha\beta\gamma\delta} = \tfrac{1}{2}\boldsymbol{\varepsilon}_{\alpha\beta\mu\nu}W^{\mu\nu}{}_{\gamma\delta}.$$

The electric and magnetic components of the Weyl tensor, referred to as the Bel–Robinson variables from here on, are now defined as

$$E(W)_{\alpha\beta} = W_{\alpha\mu\beta\nu}\partial_0^\mu\partial_0^\nu = W_{\alpha0\beta0}, \quad B(W)_{\alpha\beta} = W^*_{\alpha\mu\beta\nu}\partial_0^\mu\partial_0^\nu = W^*_{\alpha0\beta0}.$$

We note that, conversely, the Weyl tensor can be fully reconstructed from E and B since the following identities hold:

$$W_{a0c0} = E_{ac}, \quad W_{abc0} = -\boldsymbol{\varepsilon}_{ab}{}^m B_{mc}, \quad W_{abcd} = -\boldsymbol{\varepsilon}_{abi}\boldsymbol{\varepsilon}_{cdj}E^{ij}. \tag{2-19}$$

By the symmetries of the Weyl tensor as a whole, E and B are symmetric and one has $E_{0\beta} = 0 = B_{0\beta}$. Hence, E and B are symmetric, tracefree Σ_t-tangent $(0,2)$-tensors, which we shall simply denote as E_{ij} and B_{ij}.

Further, we define

$$J_{\beta\gamma\delta} := \bar{\nabla}^\alpha W_{\alpha\beta\gamma\delta}, \quad J^*_{\beta\gamma\delta} := \bar{\nabla}^\alpha W^*_{\alpha\beta\gamma\delta}. \tag{2-20}$$

By applying the Bianchi identity for $\mathrm{Riem}[\bar{g}]$, we gain the explicit expression

$$J_{\beta\gamma\delta} = \tfrac{1}{2}(\bar{\nabla}_\gamma \mathrm{Ric}[\bar{g}]_{\delta\beta} - \bar{\nabla}_\delta \mathrm{Ric}[\bar{g}]_{\gamma\beta}) - \tfrac{1}{12}(\bar{g}_{\beta\delta}\bar{\nabla}_\gamma R[\bar{g}] - \bar{g}_{\beta\gamma}\bar{\nabla}_\delta R[\bar{g}]). \tag{2-21}$$

Using (1-1a), we collect

$$J_{i0j} = 4\pi\big[\nabla_i(\partial_0\phi)\nabla_j\phi + k^l{}_i\nabla_l\phi\nabla_j\phi - \partial_0\phi\nabla_i\nabla_j\phi - k_{ij}(\partial_0\phi)^2 - n^{-1}\nabla_i n\cdot\nabla_j\phi\cdot\partial_0\phi\big] - \tfrac{2\pi}{3}[\partial_0(\bar{\nabla}^\alpha\phi\bar{\nabla}_\alpha\phi)]g_{ij}, \tag{2-22}$$

$$J^*_{i0j} = 4\pi\boldsymbol{\varepsilon}_{lmj}(\nabla^l\nabla_i\phi + k^l{}_i\partial_0\phi)\nabla^m\phi + \tfrac{2\pi}{3}\boldsymbol{\varepsilon}_{imj}\nabla^m(\bar{\nabla}^\alpha\phi\bar{\nabla}_\alpha\phi). \tag{2-23}$$

Note that expressions containing $\bar{\nabla}^\alpha\phi\bar{\nabla}_\alpha\phi$ can be ignored throughout our analysis since they are either pure trace or antisymmetric and thus will cancel in inner products with E, B, $\hat{k}$ and their rescaled analogues.

The Bel–Robinson variables then behave as follows:

Lemma 2.7 (constraint and evolution equations for Bel–Robinson variables). *If $(\bar{g},\phi)$ is a classical solution to the Einstein scalar-field system* (1-1a)–(1-1b) *in CMC gauge* (*see* (2-10)), *E and B satisfy the following constraint equations*:

$$E = \mathrm{Ric}[g] + \tfrac{2}{9}\tau^2 g + \tfrac{1}{3}\tau\hat{k} - \hat{k}\odot_g\hat{k} - 4\pi(\nabla\phi\otimes\nabla\phi) - \big(\tfrac{8\pi}{3}|\partial_0\phi|^2 + \tfrac{4\pi}{3}|\nabla\phi|^2_g\big)g, \tag{2-24a}$$

$$B = -\operatorname{curl}\hat{k}. \tag{2-24b}$$

Further, they satisfy the following evolution equations:

$$\partial_t E_{ij} = n\operatorname{curl}B_{ij} - (\nabla n\wedge B)_{ij} - \tfrac{5}{2}n(E\times k)_{ij} - \tfrac{2}{3}n(E\cdot k)g_{ij} - \tfrac{1}{2}\tau n\cdot E_{ij} - \tfrac{1}{2}n(J_{i0j}+J_{j0i}), \tag{2-25a}$$

$$\partial_t B_{ij} = -n\operatorname{curl}E_{ij} + (\nabla n\wedge E)_{ij} - \tfrac{5}{2}n(B\times k)_{ij} - \tfrac{2}{3}n(B\cdot k)g_{ij} - \tfrac{1}{2}\tau n\cdot B_{ij} - \tfrac{1}{2}n(J^*_{i0j}+J^*_{j0i}). \tag{2-25b}$$

Proof. For (2-25a)–(2-25b), we refer to [Andersson and Moncrief 2004, (3.11a)–(3.11b)].[7] Equations (2-24a)–(2-24b) follow as in [Wang 2019, (3.63a)–(3.63b)] from contracting the Gauss–Codazzi constraints. □

Remark 2.8 (initial data for Bel–Robinson variables). Since the Weyl tensor vanishes over FLRW spacetimes, so do $E(W[\bar{g}_{\mathrm{FLRW}}])$ and $B(W[\bar{g}_{\mathrm{FLRW}}])$. Furthermore, note that given initial data $(M, \mathring{g}, \mathring{k}, \mathring{\pi}, \mathring{\psi})$

[7] Note that there is a minor error in the statement in [Andersson and Moncrief 2004], where the authors forget to symmetrize the J-tensors when applying the symmetrization to (3.14) of that work. This error seems to also occur in [Christodoulou and Klainerman 1993, (7.2.2jk)].

on Σ_{t_0} in the sense discussed in Section 1.2.1, and defining $\mathring{\hat{k}} = \mathring{k} - \frac{1}{3}\tau \mathring{g}$, we can use (2-24a) and (2-24b) to define the following (0, 2)-tensors:

$$\mathring{E} = \mathrm{Ric}[\mathring{g}] + \tfrac{2}{9}\tau^2 \mathring{g} + \tfrac{1}{3}\tau \mathring{\hat{k}} - \mathring{\hat{k}} \odot_{\mathring{g}} \mathring{\hat{k}} - 4\pi(\mathring{\pi} \otimes \mathring{\pi}) - \big(\tfrac{8\pi}{3}|\mathring{\psi}|^2_{\mathring{g}} + \tfrac{4\pi}{3}|\mathring{\pi}|^2_{\mathring{g}}\big)\mathring{g}, \tag{2-26a}$$

$$\mathring{B} = -\operatorname{curl}_{\mathring{g}} \mathring{\hat{k}}. \tag{2-26b}$$

These are easily seen to be symmetric, and the constraints (1-3a) and (1-3b) on the initial data ensure that they are also tracefree. Hence, any choice of initial data for the Cauchy problem immediately also contains a unique choice of initial data for the Bel–Robinson variables that is consistent with solutions to the Einstein scalar-field equations in CMC gauge.

2.5. ***Rescaled variables and equations.*** It will be more convenient to work the rescaled and shifted solution variables to measure their distance from the FLRW reference solution. In this subsection, we introduce the renormalized solution variables and restate the Einstein scalar-field system in terms of these variables.

Definition 2.9 (rescaled variables for big bang stability). We will consider the rescaled variables

$$G_{ij} = a^{-2} g_{ij}, \quad (G^{-1})^{ij} = a^2 g^{ij}, \quad \Sigma_{ij} = a\hat{k}_{ij}, \tag{2-27a}$$

$$N = n - 1, \tag{2-27b}$$

$$\boldsymbol{E}_{ij} = a^4 \cdot E_{ij}, \quad \boldsymbol{B}_{ij} = a^4 \cdot B_{ij}, \tag{2-27c}$$

$$\Psi = a^3 \partial_0 \phi - C. \tag{2-27d}$$

We note that the scaling of $\boldsymbol{B}$ in (2-27c) is *not* the asymptotic rescaling of B — in fact, we expect B to have (approximate) leading order a^{-2}, as one can see in (4-4g). However, keeping this scaling parallel to that of $\boldsymbol{E}$ makes the structurally very similar evolution equations significantly easier to deal with. We also do not rescale N asymptotically, unlike [Rodnianski and Speck 2018b; Speck 2018], but note that N converges to 0 at an order slightly below a^4 at low orders (see (3-17h) and (8-3a)).

Proposition 2.10 (the rescaled Einstein scalar-field system). *The Einstein scalar-field system in CMC gauge as in Proposition 2.6 are solved by* $(g, \hat{k}, n, \nabla\phi, \partial_0\phi)$ *if the rescaled variables* $(G, \Sigma, N, \nabla\phi, \Psi, \boldsymbol{E}, \boldsymbol{B})$ *as in Definition 2.9 solve the following set of equations*:[8]

The ***rescaled metric evolution equations***

$$\partial_t G_{ij} = -2(N+1)a^{-3}\Sigma_{ij} + 2N\frac{\dot{a}}{a}G_{ij}, \tag{2-28a}$$

$$\partial_t (G^{-1})^{ij} = 2(N+1)a^{-3}(\Sigma^\sharp)^{ij} - 2N\frac{\dot{a}}{a}(G^{-1})^{ij}, \tag{2-28b}$$

$$\begin{aligned}\partial_t \Sigma_{ij} = {} & -a\nabla_i\nabla_j N + (N+1)\big[a\,\mathrm{Ric}[G]_{ij} - 2a^{-3}(\Sigma \odot_G \Sigma)_{ij} - 8\pi a \nabla_i\phi\nabla_j\phi\big] \\ & + 4\pi C^2 a^{-3}\cdot N G_{ij} + \tfrac{1}{9}(3N+2)aG_{ij} + N\frac{\dot{a}}{a}\Sigma_{ij},\end{aligned} \tag{2-28c}$$

$$\begin{aligned}\partial_t (\Sigma^\sharp)^a_b = {} & \tau N(\Sigma^\sharp)^a_b - a\nabla^{\sharp a}\nabla_b N + (N+1)a\big[(\mathrm{Ric}[G]^\sharp)^a_b + \tfrac{2}{9}\mathbb{I}^a_b\big] \\ & - 8\pi(N+1)a\nabla^{\sharp a}\phi\nabla_b\phi + N\big(4\pi C^2 a^{-3} + \tfrac{1}{9}a\big)\mathbb{I}^a_b,\end{aligned} \tag{2-28d}$$

[8] We refer to Lemma A.3 for the scalings that occur when switching between tensor field operations like $\wedge$ and $\wedge_G$.

the ***rescaled Hamiltonian and momentum constraints***

$$R[G] + \tfrac{2}{3} - a^{-4}\langle \Sigma, \Sigma\rangle_G = 8\pi[a^{-4}\Psi^2 + 2Ca^{-4}\Psi + |\nabla\phi|_G^2], \tag{2-29a}$$

$$\nabla^{\sharp m}\Sigma_{ml} = -8\pi\nabla_l\phi(\Psi + C), \tag{2-29b}$$

with their Bel–Robinson analogues

$$\boldsymbol{E}_{ij} = a^4\big(\mathrm{Ric}[G]_{ij} + \tfrac{2}{9}G_{ij}\big) + \tfrac{1}{3}\tau a^3\Sigma_{ij} - (\Sigma \odot_G \Sigma)_{ij} - 4\pi a^4\nabla_i\phi\nabla_j\phi - \big[\tfrac{4\pi}{3}a^4|\nabla\phi|_G^2 + \tfrac{8\pi}{3}\Psi^2 + \tfrac{16\pi}{3}C\Psi\big]G_{ij}, \tag{2-29c}$$

$$\boldsymbol{B}_{ij} = -a^2\,\mathrm{curl}_G\,\Sigma_{ij}, \tag{2-29d}$$

the ***rescaled lapse equation***

$$\Delta N = \big(12\pi C^2a^{-4} + \tfrac{1}{3}\big)N + (N+1)a^{-4}[\langle\Sigma, \Sigma\rangle_G + 8\pi\Psi^2 + 16\pi C\Psi], \tag{2-30a}$$

or equivalently

$$\Delta N = \big(12\pi C^2a^{-4} + \tfrac{1}{3}\big)N + (N+1)\big[R[G] + \tfrac{2}{3} - 8\pi|\nabla\phi|_G^2\big], \tag{2-30b}$$

the ***rescaled evolution equations for the Bel–Robinson variables***

$$\begin{aligned}\partial_t\boldsymbol{E}_{ij} = {}& (3-N)\frac{\dot a}{a}\boldsymbol{E}_{ij} - a^{-1}(\nabla N \wedge_G \boldsymbol{B})_{ij} + (N+1)a^{-1}\,\mathrm{curl}_G\,\boldsymbol{B}_{ij}\\ & - (N+1)\big[\tfrac{5}{2}a^{-3}(\boldsymbol{E}\times_G\Sigma)_{ij} + \tfrac{2}{3}a^{-3}\langle\boldsymbol{E},\Sigma\rangle_G G_{ij}\big]\\ & + 4\pi(N+1)a^{-3}(\Psi+C)^2\Sigma_{ij} - 4\pi(N+1)\dot a a^3\nabla_i\phi\nabla_j\phi + 4\pi a\nabla_{(i}N\nabla_{j)}\phi(\Psi+C)\\ & - 4\pi a(N+1)\big[\nabla_i\Psi\nabla_j\phi + \nabla_j\Psi\nabla_i\phi + (\Sigma^\sharp)^l_{(i}\nabla_{j)}\phi\nabla_l\phi - (\Psi+C)\nabla_i\nabla_j\phi\big]\\ & + (N+1)\big[\tfrac{2\pi}{3}a^6\partial_0(a^{-6}(\Psi+C)^2 + a^{-2}|\nabla\phi|_G^2) + 4\pi\tfrac{\dot a}{a}(\Psi+C)^2\big]G_{ij},\end{aligned} \tag{2-31a}$$

$$\begin{aligned}\partial_t\boldsymbol{B}_{ij} = {}& \frac{\dot a}{a}(3-N)\boldsymbol{B}_{ij} + a^{-1}(\nabla N\wedge_G\boldsymbol{E})_{ij} - (N+1)a^{-1}\,\mathrm{curl}_G\,\boldsymbol{E}_{ij}\\ & - (N+1)\big[\tfrac{5}{2}a^{-3}(\boldsymbol{B}\times_G\Sigma)_{ij} + \tfrac{2}{3}a^{-3}\langle\boldsymbol{B},\Sigma\rangle_G G_{ij}\big]\\ & - 4\pi(N+1)\boldsymbol{\varepsilon}[G]_{lmj}\big(a^3\nabla^{\sharp l}\nabla_j\phi\nabla^{\sharp m}\phi + a^{-1}(\Sigma^\sharp)^l{}_i\nabla^{\sharp m}\phi(\Psi+C)\big),\end{aligned} \tag{2-31b}$$

and the ***rescaled wave equation***

$$\partial_t\Psi = a\langle\nabla N, \nabla\phi\rangle_G + a(N+1)\Delta\phi - 3\frac{\dot a}{a}N(\Psi+C), \tag{2-32a}$$

along with the evolution equation

$$\partial_t\nabla\phi = a^{-3}(N+1)\nabla\Psi + a^{-3}(\Psi+C)\nabla N. \tag{2-32b}$$

Finally, we collect the ***rescaled Ricci evolution equation***

$$\begin{aligned}\partial_t\,\mathrm{Ric}[G]_{ab} = {}& a^{-3}(N+1)(\Delta_G\Sigma_{ab} - \nabla^{\sharp d}\nabla_a\Sigma_{db} - \nabla^{\sharp d}\nabla_b\Sigma_{da})\\ & + a^{-3}\nabla^{\sharp d}N(2\nabla_d\Sigma_{ab} - \nabla_a\Sigma_{db} - \nabla_b\Sigma_{da})\\ & - a^{-3}(\nabla_aN(\mathrm{div}_G\,\Sigma)_b + \nabla_b(\mathrm{div}_G\,\Sigma)_a) + \Delta_G N\big(a^{-3}\Sigma_{ab} + \tfrac{1}{3}\tau G_{ab}\big)\\ & - a^{-3}(\nabla^{\sharp d}\nabla_aN\cdot\Sigma_{db} + \nabla^{\sharp d}\nabla_bN\cdot\Sigma_{da}) + \tfrac{1}{3}\tau\nabla_a\nabla_bN,\end{aligned} \tag{2-33}$$

as well as (*in a coordinate neighbourhood*) *the Christoffel evolution equation*

$$\begin{aligned}\partial_t \Gamma^k_{ij}[G] &= \tfrac{1}{2}(G^{-1})^{kl}\big(\nabla_i(\partial_t G_{jl}) + \nabla_j(\partial_t G_{il}) - \nabla_l(\partial_t G_{ij})\big)\\ &= -(N+1)a^{-3}[\nabla_i(\Sigma^\sharp)^k{}_j + \nabla_j(\Sigma^\sharp)^k{}_i - \nabla^{\sharp k}\Sigma_{ij}]\\ &\quad - a^{-3}[\nabla_i N(\Sigma^\sharp)^k{}_j + \nabla_j N(\Sigma^\sharp)^k{}_i - \nabla^{\sharp k}N\Sigma_{ij}] + \frac{\dot a}{a}[\nabla_i N\cdot \mathbb{I}^k_j + \nabla_j N\cdot \mathbb{I}^k_i - \nabla^{\sharp k}N\cdot G_{ij}]. \quad (2\text{-}34)\end{aligned}$$

Proof. For the first identity in (2-34), we refer to [Chow et al. 2006, Lemma 2.27] and insert the evolution equation (2-28a). Otherwise, all equations simply follow by computing the effects of rescaling on the equations from Proposition 2.6 (respectively the Ricci evolution equation as in [Rendall 2008, Chapter 2.3, (2.32)]), as well as the Bel–Robinson evolution equations (2-25a)–(2-25b) and constraint equations (2-24a)–(2-24b). Notice that one already finds a solution to the system in Proposition 2.6 with the rescaled variables excluding (2-31a), (2-31b), (2-29c) and (2-29d). Conversely, all of the rescaled equations are satisfied by a solution to Proposition 2.6 at sufficiently high regularity. Hence, solving the full rescaled system is always sufficient to solve the Einstein system in Proposition 2.6 and they are equivalent if the initial data is regular enough to ensure sufficiently high regularity of solutions. □

2.6. ***Commuted equations.*** We collect Laplace-commuted versions of the equations for the rescaled variables in Proposition 2.10 in this subsection. For the sake of brevity, we will not state all possible commutations for every equation, but restrict ourselves to the ones we actually need within the bootstrap argument. We also refer to Section A.2 for expressions for commutators of spatial differential operators with each other and with ∂_t.

The terms written down explicitly in Lemma 2.11 are ones that dominate the evolution behaviour or that are the largest higher-order terms, both of which require careful treatment within the bootstrap argument. The error terms are broadly categorized into three groups:

- "Borderline" terms are terms that critically contribute to the fact that the energies diverge toward the big bang singularity. This almost always takes the form of adding energy terms at the same order as the evolved variable scaled by factors of the type εa^{-3} or $\varepsilon a^{-3-c\sqrt{\varepsilon}}$, which causes the energies to slightly diverge since a^{-3} is barely not integrable (see (2-6)).

- "Junk" terms are terms that are subcritical in the sense that they lead to integrable error terms, or terms that only contain lower-order derivatives of the solution variables.

- "Top" order terms (which only appear in (2-38a) and (2-38b)) are terms that are junk terms for low-order energies, but become borderline terms at top order.

All of these error terms are tracked schematically in Section A.3. Since we will only need L^2_G-bounds on these error terms, which are given in Section A.4, we will treat them as notational "black boxes" outside of the appendix.

Lemma 2.11 (Laplace-commuted rescaled equations). *Let $L \in 2\mathbb{N}$, $L \geq 2$. With error terms as defined in Appendix A.3, the system in Proposition 2.10 leads to the following Laplace-commuted equations*:

*The **Laplace-commuted rescaled evolution equations for the second fundamental form***

$$\partial_t \Delta^{L/2}\Sigma = -a\nabla^2\Delta^{L/2}N + a(N+1)\Delta^{L/2}\operatorname{Ric}[G] + \mathfrak{S}_{L,\text{Border}} + \mathfrak{S}_{L,\text{Junk}}, \tag{2-35}$$

*the **Laplace-commuted rescaled momentum constraint equations***

$$\operatorname{div}_G \Delta^{L/2}\Sigma = -8\pi(\Psi+C)\big[\nabla\Delta^{L/2}\phi + \Delta^{L/2-1}\operatorname{Ric}[G]*\nabla\phi\big] + \nabla\Delta^{L/2-1}\operatorname{Ric}[G]*\Sigma + \mathfrak{M}_{L,\text{Junk}} \tag{2-36a}$$

and

$$\operatorname{curl}_G \Delta^{L/2}\Sigma = -a^{-2}\Delta^{L/2}\boldsymbol{B} + \boldsymbol{\varepsilon}[G]*\nabla\Delta^{L/2-1}\operatorname{Ric}[G]*\Sigma + \widetilde{\mathfrak{M}}_{L,\text{Junk}}, \tag{2-36b}$$

*the **Laplace-commuted rescaled Hamiltonian constraint equations***

$$\begin{aligned}\Delta^{L/2}R[G] + a^{-4}\sum_{I_1+I_2=L}\nabla^{I_1}\Sigma * \nabla^{I_2}\Sigma \\ = 16\pi C a^{-4}\Delta^{L/2}\Psi + a^{-4}\sum_{I_1+I_2=L}[\nabla^{I_1}\Psi * \nabla^{I_2}\Psi + \nabla^{I_1+1}\phi * \nabla^{I_2+1}\phi]\end{aligned} \tag{2-36c}$$

and

$$\Delta^{L/2}\operatorname{Ric}[G] = a^{-4}\Delta^{L/2}\boldsymbol{E} - \tfrac{1}{3}\tau a^{-1}\Delta^{L/2}\Sigma + \mathfrak{H}_{L,\text{Border}} + \mathfrak{H}_{L,\text{Junk}}, \tag{2-36d}$$

*the **Laplace-commuted rescaled lapse equations***

$$\Delta^{L/2+1}N = \big(12\pi C^2 a^{-4} + \tfrac{1}{3}\big)\Delta^{L/2}N + 16\pi C a^{-4}\cdot\Delta^{L/2}\Psi + \mathfrak{N}_{L,\text{Border}} + \mathfrak{N}_{L,\text{Junk}}, \tag{2-37a}$$

$$\nabla\Delta^{L/2+1}N = \big(12\pi C^2 a^{-4} + \tfrac{1}{3}\big)\nabla\Delta^{L/2}N + 16\pi C a^{-4}\cdot\nabla\Delta^{L/2}\Psi + \mathfrak{N}_{L+1,\text{Border}} + \mathfrak{N}_{L+1,\text{Junk}}, \tag{2-37b}$$

*the **Laplace-commuted rescaled Bel–Robinson evolution equations***

$$\begin{aligned}\partial_t\Delta^{L/2}\boldsymbol{E} = {}& \frac{\dot a}{a}(3-N)\Delta^{L/2}\boldsymbol{E} + (N+1)a^{-1}\operatorname{curl}_G\Delta^{L/2}\boldsymbol{B} - a^{-1}\nabla\Delta^{L/2}N \wedge_G \boldsymbol{B} \\ & + 4\pi C^2 a^{-3}(N+1)\Delta^{L/2}\Sigma + 4\pi a(\Psi+C)\nabla\Delta^{L/2}N\otimes\nabla\phi \\ & + 4\pi a(\Psi+C)(N+1)\nabla^2\Delta^{L/2}\phi - 8\pi a(N+1)(\nabla\phi\otimes\nabla\Delta^{L/2}\Psi) \\ & + \mathfrak{E}_{L,\text{Border}} + \mathfrak{E}_{L,\text{top}} + \mathfrak{E}_{L,\text{Junk}},\end{aligned} \tag{2-38a}$$

$$\begin{aligned}\partial_t\Delta^{L/2}\boldsymbol{B} = {}& \frac{\dot a}{a}(3-N)\Delta^{L/2}\boldsymbol{B} - (N+1)a^{-1}\operatorname{curl}_G\Delta^{L/2}\boldsymbol{E} + a^{-1}\nabla\Delta^{L/2}N\wedge_G\boldsymbol{E} \\ & + a^3\boldsymbol{\varepsilon}[G]*\nabla^2\Delta^{L/2}\phi*\nabla\phi + \mathfrak{B}_{L,\text{Border}} + \mathfrak{B}_{L,\text{top}} + \mathfrak{B}_{L,\text{Junk}},\end{aligned} \tag{2-38b}$$

*the **Laplace-commuted rescaled matter evolution equations***

$$\partial_t\Delta^{L/2}\Psi = a\langle\nabla\Delta^{L/2}N, \nabla\phi\rangle_G + a(N+1)\Delta^{L/2+1}\phi - 3C\frac{\dot a}{a}\Delta^{L/2}N + \mathfrak{P}_{L,\text{Border}} + \mathfrak{P}_{L,\text{Junk}}, \tag{2-39a}$$

$$\partial_t\nabla\Delta^{L/2}\phi = a^{-3}(N+1)\nabla\Delta^{L/2}\Psi + Ca^{-3}\nabla\Delta^{L/2}N + \mathfrak{Q}_{L,\text{Border}} + \mathfrak{Q}_{L,\text{Junk}}, \tag{2-39b}$$

as well as (*also allowing* $L=0$ *for* (2-39d))

$$\begin{aligned}\partial_t\nabla_l\Delta^{L/2}\Psi = {}& a\nabla_l\nabla^{\sharp j}\Delta^{L/2}N\nabla_j\phi + a(N+1)\nabla_l\Delta^{L/2+1}\phi - 3C\frac{\dot a}{a}\nabla_l\Delta^{L/2}N \\ & + (\mathfrak{P}_{L+1,\text{Border}})_l + (\mathfrak{P}_{L+1,\text{Junk}})_l,\end{aligned} \tag{2-39c}$$

$$\partial_t\Delta^{L/2+1}\phi = a^{-3}(N+1)\Delta^{L/2+1}\Psi + Ca^{-3}\Delta^{L/2+1}N + \mathfrak{Q}_{L+1,\text{Border}} + \mathfrak{Q}_{L+1,\text{Junk}}, \tag{2-39d}$$

and the ***Laplace-commuted rescaled Ricci evolution equations***

$$\partial_t \Delta^{L/2}\mathrm{Ric}[G]_{ij} = a^{-3}(\Delta^{L/2+1}\Sigma_{ij} - 2\nabla^{\sharp m}\nabla_{(i}\Delta^{L/2}\Sigma_{j)m}) - \frac{\dot{a}}{a}(\nabla_i\nabla_j\Delta^{L/2}N + \Delta^{L/2+1}N\cdot G_{ij}) + (\mathfrak{R}_{L,\mathrm{Border}})_{ij} + (\mathfrak{R}_{L,\mathrm{Junk}})_{ij}, \tag{2-40a}$$

$$\partial_t \nabla_k\Delta^{L/2}\mathrm{Ric}[G]_{ij} = a^{-3}(\nabla_k\Delta^{L/2+1}\Sigma_{ij} - 2\nabla_k\nabla^{\sharp m}\nabla_{(i}\Delta^{L/2}\Sigma_{j)m}) - \frac{\dot{a}}{a}(\nabla_k\nabla_i\nabla_j\Delta^{L/2}N + \nabla_k\Delta^{L/2+1}N\cdot G_{ij}) + (\mathfrak{R}_{L+1,\mathrm{Border}})_{ijk} + (\mathfrak{R}_{L+1,\mathrm{Junk}})_{ijk}. \tag{2-40b}$$

Proof. The equations (2-36c),(2-36d) and (2-37a) are obtained by simply applying $\Delta^{L/2}$ on both sides of (2-29a), (2-29c) and (2-30a) respectively. For (2-36a) and (2-36b), we additionally use the commutator formulas (A-7e) and (A-7f), while for the evolution equations, we apply the respective commutator of ∂_t and spatial derivatives as collected in Lemma A.7 and commute Laplacians past ∇ and curl where needed (see the commutators in Lemma A.5). The commutators with ∂_t only cause additional borderline and junk terms that do not substantially influence the behaviour, while the spatial commutators often lead to high-order curvature terms, for example the Ricci terms in (2-36a), that need to be more carefully tracked. □

Remark 2.12 (simplified junk term notation). For junk terms that occur in an inner product with a tracefree symmetric tensor, any terms that are pure trace will immediately cancel and thus do not need to be taken into consideration for the following estimates, even if they have to be written down in the junk terms. Hence, we will denote with a superscript "$\|$" (for example $\mathfrak{H}^{\|}_{L,\mathrm{Junk}}$) on a schematic error term the expressions that arise when dropping all terms of the form $\zeta \cdot G$ for some scalar function ζ that occur in this term's definition (see, for example, (A-11d)).

3. Big bang stability: norms, energies and bootstrap assumptions

Herein, we state the norms and energies we use to control the solution variables. These will allow us to state our initial data and bootstrap assumptions, and we then provide which improvement we aim to achieve for the latter. Note that we do not provide the coerciveness of our energies immediately (and actually cannot, at least not in a manner useful to our analysis), but will establish Sobolev norm control in the proof of Corollary 7.3, the key ingredient being Lemma 4.5. Furthermore, we collect a local well-posedness statement from previous work in Section 3.4.

3.1. ***Norms.*** Recall that γ is the hyperbolic spatial reference metric on M introduced in Definition 2.1, which we view as a metric on any foliation hypersurface Σ_t (see Section 1.2.1), and G is the rescaled spatial metric arising from the evolution (see Definition 2.9).

Definition 3.1 (pointwise norms and volume forms). We denote by $|\cdot|_\gamma$ (resp. $|\cdot|_{G(t,\cdot)}$) the pointwise norm with regard to γ (resp. $G(t,\cdot)$). For the sake of simplicity, we define $|\zeta|_\gamma = |\zeta|_{G(t,\cdot)} = |\zeta(t,\cdot)|$ for any scalar function ζ on Σ_t. The volume forms on Σ_t with respect to γ and $G(t,\cdot)$ are written as vol_γ and $\mathrm{vol}_{G(t,\cdot)}$ (or just vol_G).

Definition 3.2 (L^2-norms). Let $\mathfrak{T}$ be a Σ_t-tangent (r,s)-tensor field (for $r,s\geq 0$). Then, we define

$$\|\mathfrak{T}\|^2_{L^2_\gamma(\Sigma_t)} = \|\mathfrak{T}(t,\cdot)\|^2_{L^2_\gamma(\Sigma_t)} := \int_M |\mathfrak{T}(t,\cdot)|^2_\gamma \,\mathrm{vol}_\gamma, \tag{3-1a}$$

$$\|\mathfrak{T}\|^2_{L^2_G(\Sigma_t)} = \|\mathfrak{T}(t,\cdot)\|^2_{L^2_G(\Sigma_t)} := \int_M |\mathfrak{T}(t,\cdot)|^2_{G(t,\cdot)} \,\mathrm{vol}_{G(t,\cdot)}. \tag{3-1b}$$

Definition 3.3 (Sobolev norms). Let $\mathfrak{T}$ be as above and $J\in\mathbb{N}_0$. We define

$$\|\mathfrak{T}\|^2_{\dot H^J_\gamma(\Sigma_t)} = \|\mathfrak{T}(t,\cdot)\|^2_{\dot H^J_\gamma} = \int_{\Sigma_t} |\hat\nabla^J \mathfrak{T}(t,\cdot)|^2_\gamma \,\mathrm{vol}_\gamma, \tag{3-2a}$$

$$\|\mathfrak{T}\|^2_{\dot H^J_G(\Sigma_t)} = \|\mathfrak{T}(t,\cdot)\|^2_{\dot H^J_G} = \int_{\Sigma_t} |\nabla^J \mathfrak{T}(t,\cdot)|^2_{G(t,\cdot)} \,\mathrm{vol}_{G(t,\cdot)}, \tag{3-2b}$$

$$\|\mathfrak{T}\|^2_{H^J_\gamma(\Sigma_t)} = \|\mathfrak{T}(t,\cdot)\|^2_{H^J_\gamma} = \sum_{k=0}^J \|\mathfrak{T}\|^2_{\dot H^k_\gamma(\Sigma_t)}, \tag{3-2c}$$

$$\|\mathfrak{T}\|^2_{H^J_G(\Sigma_t)} = \|\mathfrak{T}(t,\cdot)\|^2_{H^J_G} = \sum_{k=0}^J \|\mathfrak{T}\|^2_{\dot H^k_G(\Sigma_t)}. \tag{3-2d}$$

Definition 3.4 (supremum norms). For $\mathfrak{T}$ as above and $J\in\mathbb{N}_0$, we set

$$\|\mathfrak{T}\|_{\dot C^J_\gamma(\Sigma_t)} = \|\mathfrak{T}(t,\cdot)\|_{\dot C^J_\gamma} = \sup_{p\in\Sigma_t} |\hat\nabla^J \mathfrak{T}(t,\cdot)|_\gamma, \qquad \|\mathfrak{T}\|_{C^J_\gamma(\Sigma_t)} = \sum_{k=0}^J \|\mathfrak{T}\|_{\dot C^k_\gamma(\Sigma_t)}, \tag{3-3a}$$

$$\|\mathfrak{T}\|_{\dot C^J_G(\Sigma_t)} = \|\mathfrak{T}(t,\cdot)\|_{\dot C^J_G} = \sup_{p\in\Sigma_t} |\nabla^J \mathfrak{T}(t,\cdot)|_{G(t,\cdot)}, \quad \|\mathfrak{T}\|_{C^J_G(\Sigma_t)} = \sum_{k=0}^J \|\mathfrak{T}\|_{\dot C^k_G(\Sigma_t)}. \tag{3-3b}$$

Remark 3.5 (time-dependence is suppressed in notation). When the choice of t and Σ_t is clear from context, we will often drop time-dependences of G, $|\cdot|_G$, vol_G and $\mathfrak{T}$, suppress the hypersurface Σ_t in the Sobolev and supremum norms, and simply write $\int_M$ instead of $\int_{\Sigma_t}$. For example, we write

$$\|\mathfrak{T}\|^2_{L^2_G} = \int_M |\mathfrak{T}|^2_G \,\mathrm{vol}_G.$$

Definition 3.6 (solution norms). We define the following norms to measure the size of near-FLRW solutions:

$$\begin{aligned}\mathcal{H} = {}& \|\Psi\|_{H^{18}_G} + \|\nabla\phi\|_{H^{17}_G} + a^2\|\nabla\phi\|_{\dot H^{18}_G} + \|\Sigma\|_{H^{18}_G} + \|\boldsymbol{E}\|_{H^{18}_G} + \|\boldsymbol{B}\|_{H^{18}_G} \\ &+ \|G-\gamma\|_{H^{18}_G} + \|\mathrm{Ric}[G] + \tfrac{2}{9}G\|_{H^{16}_G} + a^{-4}\|N\|_{H^{16}_G} + a^{-2}\|N\|_{\dot H^{17}_G} + \|N\|_{\dot H^{18}_G},\end{aligned} \tag{3-4a}$$

$$\mathcal{H}_{\mathrm{top}} = a^2\|\Psi\|_{\dot H^{19}_G} + a^4\|\nabla\phi\|_{\dot H^{19}_G} + a^2\|\Sigma\|_{\dot H^{19}_G} + a^2\|\mathrm{Ric}[G] + \tfrac{2}{9}G\|_{\dot H^{17}_G}, \tag{3-4b}$$

$$\begin{aligned}\mathcal{C} = {}& \|\Psi\|_{C^{16}_G} + \|\nabla\phi\|_{C^{15}_G} + \|\Sigma\|_{C^{16}_G} + \|\boldsymbol{E}\|_{C^{16}_G} + \|\boldsymbol{B}\|_{C^{16}_G} \\ &+ \|G-\gamma\|_{C^{16}_G} + \|\mathrm{Ric}[G] + \tfrac{2}{9}G\|_{C^{14}_G} + a^{-4}\|N\|_{C^{14}_G} + a^{-2}\|N\|_{\dot C^{15}_G} + \|N\|_{\dot C^{16}_G},\end{aligned} \tag{3-4c}$$

$$\begin{aligned}\mathcal{C}_\gamma = {}& \|\Psi\|_{C^{16}_\gamma} + \|\nabla\phi\|_{C^{15}_\gamma} + \|\Sigma\|_{C^{16}_\gamma} + \|\boldsymbol{E}\|_{C^{16}_\gamma} + \|\boldsymbol{B}\|_{C^{16}_\gamma} \\ &+ \|G-\gamma\|_{C^{16}_\gamma} + \|\mathrm{Ric}[G] + \tfrac{2}{9}G\|_{C^{14}_\gamma} + a^{-4}\|N\|_{C^{14}_\gamma} + a^{-2}\|N\|_{\dot C^{15}_\gamma} + \|N\|_{\dot C^{16}_\gamma}.\end{aligned} \tag{3-4d}$$

Remark 3.7 (choice of metric and controlling Christoffel symbols). We could equivalently also phrase $\mathcal{H}$ in terms of γ-norms, or predominately use $\mathcal{C}_\gamma$ instead of $\mathcal{C}$, since we include the norms on $G-\gamma$ and

$\mathrm{Ric}[G]+\frac{2}{9}G=\big(\mathrm{Ric}[G]+\frac{2}{9}\gamma\big)+\frac{2}{9}(G-\gamma)$. We will demonstrate in Lemma 7.4 how H_G and C_γ norms can be used to control H_γ and C_G norms. We will also indicate how the initial data and bootstrap assumptions for $\mathcal{C}_\gamma$ and $\mathcal{C}$ are equivalent in Remarks 3.11 and 3.18. The main reason for this is that, by successively replacing local coordinates in the expressions of $\Gamma-\widehat{\Gamma}$ by $\hat{\nabla}$, one has

$$\|\Gamma-\widehat{\Gamma}\|_{C^{l-1}_\gamma(M)}\lesssim P_l(\|G-\gamma\|_{C^l_\gamma(M)},\|G^{-1}-\gamma^{-1}\|_{C^l_\gamma(M)}). \tag{3-5}$$

We choose to work predominately with norms in terms of the rescaled metric since quantities appearing in the Einstein system are naturally contracted by G, not γ, and we commute with differential operators associated with G.

Remark 3.8 (redundancies in the solution norms). The solution norms $\mathcal{H}$, $\mathcal{C}$ and $\mathcal{C}_\gamma$ are not "optimal" in the sense that controlling the norms of Ψ, $\nabla\phi$, Σ and $G-\gamma$ is entirely sufficient to gain the claimed control (up to constant) on N via the lapse equation, $\boldsymbol{E}$ and $\boldsymbol{B}$ via to the constraint equations and $\mathrm{Ric}[G]+\frac{2}{9}G$ via local coordinates. We choose to include all variables in the norms and subsequent assumptions mainly for the sake of convenience.

3.2. ***Energies.*** The fundamental objects used to control the solution variables are the energies that take the following form:

Definition 3.9 (energies). Let $l\in\mathbb{N}_0$. We define

$$\begin{aligned}\mathcal{E}^{(l)}(\phi,t)&=(-1)^l\int_M\Psi\Delta^l\Psi-a^4\phi\Delta^{l+1}\phi\,\mathrm{vol}_G\\&=\begin{cases}\int_M|\Delta^{l/2}\Psi|^2+a^4|\nabla\Delta^{l/2}\phi|^2_G\,\mathrm{vol}_G, & l\text{ even},\\ \int_M|\nabla\Delta^{(l-1)/2}\Psi|^2_G+a^4|\Delta^{(l+1)/2}\phi|^2_G\,\mathrm{vol}_G, & l\text{ odd},\end{cases}\end{aligned} \tag{3-6a}$$

$$\mathcal{E}^{(l)}(W,t)=(-1)^l\int_M\langle\boldsymbol{E},\Delta^l\boldsymbol{E}\rangle_G+\langle\boldsymbol{B},\Delta^l\boldsymbol{B}\rangle_G\,\mathrm{vol}_G, \tag{3-6b}$$

$$\mathcal{E}^{(l)}(\Sigma,t)=(-1)^l\int_M\langle\Sigma,\Delta^l\Sigma\rangle_G\,\mathrm{vol}_G, \tag{3-6c}$$

$$\mathcal{E}^{(l)}(\mathrm{Ric},\cdot)=(-1)^l\int_M\big\langle\mathrm{Ric}[G]+\tfrac{2}{9}G,\Delta^l\big(\mathrm{Ric}[G]+\tfrac{2}{9}G\big)\big\rangle_G\,\mathrm{vol}_G, \tag{3-6d}$$

$$\mathcal{E}^{(l)}(N,\cdot)=(-1)^l\int_M\langle N,\Delta^l N\rangle_G\,\mathrm{vol}_G. \tag{3-6e}$$

Usually, we will use integration by parts to distribute derivatives within the energies as in (3-6a). Further, we introduce the notation

$$\mathcal{E}^{(\leq l)}=\sum_{i=0}^{l}\mathcal{E}^{(i)} \tag{3-7}$$

for any of the energies above.

For any $l\in\mathbb{N}_0$ and any smooth functions $f_1,f_2\in C^\infty(\mathbb{R}_+)$, we have

$$f_1\cdot f_2\cdot\mathcal{E}^{(2l+1)}\leq\tfrac{1}{2}f_1^2\mathcal{E}^{(2l)}+\tfrac{1}{2}f_2^2\mathcal{E}^{(2l+2)}. \tag{3-8}$$

Performing the calculation for Σ as an example, we have

$$\mathcal{E}^{(2l+1)}(\Sigma,\cdot)=-\int_M\langle\Delta^l\Sigma,\Delta^{l+1}\Sigma\rangle_G\,\mathrm{vol}_G\leq\int_M|\Delta^l\Sigma|_G|\Delta^{l+1}\Sigma|_G\,\mathrm{vol}_G\leq\sqrt{\mathcal{E}^{(2l)}(\Sigma,\cdot)}\sqrt{\mathcal{E}^{(2l+2)}(\Sigma,\cdot)}.$$

Now, (3-8) then follows from the Young inequality. As a consequence, we also have

$$\mathcal{E}^{(\leq 2l)} \lesssim \sum_{m=0}^{l} \mathcal{E}^{(2m)}, \quad \mathcal{E}^{(\leq 2l+1)} \lesssim \sum_{m=0}^{l+1} \mathcal{E}^{(2m)}. \tag{3-9}$$

This allows us to largely restrict our analysis to even-order energies, outside of how we close the bootstrap argument at top order.

3.3. ***Assumptions on the initial data.*** With the necessary solution norms and energies now defined, we can now state what we assume near-FLRW initial data to satisfy:

Assumption 3.10 (near-FLRW initial data). *For some small enough $\varepsilon \in (0, 1)$ and the solution norms $\mathcal{H}$, $\mathcal{H}_{\mathrm{top}}$ and $\mathcal{C}$ as in Definition 3.6, we assume the rescaled initial data to be close to that of the FLRW solution in Lemma 2.3 in the following sense*:

$$\mathcal{H}(t_0) + \mathcal{H}_{\mathrm{top}}(t_0) + \mathcal{C}(t_0) \lesssim \varepsilon^2. \tag{3-10}$$

The assumptions on $\mathcal{H} + \mathcal{H}_{\mathrm{top}}$ also imply

$$\begin{aligned}\mathcal{E}^{(\leq 18)}(\phi, t_0) + \mathcal{E}^{(\leq 18)}(\Sigma, t_0) + \mathcal{E}^{(\leq 18)}(W, t_0) + \mathcal{E}^{(\leq 16)}(\mathrm{Ric}, t_0) \\ + \|\nabla\phi\|^2_{H^{18}_G} + \mathcal{E}^{(18)}(N, t_0) + a(t_0)^{-4}\mathcal{E}^{(17)}(N, t_0) + a(t_0)^{-8}\mathcal{E}^{(\leq 16)}(N, t_0) \\ + a(t_0)^4\mathcal{E}^{(19)}(\phi, t_0) + a(t_0)^4\mathcal{E}^{(19)}(\Sigma, t_0) + a(t_0)^4\mathcal{E}^{(17)}(\mathrm{Ric}, t_0) \lesssim \varepsilon^4.\end{aligned} \tag{3-11}$$

Remark 3.11 (initial data size in $\mathcal{C}_\gamma(t_0)$). Notice that by (3-5), (3-10) implies that

$$\|\Gamma - \widehat{\Gamma}\|_{C^{15}_G(\Sigma_{t_0})} \lesssim \varepsilon^4, \tag{3-12a}$$

and arguing along similar lines and using $L^2 - L^\infty$-estimates, also

$$\|\Gamma - \widehat{\Gamma}\|_{H^{17}_G(\Sigma_{t_0})} \lesssim \varepsilon^4. \tag{3-12b}$$

In particular, since moving from C^l_γ to C^l_G only requires control on Christoffel symbols to order $l-1$ for general tensors and $l-2$ for scalar functions, as well as zero order control on $G-\gamma$, it follows from (3-10) that

$$\mathcal{C}_\gamma(t_0) \lesssim \varepsilon^2. \tag{3-13}$$

We refer to the proof of Lemma 7.4 for a more detailed term analysis and how a similar argument also applies to the Sobolev norms.

Remark 3.12 (redundancies in the initial data assumptions). Similar to Remark 3.8, one could also reduce the initial data assumptions in (3-10), especially at top order. In particular, we highlight that the Bel–Robinson energy can be entirely controlled by the other terms that occur due to the additional scaled Σ-energy at order 19, or vice versa we could drop the latter in favour of the former. This will be reflected in Lemma 6.10.

Remark 3.13 (the volume form). Let μ_G and μ_γ denote the volume elements of G and γ respectively. Since the determinant is a smooth map on invertible matrices, the initial data assumptions on $G - \gamma$ also

imply

$$\|\mu_G - \mu_\gamma\|_{C^0_G(\Sigma_{t_0})} = \|\mu_G - \mu_\gamma\|_{C^0_\gamma(\Sigma_{t_0})} \lesssim \varepsilon^2. \tag{3-14}$$

Consequently, we have

$$\|\mathrm{vol}_G - \mathrm{vol}_\gamma\|_{C^0_\gamma(\Sigma_{t_0})} = \mu_\gamma^{-1}\|\mathrm{vol}_\gamma\|_{C^0_\gamma(\Sigma_{t_0})}\|\mu_{G(t_0,\cdot)} - \mu_\gamma\|_{C^0_\gamma(\Sigma_{t_0})} \lesssim \varepsilon^2$$

and, since $\|G^{-1} - \gamma^{-1}\|_{C^0_\gamma(\Sigma_{t_0})} \lesssim \varepsilon^2$ also follows by a von Neumann series argument from the initial data assumption on $G - \gamma$,

$$\|\mathrm{vol}_G - \mathrm{vol}_\gamma\|_{C^0_G(\Sigma_{t_0})} \lesssim \varepsilon^2. \tag{3-15}$$

3.4. ***Local well-posedness and continuation criteria.*** For everything that follows, we need to establish that the initial data assumptions above also ensure local well-posedness. For the core system, we translate the local well-posedness result for stiff fluids in [Rodnianski and Speck 2018b] to the subcase of the scalar field system. While statement and proof there are for vanishing spatial sectional curvature and what corresponds to choosing $C = \sqrt{\frac{2}{3}}$, the arguments for our setting are completely analogous.

Lemma 3.14 (local well-posedness and continuation criteria for the Einstein scalar-field system (big bang version); see [Rodnianski and Speck 2018b, Theorem 14.1]). *Let $N \geq 4$ be an integer and $(M, \mathring{g}, \mathring{k}, \mathring{\pi}, \mathring{\psi})$ be geometric initial data to the Einstein scalar-field system (see Section 1.2.1) and assume that one has*

$$\|\mathring{g} - a(t_0)^2\gamma\|_{H^{N+1}_\gamma(M)} + \|\mathring{k} + \tfrac{1}{3}\tau(t_0)\cdot a(t_0)^2\gamma\|_{H^N_\gamma(M)} + \|\mathring{\pi}\|_{H^N_\gamma(M)} + \|\mathring{\psi} - Ca(t_0)^{-3}\|_{H^N_\gamma(M)} < \infty,$$

as well as, for some sufficiently small $\eta' > 0$,

$$\|\mathring{\psi} - Ca(t_0)^{-3}\|_{C^0_\gamma(M)} \leq \eta'.$$

Then, the CMC-transported Einstein scalar-field system (respectively the rescaled system) is locally well-posed in the following sense: The initial data $(\mathring{g}, \mathring{k}, \mathring{\pi}, \mathring{\psi})$ launches a unique classical solution $(g, k, n, \nabla\phi, \partial_t\phi)$ to (2-16a)–(2-16b), (2-15a)–(2-15b), (2-18) and (2-17a) on $[t_1, t_0] \times M$ for some $t_1 \in (0, t_0)$ that satisfies $k^l{}_l = -3\dot{a}a^{-1}$ and $n > 0$, launches a solution to the Einstein scalar-field system and such that the variables enjoy the following regularity:

$$\begin{aligned}
g &\in C^{N-1}_{dt^2+\gamma}([t_1, t_0] \times M) \cap C^0([t_1, t_0], H^{N+1}_\gamma(M)),\\
k &\in C^{N-2}_{dt^2+\gamma}([t_1, t_0] \times M) \cap C^0([t_1, t_0], H^{N}_\gamma(M)),\\
\nabla\phi &\in C^{N-2}_{dt^2+\gamma}([t_1, t_0] \times M) \cap C^0([t_1, t_0], H^{N}_\gamma(M)),\\
\partial_t\phi &\in C^{N-2}_{dt^2+\gamma}([t_1, t_0] \times M) \cap C^0([t_1, t_0], H^{N}_\gamma(M)),\\
n &\in C^{N}_{dt^2+\gamma}([t_1, t_0] \times M) \cap C^0([t_1, t_0], H^{N+2}_\gamma(M)).
\end{aligned}$$

The rescaled variables $(G, \Sigma, N, \nabla\phi, \Psi)$ enjoy the analogous regularity. If $(\mathfrak{t}, t_0]$ is the maximal interval on which the above statements hold, then one either has $\mathfrak{t} = 0$ or one of the following blow-up criteria are satisfied:

(1) *The smallest eigenvalue of $g(t_m, \cdot)$ converges to 0 for some sequence $(t_m, x_m) \subseteq (\mathfrak{t}, t_0] \times M$ with $t_m \downarrow \mathfrak{t}$.*

(2) *$n(t_m, x_m)$ converges to 0 for some sequence $(t_m, x_m) \subseteq (\mathfrak{t}, t_0] \times M$ with $t_m \downarrow \mathfrak{t}$.*

(3) $(|\partial_0\phi|^2 + |\nabla\phi|_g^2)(t_m, x_m)$ *converges to* 0 *for some sequence* $(t_m, x_m) \subseteq (\mathfrak{t}, t_0] \times M$ *with* $t_m \downarrow \mathfrak{t}$.

(4) $s \in (\mathfrak{t}, t_0] \mapsto \|g\|_{C^2_\gamma(\Sigma_s)} + \|k\|_{C^1_\gamma(\Sigma_s)} + \|n\|_{C^2_\gamma(\Sigma_s)} + \|\partial_t\phi\|_{C^1_\gamma(\Sigma_s)} + \|\nabla\phi\|_{C^1_\gamma(\Sigma_s)}$ *is unbounded.*

A note on the proof. Note that the additional initial data requirement in the stiff-fluid setting that the pressure is strictly positive is covered by the smallness assumption on $\mathring{\psi} - Ca(t_0)^2$, since the pressure corresponds to $|\mathring{\psi}|^2 + |\mathring{\pi}|^2_{\mathring{g}}$ and the assumptions on $\partial_t\phi$ and $\nabla\phi$ ensure that (after embedding) this quantity behaves like $C^2 a(t_0)^{-6} + \mathcal{O}(\eta')$ at Σ_{t_0}. □

Corollary 3.15 (local well-posedness for the Bel–Robinson variables). *Under the assumptions of Lemma 3.14, the Bel–Robinson variables E and B corresponding to the Lorentzian metric* $\bar{g} = -n^2\,dt^2 + g$ *satisfy* (2-24a)–(2-24b), *are the unique classical solutions to the evolution equations* (2-25a)–(2-25b), *and satisfy*

$$E,\, B \in C^{N-3}([t_1, t_0] \times M) \cap C([t_1, t_0],\, H^{N-1}_\gamma(M)).$$

Proof. That E and B satisfy the constraint equations, solve the evolution equations and have the stated regularity on the interval of existence is a direct consequence of Lemma 3.14 and the computations in Section 2.4. Furthermore, with initial data derived from the constraint equations as in Remark 2.8, the hyperbolic system (2-24a)–(2-24b) launches a unique solution satisfying the regularity above that must then be (E, B). □

For sufficiently regular initial data ($N \geq 21$), it follows that

$$\mathcal{E}^{(\leq 19)}(\phi, \cdot),\ \mathcal{E}^{(\leq 18)}(W, \cdot),\ \mathcal{E}^{(\leq 19)}(\Sigma, \cdot),\ \mathcal{E}^{(\leq 17)}(\mathrm{Ric}, \cdot),\ \|G - \gamma\|_{H^{18}_G} \in C^1([t_1, t_0]),$$

and similarly the square of any supremum norm occurring in $\mathcal{C}$ is continuously differentiable on $[t_1, t_0]$. Strictly speaking, we would need to assume this additional regularity on our initial data for the computations in the following sections (especially Section 6) to hold. However, since smooth functions are dense in $H^l(M)$ for any $l \in \mathbb{N}_0$, any bounds on $\mathcal{H}(t)$ and $\mathcal{C}(t)$ that we prove assuming sufficient regularity at Σ_{t_0} then immediately extend to data only satisfying the regularity implied by (3-10).

Thus, from here on out, we will assume without loss of generality that all energies and squared norms are continuously differentiable on the domain of existence, and similarly all variables are continuously differentiable for the lower-order C_G-norm improvements in Section 4.2.

3.5. ***Bootstrap assumption.*** To keep an overview of the entire bootstrap argument, we state all of the assumptions and comprehensively list how we intend to improve them.

Assumption 3.16 (bootstrap assumption). *Fix some* $t_{\mathrm{Boot}} \in [0, t_0)$. *Further, let* $c_0 > 0$, *let* $\sigma \in (\varepsilon^{1/8}, 1]$ *be suitably small such that* $c_0\sigma < 1$, *and* $K_0 > 0$ *a suitable constant. For any* $t \in (t_{\mathrm{Boot}}, t_0]$, *we assume*

$$\mathcal{C}(t) \leq K_0 \varepsilon a(t)^{-c_0\sigma}. \tag{3-16}$$

Remark 3.17. More explicitly, (3-16) means

$$\|\Psi\|_{C^{16}_G} \leq K_0 \varepsilon a^{-c_0\sigma}, \tag{3-17a}$$

$$\|\nabla\phi\|_{C^{15}_G} \leq K_0 \varepsilon a^{-c_0\sigma}, \tag{3-17b}$$

$$\|\Sigma\|_{C^{16}_G} \leq K_0\varepsilon a^{-c_0\sigma}, \tag{3-17c}$$

$$\|\boldsymbol{E}\|_{C^{16}_G} \leq K_0\varepsilon a^{-c_0\sigma}, \tag{3-17d}$$

$$\|\boldsymbol{B}\|_{C^{16}_G} \leq K_0\varepsilon a^{-c_0\sigma}, \tag{3-17e}$$

$$\left\|\mathrm{Ric}[G] + \tfrac{2}{9}G\right\|_{C^{14}_G} \leq K_0\varepsilon a^{-c_0\sigma}, \tag{3-17f}$$

$$\|G - \gamma\|_{C^{16}_G} \leq K_0\varepsilon a^{-c_0\sigma}, \tag{3-17g}$$

$$\|N\|_{C^{14}_G} + a^2\|N\|_{\dot{C}^{15}_G} + a^4\|N\|_{\dot{C}^{16}_G} \leq K_0\varepsilon a^{4-c_0\sigma}, \tag{3-17h}$$

$$\|\Gamma - \widehat{\Gamma}\|_{C^{15}_G} \leq K_0\varepsilon a^{-c_0\sigma}. \tag{3-17i}$$

Remark 3.18 (bootstrap assumptions with respect to γ). Note again that we could equivalently make the above bootstrap assumptions with respect to H_γ- and C_γ-norms: For example, the assumptions (3-17i) and (3-17g) imply

$$\|\zeta\|_{C^l_\gamma} \lesssim a^{-c\sigma}\|\zeta\|_{C^l_G} + \|\zeta\|_{C^{\lceil (l-1)/2\rceil}_\gamma}\varepsilon a^{-c\sigma},$$
$$\|\mathfrak{T}\|_{C^l_\gamma} \lesssim a^{-c\sigma}\|\mathfrak{T}\|_{C^l_G} + \|\mathfrak{T}\|_{C^{\lceil l/2\rceil}_\gamma}\varepsilon a^{-c\sigma}$$

for any smooth function $\zeta \in C^\infty(\Sigma_t)$, any Σ_t-tangent tensor $\mathfrak{T}$ and a constant $c > 0$. This is essentially a direct consequence of (3-5), and we will prove an improved version of this rigorously in Lemma 7.4. Applying this to each norm in $\mathcal{C}$, we get

$$\mathcal{C}_\gamma \lesssim \varepsilon a^{-c\sigma} \tag{3-18}$$

for some updated constant $c \geq c_0$.

Remark 3.19 (strategy for the bootstrap improvement). Our goal is to improve the C-norm estimate to

$$\mathcal{C} \leq K_1\varepsilon^{9/8}a^{-c_1\varepsilon^{1/8}},$$

where c_1, $K_1 > 0$ are positive constants independent of σ and ε. Notice how this is actually an improvement if we choose σ suitably and then choose ε sufficiently small: Any update between K_0 and K_1 can be balanced out since we gain at least the additional prefactor $\varepsilon^{1/8}$ in each estimate, which we can then choose to have been suitably small. Similarly, we improve the power of a if we have $\varepsilon^{1/8}\cdot\sigma^{-1} < c_0/c_1$. If we then retroactively choose σ large enough compared to ε but small overall — for example $\sigma = \varepsilon^{1/16}$ — and then ensure that $\max\{c_0, c_1\}\varepsilon^{1/16} < 1$, as well as $c_1\varepsilon^{1/16} < c_0$, are satisfied by choosing ε to have been small enough, we have strictly improved the bootstrap assumptions.

Remark 3.20 (conventions within the bootstrap argument). Throughout the rest of the argument, we tacitly assume $t \in (t_{\mathrm{Boot}}, t_0]$ if not stated otherwise, and we assume ε and σ to be sufficiently small. In the proof of Theorem 8.2, we will choose $\sigma = \varepsilon^{1/16}$, but this explicit choice will not be used or needed up to that point. Finally, we allow $c \geq c_0$ be a constant that we may update from line to line, and will similarly deal with prefactors by "$\lesssim$"-notation where the constant may change in each line. These updates will always be independent of σ and ε, but may depend on t_0, and the quantities arising from the FLRW reference solution. Hence, we not only assume $c_0\sigma < 1$, but $c\sigma < 1$ throughout the argument.

4. Big bang stability: a priori estimates

In this section, we collect strong low-order C_G-norm estimates that follow as an immediate consequence from the bootstrap assumptions, starting with key estimates at the base level and followed by weaker, but still improved estimates at higher levels. Finally, we collect a differentiation formula for integrals with respect to vol_G, as well as a Sobolev estimate that lays the groundwork for energy coercivity. In particular, using the strong C_G-norm estimates, said estimate proves that moving between energies and norms at most incurs an error involving lower-order energies of the controlled variable and curvature energies, scaled by $a^{-c\sqrt{\varepsilon}}$.

4.1. ***Strong C^0_G-estimates.*** First, we establish a pointwise bound on the lapse that actually holds irrespective of the bootstrap assumptions:

Lemma 4.1 (maximum principle for the lapse). *The lapse remains positive and bounded throughout the evolution*:

$$n = N + 1 \in (0, 3]. \tag{4-1}$$

Proof. Let $t \in \mathbb{R}_+$ be arbitrary and $n_{\min}$ be the minimum of n over Σ_t at $(t, x_{\min})$. Then, $(\Delta_g n)(t, x_{\min}) > 0$ holds. If $n_{\min}$ were nonpositive, (2-17a) would lead to the following contradiction:

$$0 \geq -12\pi C^2 a^{-6} - \tfrac{1}{3}a^{-2} + n_{\min}\big[\tfrac{1}{3}a^{-2} + 4\pi C^2 a^{-6} + \langle \hat{k}, \hat{k}\rangle_g + 8\pi |\partial_0 \phi|^2\big] = \Delta_g n(t, x_{\min}) > 0.$$

This shows $n > 0$, and the upper bound follows analogously. □

The following estimate will be essential in dealing with borderline terms throughout the bootstrap argument:

Lemma 4.2 (strong C^0_G estimates). *The following estimates hold*:

$$\|\Psi\|_{C^0_G} \lesssim \varepsilon, \tag{4-2a}$$

$$\|\Sigma\|_{C^0_G} \lesssim \varepsilon, \tag{4-2b}$$

$$\|\boldsymbol{E}\|_{C^0_G} \lesssim \varepsilon. \tag{4-2c}$$

Proof. (4-2a): From (2-32a), we obtain the following using Lemma 4.1 for n, the bootstrap assumptions (3-17h) and (3-17b) and that $\dot{a} \simeq a^{-2}$ by (2-3):

$$|\partial_t \Psi| \lesssim \varepsilon a^{5-c\sigma} + \varepsilon a^{1-c\sigma} + \varepsilon a^{1-c\sigma}|\Psi| + \varepsilon a^{1-c\sigma}.$$

After integration, we thus obtain using the initial data assumption (3-10):

$$\begin{aligned} |\Psi(t)| &\lesssim |\Psi(t_0)| + \int_t^{t_0} \varepsilon a(s)^{1-c\sigma}\, ds + \int_t^{t_0} \varepsilon a(s)^{1-c\sigma}|\Psi(s)|\, ds \\ &\lesssim \varepsilon\Big(1 + \int_t^{t_0} a(s)^{1-c\sigma}\, ds\Big) + \int_t^{t_0} \varepsilon a(s)^{1-c\sigma}|\Psi(s)|\, ds. \end{aligned}$$

By (2-6), the integral over $a^{1-c\sigma}$ is bounded since $c\sigma < 1$, so the Gronwall lemma now yields (4-2a).

(4-2b): Notice that

$$\partial_t |\Sigma|^2_G = (\partial_t \Sigma^\sharp)^l{}_m (\Sigma^\sharp)^m{}_l + (\Sigma^\sharp)^l{}_m (\partial_t \Sigma^\sharp)^m{}_l = 2(\partial_t \Sigma^\sharp)^l{}_m (\Sigma^\sharp)^m{}_l \leq 2|\partial_t \Sigma^\sharp|_G |\Sigma|_G. \tag{4-3}$$

Now, we consider (2-28d) and, using the bootstrap assumptions (3-17h), (3-17f) and (3-17b), get

$$|\partial_t \Sigma^\sharp|_G \lesssim \tau|N||\Sigma^\sharp|_G + |\nabla^\sharp \nabla N|_G a + |N+1|a\big|\mathrm{Ric}[G]^\sharp + \tfrac{2}{9}G^\sharp\big|_G + |N+1|a|\nabla^\sharp\phi\nabla\phi|_G + \sqrt{3}|N|\cdot\big(4\pi C^2 a^{-3} + \tfrac{1}{9}a\big)$$
$$\lesssim \varepsilon a^{1-c\sigma}|\Sigma|_G + \varepsilon a^{1-c\sigma}.$$

We can now apply Lemma A.2 with $f = |\Sigma|_G^2$, and thus have along with (3-10) and (4-3)

$$|\Sigma|_G(t) \leq |\Sigma|_G(t_0) + \int_t^{t_0} |\partial_t \Sigma^\sharp|_G(s)\, ds \lesssim \varepsilon.$$

(4-2c): Using the constraint equation (2-29c) and that $\langle G, \boldsymbol{E}\rangle_G = \mathrm{tr}_G \boldsymbol{E} = 0$, one sees

$$|\boldsymbol{E}|_G^2 = \big\langle a^4\big(\mathrm{Ric}[G] + \tfrac{2}{9}G\big) - \dot{a}a^2\Sigma - \Sigma\odot_G\Sigma - 4\pi a^4\nabla\phi\nabla\phi,\, \boldsymbol{E}\big\rangle_G.$$

Then, applying the bootstrap assumptions (3-17f) and (3-17b) shows the Ricci and matter terms are bounded by $\varepsilon a^{4-c\sigma}|\boldsymbol{E}|_G$, and the a priori estimate (4-2b) along with $\dot{a}a^2 \simeq 1$ by (2-3) bounds the remaining terms by $\varepsilon|\boldsymbol{E}|_G$. The statement then follows by dividing by $|\boldsymbol{E}|_G$ and taking the supremum. □

Note that, in the proof of (4-2b), it was essential that we used (2-28d) instead of (2-28c), since using the latter would incur terms of the type $|\partial_t G|_G|\Sigma|_G^2$ and $a^{-3}|\Sigma|_G^3$ when computing the time derivative of $|\Sigma|_G^2$, which, at this point, behave like $\varepsilon a^{-3-c\sigma}|\Sigma|_G^2$, and thus not yield the sharp estimate (or even an improved estimate) that we will need to control borderline terms.

4.2. ***Strong low-order C_G-norm estimates.*** Now, we can prove the main supremum norm estimates in this section:

Lemma 4.3 (strong low-order C_G-norm estimates). *The following estimates hold*:

$$\|\Psi\|_{C_G^{13}} \lesssim \varepsilon a^{-c\sqrt{\varepsilon}}, \tag{4-4a}$$
$$\|\Sigma\|_{C_G^{12}} \lesssim \varepsilon a^{-c\sqrt{\varepsilon}}, \tag{4-4b}$$
$$\|G-\gamma\|_{C_G^{12}} \lesssim \sqrt{\varepsilon} a^{-c\sqrt{\varepsilon}}, \tag{4-4c}$$
$$\|G^{-1}-\gamma^{-1}\|_{C_G^{12}} \lesssim \sqrt{\varepsilon} a^{-c\sqrt{\varepsilon}}, \tag{4-4d}$$
$$\|\nabla\phi\|_{C_G^{12}} \lesssim \sqrt{\varepsilon} a^{-c\sqrt{\varepsilon}}, \tag{4-4e}$$
$$\big\|\mathrm{Ric}[G] + \tfrac{2}{9}G\big\|_{C_G^{10}} \lesssim \sqrt{\varepsilon} a^{-c\sqrt{\varepsilon}}, \tag{4-4f}$$
$$\|\boldsymbol{B}\|_{C_G^{11}} \lesssim \varepsilon a^{2-c\sqrt{\varepsilon}}, \tag{4-4g}$$
$$\|\boldsymbol{E}\|_{C_G^{12}} \lesssim \varepsilon a^{-c\sqrt{\varepsilon}}. \tag{4-4h}$$

Proof. Before going into the individual estimates, we collect the following commutator term estimates from the expressions in (A-10a)–(A-10b):

$$\|[\partial_t, \nabla^J]\zeta\|_{C_G^0} \lesssim a^{-3}\|N+1\|_{C_G^{J-1}}\|\Sigma\|_{C_G^{J-1}}\|\zeta\|_{C_G^{J-1}} + \frac{\dot{a}}{a}\|N\|_{C_G^{J-1}}\|\zeta\|_{C_G^{J-1}}, \tag{4-5}$$
$$\|[\partial_t, \nabla^J]\mathfrak{T}\|_{C_G^0} \lesssim a^{-3}\|N+1\|_{C_G^J}(\|\nabla^J\Sigma\|_{C_G^0}\|\mathfrak{T}\|_{C_G^0} + \|\Sigma\|_{C_G^{J-1}}\|\mathfrak{T}\|_{C_G^{J-1}}) + \frac{\dot{a}}{a}\|N\|_{C_G^J}\|\mathfrak{T}\|_{C_G^{J-1}}. \tag{4-6}$$

With this in hand, we will prove each estimate by iterating over the derivative order as long as the bootstrap assumptions can be applied. In each step, we use the previously obtained estimates at lower order to control the commutator term (with some additional care for $\mathfrak{T} = \Sigma$ which we need to consider first), while we can use similar arguments to those at order 0 to control the "core" of the evolution equations.

To start out, we apply (4-2b) on Σ and the bootstrap assumption (3-17h) on N to the rescaled evolution equations (2-28a)–(2-28b) and deduce

$$|\partial_t G^{\pm 1}|_G = |\partial_t (G^{\pm 1} - \gamma^{\pm 1})|_G \lesssim \varepsilon a^{-3} + \varepsilon a^{1-c\sigma} \lesssim \varepsilon a^{-3}. \tag{4-7}$$

(4-4b): We assume

$$\|\Sigma\|_{C_G^{J-1}} \lesssim \varepsilon a^{-c\sqrt{\varepsilon}} \tag{4-8}$$

to be satisfied for some $J \in \{1, \dots, 12\}$ (for $J = 1$, this is true by (4-2b)). Observe the following:

$$\partial_t |\nabla^J \Sigma|_G^2 = 2\langle \partial_t \nabla^J \Sigma, \nabla^J \Sigma\rangle_G + \partial_t G^{-1} * \nabla^J \Sigma * \nabla^J \Sigma.$$

Now, we commute (2-28d) with ∇^J: As before, $\nabla^J \partial_t \Sigma$ is bounded by $\varepsilon a^{-c\sigma}$ for any admissible J. Hence and using (4-7),

$$\partial_t |\nabla^J \Sigma|_G^2 \lesssim \varepsilon a^{-3} |\nabla^J \Sigma|_G^2 + (\varepsilon a^{1-c\sigma} + \|[\partial_t, \nabla^J]\Sigma\|_{C_G^0}) |\nabla^J \Sigma|_G$$

is satisfied. Looking at the commutator term using (4-6), we have with (4-2b) that

$$\|[\partial_t, \nabla^J]\Sigma\|_{C_G^0} \lesssim \varepsilon a^{-3} \|\Sigma\|_{\dot{C}_G^J} + a^{-3} \cdot \|\Sigma\|_{C_G^{J-1}}^2 + \varepsilon a^{1-c\sigma} \|\Sigma\|_{C_G^{J-1}}.$$

Altogether, we obtain

$$\partial_t |\nabla^J \Sigma|_G^2 \lesssim (\varepsilon a^{-3} \|\Sigma\|_{\dot{C}_G^J} + \varepsilon a^{-c\sigma} + \varepsilon^2 a^{-3-c\sqrt{\varepsilon}}) |\nabla^J \Sigma|_G.$$

With Lemma A.2 as well as the initial data assumption (3-10) and the integral formula (2-6) with $p = c\sqrt{\varepsilon}$, this implies

$$|\nabla^J \Sigma|_G(t) \lesssim \int_t^{t_0} \varepsilon a^{-3} \|\Sigma\|_{\dot{C}_G^J(\Sigma_s)}\, ds + \varepsilon(1 + \sqrt{\varepsilon} a^{-c\sqrt{\varepsilon}})$$

and consequently, after taking the supremum on the left and applying the Gronwall lemma,

$$\|\Sigma\|_{\dot{C}_G^J(\Sigma_s)} \lesssim \varepsilon a^{-c\sqrt{\varepsilon}}.$$

Combining this with (4-8) proves the statement up to order J, and hence shows (4-4b) by iterating the argument up to $J = 12$.

(4-4a): We again assume that

$$\|\Psi\|_{C_G^{J-1}} \lesssim \varepsilon a^{-c\sqrt{\varepsilon}} \tag{4-9}$$

holds for $J \in \{1, 2, \dots, 13\}$. Observe that

$$|\partial_t \nabla^J \Psi|_G \lesssim a\|N + 1\|_{C_G^{J+1}} \|\nabla\phi\|_{C_G^{J+1}} + \frac{\dot{a}}{a} \|\nabla N\|_{C_G^J} (1 + \|\Psi\|_{C_G^J}) + \|[\partial_t, \nabla^J]\Psi\|_{C_G^0}.$$

By (3-17h), (3-17b) and (3-17a), the first two summands can be bounded (up to constant) by $\varepsilon a^{1-c\sigma}$. By (4-9), (4-4b) and (3-17h) and using (4-5), the commutator term is bounded (up to constant) by $\varepsilon^2 a^{-3-c\sqrt{\varepsilon}}$. Altogether,

$$|\partial_t \nabla^J \Psi|_G \lesssim \varepsilon a^{1-c\sigma} + \varepsilon^2 a^{-3-c\sqrt{\varepsilon}}$$

follows. Inserting this and (4-7) into

$$|\partial_t(|\nabla^J \Psi|_G^2)| \leq |\partial_t G^{-1}|_G |\nabla^J \Psi|_G^2 + 2|\partial_t \nabla^J \Psi|_G \cdot |\nabla^J \Psi|_G$$

implies, with Lemma A.2,

$$\begin{aligned} |\nabla^J \Psi|_G(t) &\leq |\nabla^J \Psi|(t_0) + \int_t^{t_0} \big(\tfrac{1}{2}|\partial_t G^{-1}||\nabla^J \Psi|_G + |\partial_t \nabla^J \Psi|_G\big)(s)\, ds \\ &\lesssim \varepsilon^2 + \int_t^{t_0} \big(\varepsilon a(s)^{-3}|\nabla^J \Psi(s,\cdot)|_G + \varepsilon a(s)^{1-c\sigma} + \varepsilon^2 a(s)^{-3-c\sqrt{\varepsilon}}\big)\, ds. \end{aligned}$$

We obtain using (2-6)

$$|\nabla^J \Psi|_G(t) \lesssim \varepsilon a(t)^{-c\sqrt{\varepsilon}} + \int_t^{t_0} \varepsilon a(s)^{-3}|\nabla^J \Psi(s,\cdot)|_G\, ds.$$

The Gronwall lemma, applying (2-5) and taking the supremum over Σ_t then implies $|\nabla^J \Psi|_{\dot{C}^J_G} \lesssim \varepsilon a^{-c\sqrt{\varepsilon}}$. This proves (4-2a) by iterating over J and adding up the individual seminorms.

(4-4c)–(4-4d): Note that (4-7) implies (4-4c) at order 0 since one has

$$|\partial_t(|G-\gamma|_G)^2| \lesssim |\partial_t G^{-1}|_G |G-\gamma|_G^2 + |\partial_t(G-\gamma)|_G |G-\gamma|_G \lesssim \varepsilon a^{-3}(1+|G-\gamma|_G)|G-\gamma|_G,$$

which we can apply the Gronwall lemma to after integrating, along with (2-7) for the error term, as in the proof of (4-4b).

For higher orders, commuting (2-28a) with ∇^J and inserting (4-4b) and (3-17h) implies

$$\|\partial_t \nabla^J (G-\gamma)\|_{C^0_G} \lesssim \varepsilon a^{-3-c\sqrt{\varepsilon}} + \varepsilon a^{1-c\sigma} + \|[\partial_t, \nabla^J](G-\gamma)\|_{C^0_G},$$

with

$$\|[\partial_t, \nabla^J](G-\gamma)\|_{C^0_G} \lesssim (\varepsilon a^{-3-c\sqrt{\varepsilon}} + \varepsilon a^{1-c\sigma})\|G-\gamma\|_{C^{J-1}_G}.$$

Once again doing the same iterative argument over $J \leq 12$ and assuming the estimate to hold up to $J-1$, this altogether becomes

$$\|\partial_t \nabla^J (G-\gamma)\|_{C^0_G} \lesssim \varepsilon a^{-3-c\sqrt{\varepsilon}},$$

implying with (2-6)

$$\|\nabla^J (G-\gamma)\|_{C^0_G} \lesssim \varepsilon^2 + \varepsilon \int_t^{t_0} a(s)^{-3-c\sqrt{\varepsilon}}\, ds \lesssim \sqrt{\varepsilon} a^{-c\sqrt{\varepsilon}}.$$

The argument for $G^{-1} - \gamma^{-1}$ is completely analogous.

(4-4e): We only prove the statement for C^0_G, the full estimate extends from there by the same iterative arguments as above. Considering (2-32b), Lemma 4.1, (4-2a) and (2-3), we have

$$|\partial_t \nabla \phi|_G \lesssim a^{-3}(|\nabla \Psi|_G + |\nabla N|_G)$$

and thus, with (4-4a) and the bootstrap assumption (3-17h),

$$|\partial_t \nabla\phi|_G \lesssim \varepsilon a^{-3-c\sqrt{\varepsilon}}.$$

With (4-7), this implies

$$|\partial_t |\nabla\phi|_G|^2 \lesssim \varepsilon a^{-3}|\nabla\phi|_G^2 + \varepsilon a^{-3-c\sqrt{\varepsilon}}|\nabla\phi|_G$$

and the statement follows as usual by applying Lemma A.2, (2-7) and the Gronwall lemma.

(4-4f): This follows as in the proof of (4-4c) using (2-33) and (2-28a) and their commuted analogues.

Once again, for C^0_G, we have (4-4g): This is obtained immediately from commuting (2-29d) with ∇^J and applying (4-4b). Notice that the Levi-Civita tensor can be absorbed into the implicit constants since $|\varepsilon[G]|_G = \sqrt{6}$ holds (see (A-2c)).

(4-4h): This follows like in the proof of (4-2c) from applying (3-17f), (3-17b) and (4-4b) to the constraint equation (2-29c) commuted with ∇^J. □

4.3. ***Other useful a priori observations.*** Before moving on to the energy estimates, we collect a differentiation identity and lay the groundwork for energy coercivity:

Lemma 4.4 (the volume form and differentiation of integrals). *Let $\mu_G = \sqrt{\det G}$ denote the volume element with regard to G. It satisfies*

$$\partial_t \mu_G = \tfrac{1}{2}\mu_G (G^{-1})^{ij}\partial_t G_{ij} = -N\tau\mu_G\,, \tag{4-10}$$

and hence one has

$$\|\mu_G - \mu_\gamma\|_{C^0_G} \lesssim \varepsilon \tag{4-11}$$

on $(\Sigma_t)_{t\in(t_{\mathrm{Boot}},t_0]}$. Further, for any differentiable function ζ, one has

$$\partial_t \int_M \zeta \,\mathrm{vol}_G = \int_M \partial_t\zeta\,\mathrm{vol}_G - \int_M N\tau\cdot\zeta\,\mathrm{vol}_G. \tag{4-12}$$

Proof. From (2-28a), we obtain $(G^{-1})^{ij}\partial_t G_{ij} = -2N\tau$, and (4-10) follows by

$$\partial_t\mu_G = \tfrac{1}{2}\sqrt{\det G}(G^{-1})^{ij}\partial_t G_{ij} = -N\tau\mu_G.$$

Hence, we have using (3-17h) and the initial data estimate (3-14) that

$$|\mu_G - \mu_\gamma|(t,\cdot) \lesssim \varepsilon + \int_t^{t_0} \varepsilon a(s)^{1-c\sigma}|\mu_G - \mu_\gamma|(s,\cdot)\,ds$$

holds, and thus (4-11) after applying the Gronwall lemma.

Finally, we obtain (4-12) by writing $\mathrm{vol}_G = (\mu_G/\mu_\gamma)\mathrm{vol}_\gamma$ and inserting (4-10). □

Lemma 4.5 (preliminary Sobolev norm estimates). *Let ζ be a scalar function and $\mathfrak{T}$ be a symmetric Σ_t-tangent $(0,2)$-tensor, and let $l \in \{1,\dots,9\}$. Then, on $(t_{\mathrm{Boot}}, t_0]$, the following estimates are satisfied: For $l > 5$, one has*

$$\|\nabla^2\zeta\|^2_{L^2_G} \lesssim \|\Delta\zeta\|^2_{L^2_G} + a^{-c\sqrt{\varepsilon}}\|\nabla\zeta\|^2_{L^2_G}, \tag{4-13a}$$

$$\|\zeta\|^2_{H^{2l}_G} \lesssim \|\Delta^l\zeta\|^2_{L^2_G} + a^{-c\sqrt{\varepsilon}}\Big(\sum_{m=0}^{l-1}\|\Delta^m\zeta\|^2_{L^2_G} + \|\zeta\|^2_{C^{2l-12}_G}\mathcal{E}^{(\le 2l-3)}(\mathrm{Ric},\cdot)\Big), \tag{4-13b}$$

$$\sum_{m=1}^{2l+1}\|\zeta\|^2_{\dot{H}^m_G}\lesssim\|\nabla\Delta^l\zeta\|^2_{L^2_G}+a^{-c\sqrt{\varepsilon}}\Big(\sum_{m=0}^{l-1}\|\nabla\Delta^m\zeta\|^2_{L^2_G}+\|\nabla\zeta\|^2_{C^{2l-12}_G}\mathcal{E}^{(\leq 2l-2)}(\mathrm{Ric},\cdot)\Big),\tag{4-13c}$$

$$\sum_{m=1}^{2l}\|\nabla\zeta\|^2_{\dot{H}^m_G}\lesssim\|\nabla\Delta^l\zeta\|^2_{L^2_G}+a^{-c\sqrt{\varepsilon}}\Big(\sum_{m=0}^{l-1}\|\nabla\Delta^m\zeta\|^2_{L^2_G}+\|\nabla\zeta\|^2_{C^{2l-11}_G}\mathcal{E}^{(\leq 2l-2)}(\mathrm{Ric},\cdot)\Big)\tag{4-13d}$$

and

$$\|\mathfrak{T}\|^2_{H^{2l}_G}\lesssim\|\Delta^l\mathfrak{T}\|^2_{L^2_G}+a^{-c\sqrt{\varepsilon}}\Big(\sum_{m=0}^{l-1}\|\Delta^m\mathfrak{T}\|^2_{L^2_G}+\|\mathfrak{T}\|^2_{C^{2l-11}_G}\mathcal{E}^{(\leq 2l-2)}(\mathrm{Ric},\cdot)\Big),\tag{4-14a}$$

$$\sum_{m=1}^{2l+1}\|\mathfrak{T}\|^2_{\dot{H}^m_G}\lesssim\|\nabla\Delta^l\mathfrak{T}\|^2_{L^2_G}+a^{-c\sqrt{\varepsilon}}\Big(\sum_{m=0}^{l-1}\|\nabla\Delta^m\mathfrak{T}\|^2_{L^2_G}+\|\mathfrak{T}\|^2_{C^{2l-10}_G}\mathcal{E}^{(\leq 2l-1)}(\mathrm{Ric},\cdot)\Big).\tag{4-14b}$$

More precisely, the Ricci energy terms can be dropped in all of the above estimates for $l\leq 5$.

Remark 4.6. We stress that Lemma 4.5 is crucial for everything that follows in multiple ways:

Firstly, the L^2_G-norms containing ζ and $\mathfrak{T}$ on the right-hand sides above except (4-13d) are in precisely the form the energies in Definition 3.9 take. Hence, this is what will actually yield near-coercivity of our energies since the C_G-norms can be controlled by $a^{-c\sqrt{\varepsilon}}$ or better using the a priori estimates from Lemma 4.3, as well as (3-17h) for the lapse. This will be shown more explicitly as an intermediary step in improving the bootstrap assumptions for $\mathcal{C}$ (see proof of Corollary 7.3).

Secondly, a downside of using Δ as the main differential operator to commute with the Einstein scalar-field system is that it creates error terms that we can only bound by Sobolev norms and not directly express as energies. Thus, we need a way to translate this information back to energies to formulate energy inequalities. A lot of this is done "under the hood" in the error term estimates in Section A.4.

Finally, some top-order terms also do not appear in a way that their L^2-norm is directly the square root of an energy (see, for example, the term $a\nabla^2\Delta^{L/2}N$ in (2-35)), and some borderline terms would lead to nonintegrable divergences if we were to incur additional divergences in estimation (see, for example, the first term in (A-14a)). Lemma 4.5 precisely provides a way to relate these terms to energies . Additionally, by applying these estimates for terms of the form $\Delta^{L/2}\zeta$ and $\Delta^{L/2}\mathfrak{T}$, one can avoid high-order curvature energies that run the risk of breaking the energy hierarchy.

Proof. Since the arguments for all of the inequalities above are very similar, we only prove (4-14a) in full and then briefly address the other estimates.

Letting $\widetilde{\mathfrak{T}}_{i_1\cdots i_{2l}k_1k_2}=\nabla_{i_1}\cdots\nabla_{i_{2l}}\mathfrak{T}_{k_1k_2}$, we compute with the commutator formula (A-6c) and strong C_G-norm estimate (4-4f):

$$\begin{aligned}\int_M|\nabla^2\widetilde{\mathfrak{T}}|^2_G&=-\int_M\langle\nabla\widetilde{\mathfrak{T}},\Delta\nabla\widetilde{\mathfrak{T}}\rangle_G\,\mathrm{vol}_G\\&=-\int_M\langle\nabla\widetilde{\mathfrak{T}},\nabla\Delta\widetilde{\mathfrak{T}}\rangle_G\,\mathrm{vol}_G+\int_M\nabla\widetilde{\mathfrak{T}}*[\nabla\,\mathrm{Ric}[G]*\widetilde{\mathfrak{T}}+\mathrm{Ric}[G]*\nabla\widetilde{\mathfrak{T}}]\,\mathrm{vol}_G\\&\lesssim\int_M|\Delta\widetilde{\mathfrak{T}}|^2_G\,\mathrm{vol}_G+\big(1+\big\|\mathrm{Ric}[G]+\tfrac{2}{9}G\big\|_{C^1_G}\big)\cdot\Big[\int_M|\widetilde{\mathfrak{T}}|^2_G\,\mathrm{vol}_G+\int_M|\nabla\widetilde{\mathfrak{T}}|^2_G\,\mathrm{vol}_G\Big]\\&\lesssim\int_M|\Delta\widetilde{\mathfrak{T}}|^2_G+a^{-c\sqrt{\varepsilon}}\int_M|\widetilde{\mathfrak{T}}|^2_G\,\mathrm{vol}_G.\end{aligned}$$

In the final step, we used integration by parts to obtain

$$a^{-c\sqrt{\varepsilon}}\int_M |\nabla\widetilde{\mathfrak{T}}|_G^2 \operatorname{vol}_G \le \int_M |\Delta\widetilde{\mathfrak{T}}|_G \cdot a^{-c\sqrt{\varepsilon}}|\widetilde{\mathfrak{T}}|_G \operatorname{vol}_G \lesssim \int_M (|\Delta\widetilde{\mathfrak{T}}|_G^2 + a^{-2c\sqrt{\varepsilon}}|\widetilde{\mathfrak{T}}|_G^2)\operatorname{vol}_G$$

and updated c. This already shows (4-14a) for $l = 1$. Assume now that (4-14a) holds up to some $l \in \mathbb{N}$, $l \le 9$ and *any* symmetric Σ_t-tangent $(0,2)$-tensor field. By applying (A-6d), we have

$$\begin{aligned}\Delta\widetilde{\mathfrak{T}} &= \nabla^{2l}\Delta\mathfrak{T} + [\Delta,\nabla^2]\nabla^{2l-2}\mathfrak{T} + \cdots + \nabla^{2l-2}[\Delta,\nabla^2]\mathfrak{T}\\ &= \nabla^{2l}\Delta\mathfrak{T} + \sum_{I_{\mathrm{Ric}}+I_{\mathfrak{T}}=2l} \nabla^{I_{\mathrm{Ric}}}\big(\mathrm{Ric}[G] + \tfrac{2}{9}G\big) * \nabla^{I_{\mathfrak{T}}}\mathfrak{T} + G * \nabla^{2l}\mathfrak{T}. \end{aligned} \tag{4-15}$$

Subsequently, we have for $l > 5$ using the strong C_G-norm estimate (4-4f) for any Ricci term of order 10 or lower that

$$\|\mathfrak{T}\|^2_{\dot{H}^{2(l+1)}_G} = \int_M |\nabla^2\widetilde{\mathfrak{T}}|_G^2 \operatorname{vol}_G \lesssim \|\Delta\mathfrak{T}\|^2_{\dot{H}^{2l}_G} + (1+\sqrt{\varepsilon}a^{-c\sqrt{\varepsilon}})\|\mathfrak{T}\|^2_{H^{2l}_G} + \|\mathfrak{T}\|^2_{C^{2l-11}_G}\big\|\mathrm{Ric}[G] + \tfrac{2}{9}G\big\|^2_{H^{2l}_G},$$

and get the same estimate without the final term for $l \le 5$. By assumption, we can estimate $\|\Delta\mathfrak{T}\|^2_{H^{2l}_G}$, $\|\mathfrak{T}\|^2_{H^{2l}_G}$ and $\big\|\mathrm{Ric}[G] + \frac{2}{9}G\big\|^2_{H^{2l}_G}$ as in (4-14a), and get the following for $l > 5$:

$$\begin{aligned}\int_M |\nabla^{2l+2}\mathfrak{T}|_G^2 \operatorname{vol}_G &\lesssim \Big[\|\Delta^l\Delta\mathfrak{T}\|^2_{L^2_G} + a^{-c\sqrt{\varepsilon}}\sum_{m=0}^{l-1}\|\Delta^{m+1}\mathfrak{T}\|^2_{L^2_G} + a^{-c\sqrt{\varepsilon}}\|\Delta\mathfrak{T}\|^2_{C^{2l-12}_G}\mathcal{E}^{(\le 2l-2)}(\mathrm{Ric},\cdot)\Big]\\ &\quad + \Big[a^{-c\sqrt{\varepsilon}}\|\Delta^l\mathfrak{T}\|^2_{L^2_G} + a^{-c\sqrt{\varepsilon}}\sum_{m=0}^{l-1}\|\Delta^m\mathfrak{T}\|^2_{L^2_G} + a^{-c\sqrt{\varepsilon}}\|\mathfrak{T}\|^2_{C^{2l-12}_G}\mathcal{E}^{(\le 2l-2)}(\mathrm{Ric},\cdot)\Big]\\ &\quad + a^{-c\sqrt{\varepsilon}}\|\mathfrak{T}\|^2_{C^{2l-11}_G}\big[\mathcal{E}^{(2l)}(\mathrm{Ric},\cdot) + \big(\mathcal{E}^{(\le 2l-2)}(\mathrm{Ric},\cdot) + \varepsilon a^{-c\sqrt{\varepsilon}}\mathcal{E}^{(\le 2l-2)}(\mathrm{Ric},\cdot)\big)\big]\\ &\lesssim \|\Delta^{l+1}\mathfrak{T}\|^2_{L^2_G} + a^{-c\sqrt{\varepsilon}}\Big(\sum_{m=0}^{l}\|\Delta^m\mathfrak{T}\|^2_{L^2_G} + \|\mathfrak{T}\|^2_{C^{2l-10_G}}\mathcal{E}^{(\le 2l)}(\mathrm{Ric},\cdot)\Big).\end{aligned}$$

For $l = 5$, we get analogous estimates dropping the Ricci energies in the first two lines, and for $l = 4$, the same with all curvature terms dropped.

To prove the statement for $l+1$, it now remains to be shown that $\|\mathfrak{T}\|^2_{\dot{H}^{2l+1}_G}$ can be bounded by the same right-hand side (up to constant) as above. By integration by parts, one has

$$\|\nabla^{2l+1}\mathfrak{T}\|^2_{L^2_G} \lesssim \|\nabla^{2l}\mathfrak{T}\|^2_{L^2_G} + \|\Delta\nabla^{2l}\mathfrak{T}\|^2_{L^2_G},$$

where the latter tensor is precisely $\Delta\widetilde{\mathfrak{T}}$, which we just treated, and the former is covered by the induction assumption at order $2l$. So, (4-14a) now follows for $l+1$, and thus by iteration up to $l = 10$.

The proof of (4-14b) is analogous — we note that since we actually only needed a strong estimate on $\|\mathrm{Ric}[G] + 2G\|_{C^9_G}$ for the previous inequality, but (4-4f) holds at C^{10}_G, this gives enough room to extend the argument in full despite the extra derivative order.

For both, note that we only need to estimate the Ricci terms in the L^2_G-norm if one cannot apply the a priori estimate (4-4f) to all $\nabla^{I_{\mathrm{Ric}}}\mathrm{Ric}[G]$ that occur in (4-15), and thus we could easily adjust the proof such that the Ricci energy does not occur in any of the proofs as long as $2l-1 \le 10$ is satisfied, so for $l \le 5$.

The estimates (4-13b)–(4-13d) are proved identically, the only difference being that one order of curvature less enters in the commutator terms in (4-15), leading to one order less in curvature in total.

For (4-13a), we note that we can avoid incurring any L^2-norm by carefully repeating the argument we made for $\widetilde{\mathfrak{T}}$ using (A-6a):

$$\begin{aligned}\int_M |\nabla^2\zeta|_G^2 \operatorname{vol}_G &= \int_M -\langle \nabla\zeta, \nabla\Delta\zeta\rangle_G \operatorname{vol}_G + \int_M \operatorname{Ric}[G] * \nabla\zeta * \nabla\zeta \operatorname{vol}_G \\ &\lesssim \int_M |\Delta\zeta|_G^2 \operatorname{vol}_G + a^{-c\sqrt{\varepsilon}} \int_M |\nabla\zeta|_G^2 \operatorname{vol}_G. \qquad \square\end{aligned}$$

5. Big bang stability: elliptic lapse estimates

In this section, we study the elliptic structure of (2-30a)–(2-30b), which admit estimates controlling (time-scaled) lapse energies by other energy quantities. To this end, we recast these equations as follows:

Definition 5.1 (elliptic operators). For any (sufficiently regular) scalar function ζ on Σ_t, we define the differential operators

$$\mathcal{L}\zeta = a^4\Delta\zeta - f\cdot\zeta, \quad f = \tfrac{1}{3}a^4 + 12\pi C^2 + \underbrace{\langle\Sigma, \Sigma\rangle_G + 8\pi\Psi^2 + 16\pi C\Psi}_{=:F}, \tag{5-1a}$$

$$\widetilde{\mathcal{L}}\zeta = a^4\Delta\zeta - \tilde{f}\cdot\zeta, \quad \tilde{f} = \tfrac{1}{3}a^4 + 12\pi C^2 + \underbrace{a^4\big[R[G] + \tfrac{2}{3} - 8\pi|\nabla\phi|_G^2\big]}_{=\widetilde{F}}. \tag{5-1b}$$

Note that the lapse equations (2-30a), respectively (2-30b), now read

$$\mathcal{L}N = F, \quad \text{respectively} \quad \widetilde{\mathcal{L}}N = \widetilde{F}. \tag{5-2}$$

Furthermore, observe that

$$[\mathcal{L}, \Delta]\zeta = \Delta f\cdot\zeta + 2\langle\nabla f, \nabla\zeta\rangle_G = \Delta F\cdot\zeta + 2\langle\nabla F, \nabla\zeta\rangle_G, \tag{5-3a}$$

$$[\widetilde{\mathcal{L}}, \Delta]\zeta = \Delta\widetilde{F}\cdot\zeta + 2\langle\nabla\widetilde{F}, \nabla\zeta\rangle_G. \tag{5-3b}$$

5.1. ***Elliptic lapse estimates with $\mathcal{L}$.*** We first study the elliptic operator $\mathcal{L}$, which will admit weak lapse energy estimates in terms of scalar field quantities and Σ, up to curvature errors, that can in particular be utilized at high orders without having to resort to higher derivative levels. Before moving on to the estimates themselves, we collect a couple of inequalities we can deduce from the bootstrap assumptions and strong C_G-norm estimates.

Remark 5.2. There exists a constant $K > 0$ such that, for $\varepsilon > 0$ small enough, the following estimates hold:

- $F \geq -K\varepsilon$, and equivalently $f \geq 12\pi C^2 - K\varepsilon$. This is ensured by (4-2b) and (4-2a). In particular, we can assume ε to have been small enough such that $f - 6\pi C^2$ can be bounded from below by a positive constant that is independent of ε (for example, $3\pi C^2$).
- $|\nabla f|_G = |\nabla F|_G \leq K\varepsilon a^{-c\sqrt{\varepsilon}}$. This is given by (4-4b) and (4-4a).

Lemma 5.3 (elliptic estimates with $\mathcal{L}$). *Consider scalar functions ζ, Z on Σ_t that satisfy*

$$\mathcal{L}\zeta = Z. \tag{5-4}$$

Then,

$$a^4\|\Delta\zeta\|_{L^2_G} + a^2\|\nabla\zeta\|_{L^2_G} + \|\zeta\|_{L^2_G} \lesssim \|Z\|_{L^2_G}. \tag{5-5}$$

Proof. The proof follows along the same lines as that of [Speck 2018, Lemma 16.5]: First, we obtain the following by multiplying (5-4) with $-\zeta$ and integrating, and using that $f - 6\pi C^2$ is bounded from below by a positive constant (see the first point in Remark 5.2):

$$\int_M (a^4|\nabla\zeta|_G^2 + |\zeta|^2)\mathrm{vol}_G \lesssim \|Z\|_{L_G^2}^2.$$

Next, we multiply (5-4) with $a^4\Delta\zeta$ and obtain

$$\int_M (a^8|\Delta\zeta|^2 + a^4|\nabla\zeta|_G^2 f)\,\mathrm{vol}_G \leq \int_M \tfrac{1}{2}|Z|^2 + \tfrac{1}{2}a^8|\Delta\zeta|^2 + \big(\tfrac{1}{2}|\zeta|^2 + \tfrac{1}{2}a^4|\nabla\zeta|_G^2\big)a^2\|\nabla f\|_{L_G^\infty}\,\mathrm{vol}_G.$$

Using the second point in Remark 5.2, as well as the previous step, we can now conclude

$$\int_M (a^8|\Delta\zeta|^2 + a^4|\nabla\zeta|_G^2)\,\mathrm{vol}_G \lesssim (1+\varepsilon)\|Z\|_{L_G^2}^2,$$

and thus the statement after rearranging. □

Corollary 5.4 (intermediary elliptic lapse estimate with $\mathcal{L}$). *The following estimates hold for any* $l \in \{0, \dots, 10\}$:

$$a^4\|\Delta^{l+1}N\|_{L_G^2} + a^2\|\nabla\Delta^l N\|_{L_G^2} + \|\Delta^l N\|_{L_G^2} \lesssim \|\Delta^l F\|_{L_G^2} + \underbrace{\varepsilon a^{-c\sqrt{\varepsilon}}\|F\|_{H_G^{2(l-1)}}}_{\text{not present for } l=0} + \underbrace{\varepsilon^2 a^{4-c\sigma}\sqrt{\mathcal{E}^{(\leq 2l-4)}(\mathrm{Ric},\cdot)}}_{\text{not present for } l\leq 1}.$$

Proof. We prove the statement by induction over $l \in \mathbb{N}$: For $l = 0$, the estimates immediately follow from (5-2) and Lemma 5.3. Assume the statement to be satisfied up to $l-1$ for some $l \in \mathbb{N}_0$, $l \leq 11$. We get, applying (5-3a) iteratively,

$$\mathcal{L}\Delta^l N = \sum_{I=1}^{2l-1} \nabla^I F * \nabla^{2l-I} N + (N+1)\Delta^l F.$$

Applying Lemma 5.3 to $\zeta = \Delta^l N$, as well as Lemma 4.1, yields

$$a^4\|\Delta^{l+1}N\|_{L_G^2} + a^2\|\nabla\Delta^l N\|_{L_G^2} + \|\Delta^l N\|_{L_G^2} \lesssim \sum_{I=1}^{2l-1} \|\nabla^I F * \nabla^{2l-I} N\|_{L_G^2} + \|\Delta^l F\|_{L_G^2}.$$

Hence, using (4-13c) (replacing l with $l-1$) and (3-17h), (4-4a) and (4-4b) to estimate low-order terms, we get

$$\begin{aligned} a^4\|\Delta^{l+1}N\|_{L_G^2} + a^2\|\nabla\Delta^l N\|_{L_G^2} + \|\Delta^l N\|_{L_G^2} \\ \lesssim \|\Delta^l F\|_{L_G^2} + \varepsilon a^{-c\sqrt{\varepsilon}}\Big(\sum_{m=0}^{l-1}\|\nabla\Delta^m N\|_{L_G^2} + \varepsilon a^{4-c\sigma}\sqrt{\mathcal{E}^{(\leq 2l-4)}(\mathrm{Ric},\cdot)}\Big) \\ + \varepsilon a^{4-c\sigma}\big(\|F\|_{H_G^{2(l-1)}} + \|\nabla\Delta^{l-1}F\|_{L_G^2} + \varepsilon a^{-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq 2l-4)}(\mathrm{Ric},\cdot)}\big). \end{aligned}$$

For the top-order lapse term, we can redistribute the divergent prefactor as follows:

$$\varepsilon a^{-c\sqrt{\varepsilon}}\|\nabla\Delta^{l-1}N\|_{L_G^2} \lesssim \varepsilon\|\Delta^l N\|_{L_G^2} + \varepsilon a^{-2c\sqrt{\varepsilon}}\|\Delta^{l-1}N\|_{L_G^2}.$$

The lower-order lapse terms, as well as $\|\nabla\Delta^{l-1}F\|_{L_G^2}$, can be estimated similarly, just without having to redistribute the prefactor. Updating $c > 0$ and rearranging then yields the statement at order l for suitably small $\varepsilon > 0$, and thus the entire statement after iteration. □

Corollary 5.5 (lapse energy estimates with $\mathcal{L}$). *For any $l \in \{0, \dots, 9\}$, one has*

$$\begin{aligned} &a^8\mathcal{E}^{(2(l+1))}(N,\cdot) + a^4\mathcal{E}^{(2l+1)}(N,\cdot) + \mathcal{E}^{(2l)}(N,\cdot) \\ &\quad\lesssim \varepsilon^2\mathcal{E}^{(2l)}(\Sigma,\cdot) + \mathcal{E}^{(2l)}(\phi,\cdot) \\ &\qquad + \underbrace{\varepsilon^2 a^{-c\sqrt{\varepsilon}}[\mathcal{E}^{(\leq 2(l-1))}(\Sigma,\cdot) + \mathcal{E}^{(\leq 2(l-1))}(\phi,\cdot)]}_{\text{not present for } l=0} + \underbrace{(\varepsilon^4 a^{-c\sqrt{\varepsilon}} + \varepsilon^2 a^{8-c\sigma})\mathcal{E}^{(\leq 2l-3)}(\mathrm{Ric},\cdot)}_{\text{not present for } l\leq 1}. \end{aligned}$$

Proof. Note that, by Corollary 5.4, all that needs to be done is to relate all Sobolev norms of F that occur to the respective energies. Schematically, we have

$$\Delta^l F = 16\pi(\Delta^l\Psi)(\Psi + C) + 2\langle\Delta^l\Sigma, \Sigma\rangle_G + \sum_{I=1}^{2l-1}(\nabla^I\Psi * \nabla^{2l-I}\Psi + \nabla^I\Sigma * \nabla^{2l-I}\Sigma).$$

For the first two terms, we can use (4-2a) and (4-2b) to bound $|\Sigma|_G$ and $|\Psi + C|$ by ε and 1 up to constant, respectively. For the remaining terms, we similarly always bound the lower order in L^∞_G with (4-4a)–(4-4b) and bound the higher order with the energy estimates in Lemma 4.5. Further, we can use (3-8) to redistribute divergent prefactors onto energies of order $l-2$ and lower. This already incurs the terms on the right-hand side of the claimed estimate, and the lower-order norms of F only incur at equivalent or weaker error terms. □

5.2. ***Elliptic lapse estimates with $\widetilde{\mathcal{L}}$.*** While the estimates in the previous subsection are useful at high orders, they are not enough to close the bootstrap assumptions for N. This can be achieved by deriving estimates in terms of $\widetilde{\mathcal{L}}$ — however, due to the explicit presence of Ricci terms in this version of the lapse equation, we use this to bound N at lower orders. Since the arguments are largely identical to the ones above, we only sketch the proofs.

Remark 5.6. Note that when replacing f by $\tilde{f}$ and F by $\widetilde{F}$ in Remark 5.2, the same statements hold for a suitable constant K. In fact, the bootstrap assumptions on $\mathrm{Ric}[G]$ and $\nabla\phi$ even imply $\|\widetilde{F}\|_{C^1_G} \lesssim \varepsilon a^{4-c\sigma}$, noting

$$\big|R[G] + \tfrac{2}{3}\big|_G \leq |G^{-1}|_G\big|\mathrm{Ric}[G] + \tfrac{2}{9}G\big|_G \lesssim \big|\mathrm{Ric}[G] + \tfrac{2}{9}G\big|_G$$

and

$$|\nabla R[G]|_G = \big|\nabla\big(R[G] + \tfrac{2}{3}\big)\big|_G \lesssim \big|\nabla\big(\mathrm{Ric}[G] + \tfrac{2}{9}G\big)\big|_G.$$

Lemma 5.7. *Any scalar functions ζ and Z such that*

$$\widetilde{\mathcal{L}}\zeta = Z$$

holds satisfy the estimate

$$a^4\|\Delta\zeta\|_{L^2_G} + a^2\|\nabla\zeta\|_{L^2_G} + \|\zeta\|_{L^2_G} \lesssim \|Z\|_{L^2_G}.$$

Proof. The proof follows identically to Lemma 5.3 since all tools relating to f and F used in proving these statements were collected in Remark 5.2, and these extend to $\widetilde{\mathcal{L}}$ by Remark 5.6. □

Corollary 5.8. *For $l \in \{0, \dots, 8\}$,*

$$a^8\mathcal{E}^{(2(l+1))}(N,\cdot) + a^4\mathcal{E}^{(2l+1)}(N,\cdot) + \mathcal{E}^{(2l)}(N,\cdot) \lesssim a^8\mathcal{E}^{(\leq 2l)}(\mathrm{Ric},\cdot) + \varepsilon a^{8-c\sqrt{\varepsilon}}\|\nabla\phi\|^2_{H^{2l}_G}.$$

Proof. As in the proof of Corollary 5.4, this follows by commuting $\widetilde{\mathcal{L}}$ with Δ^l iteratively and applying (4-4e) and (4-4f) to bound lower-order terms within the nonlinearities. □

6. Big bang stability: energy and norm estimates

In this section, we derive energy estimates for matter variables and the geometric quantities, as well as Sobolev norm estimates for spatial derivatives of ϕ and for metric quantities. To derive all of the inequalities in this section beside the elliptic inequality in Lemma 6.10 and the bound on $\nabla\phi$ in Lemma 6.5, we will use the same basic strategy. Hence, we give a brief overview on the form our integral inequalities are going to take and how we intend to obtain improved energy bounds from there:

Remark 6.1 (integral inequalities and the Gronwall argument). Let $\mathcal{F}_L$ denote an energy or a squared Sobolev(-type) norm at derivative level $L \in 2\mathbb{N}$, for example $\mathcal{E}^{(L)}(\phi, \cdot)$. To derive an integral inequality for $\mathcal{F}_L$, we will take its time derivative, apply the respective commuted evolution equations in the integrand, estimate the resulting terms and integrate that inequality. Schematically, the resulting integral inequalities for $\mathcal{F}_L$ then take the following form:

$$\begin{aligned}\mathcal{F}_L(t) + \int_t^{t_0} &\langle \text{ultimately nonnegative contributions}\rangle \, ds \\ &\lesssim \mathcal{F}_L(t_0) + \int_t^{t_0} (\varepsilon^{1/8} a(s)^{-3} + a(s)^{-1-c\sqrt{\varepsilon}}) \mathcal{F}_L(s) \, ds \\ &\quad + \int_t^{t_0} a(s)^{-3} \langle \text{other energies/squared Sobolev norms at same derivative level}\rangle \, ds \\ &\quad + \int_t^{t_0} a(s)^{-3-c\sqrt{\varepsilon}} \langle \text{energies/squared Sobolev norms at derivative levels up to } L-2\rangle \, ds.\end{aligned}$$

For some inequalities, we will not be able to derive any beneficial ε-prefactors in the penultimate line. For example, for $\mathcal{E}^{(L)}(\Sigma, \cdot)$, linear lapse terms in the evolution of Σ incur a term of the form

$$\int_t^{t_0} a(s)^{-3} \cdot a(s)^4 \| \Delta^{L/2} N \|_{\dot{H}^2_G} \cdot \sqrt{\mathcal{E}^{(L)}(\Sigma, s)} \, ds$$

on the right-hand side, which after applying lapse energy estimates creates $\varepsilon^{-1/8} \mathcal{E}^{(L)}(\phi, \cdot)$ on the right. However, combining the respective inequalities for the core energy mechanism at each derivative level with appropriate ε-weights, this will then combine to an inequality of the following form for a total energy, which we informally denote by $\mathcal{F}_{\text{total},L}$:

$$\begin{aligned}\mathcal{F}_{\text{total},L}(t) + \int_t^{t_0} &\langle \text{nonnegative quantity}\rangle \, ds \\ &\lesssim \mathcal{F}_{\text{total},L}(t_0) + \int_t^{t_0} (\varepsilon^{1/8} a(s)^{-3} + a(s)^{-1-c\sqrt{\varepsilon}}) \mathcal{F}_{\text{total},L}(s) \, ds \\ &\quad + \underbrace{\int_t^{t_0} \varepsilon^{1/8} a(s)^{-3-c\varepsilon^{1/8}} \langle \text{already improved terms}\rangle \, ds}_{\text{not present for } L=0} + \varepsilon^{1/4} \mathcal{F}_{\text{total},L}(t) \\ &\quad + \underbrace{\sqrt{\varepsilon} \cdot \langle \text{small lower-order terms}\rangle(t)}_{\text{not present for } L=0}. \qquad (6\text{-}1)\end{aligned}$$

In the mentioned example, multiplying $\mathcal{E}^{(L)}(\Sigma, \cdot)$ with the weight $\varepsilon^{1/4}$, in turn, mitigates the otherwise offending term to $\varepsilon^{1/8} a(s)^{-3} \mathcal{E}^{(L)}(\phi, s)$, which can be absorbed into the first line.

Furthermore, $\varepsilon^{1/4}\mathcal{F}_{\text{total},L}(t)$ in the penultimate line of (6-1) can be absorbed into the left-hand side after updating the implicit constant in "$\lesssim$". Applying the Gronwall lemma (see Lemma A.1) and the initial data assumption (which implies $\mathcal{F}_{\text{total},L}(t_0) \lesssim \varepsilon^4$) then yields

$$\mathcal{F}_{\text{total},L}(t) \lesssim \Big(\varepsilon^4 + \underbrace{\int_t^{t_0} \varepsilon^{1/8} a(s)^{-3-c\varepsilon^{1/8}} \langle \text{already improved terms} \rangle \, ds + \sqrt{\varepsilon} \cdot \langle \text{lower-order terms} \rangle (t)}_{\text{not present for } L=0}\Big) \cdot \exp\Big(K \cdot \int_t^{t_0} \varepsilon^{1/8} a(s)^{-3} + a(s)^{-1-c\sqrt{\varepsilon}} \, ds\Big)$$

for some constant $K > 0$. By (2-7) and (2-6), the exponential factor can be bounded by $a^{-c\varepsilon^{1/8}}$, up to constant and updating $c > 0$.

Hence, for $L = 0$, this implies $\mathcal{F}_{\text{total},0} \lesssim \varepsilon^4 a^{-c\varepsilon^{1/8}}$, and thus leads to improved bounds for base level energy quantities (see Remark 3.19 for the precise scaling hierarchy that will achieve). By iterating this argument for $L > 0$, the already improved terms will then be bounded (at worst) by $\varepsilon^4 a^{-c\varepsilon^{1/8}}$, and (2-6) shows that the first line can be bounded by $\varepsilon^4 a^{-c\varepsilon^{1/8}}$ after updating c. This allows us to bound $\mathcal{F}_{\text{total},L}$ by $\varepsilon^4 a^{-c\varepsilon^{1/8}}$ for any L up to and including top order.

Finally, we mention that, to control energies at order L, we need to consider scaled energies at order $L+1$ within $\mathcal{F}_{\text{total},L}$ — this arises since the scalar field occurs at first order in the evolution equations for $\boldsymbol{E}$ and $\boldsymbol{B}$. We avoid losing derivatives by employing the div-curl-estimate in Lemma 6.10 at order $L+1$, which allows us to control $a^4\mathcal{E}^{(L+1)}(\Sigma, \cdot)$ by quantities at order L. This is precisely what generates the nonintegral terms in the schematics above. We note that it is crucial that the scalar field occurs at no worse scaling than a^{-1} in (2-38a)–(2-38b) — else, moving to these time-scaled estimates at order $L+1$ would lose too many powers of a and lead to exponentially divergent terms after applying the Gronwall argument.

Recall that L^2_G-norm estimates for error terms arising in the Laplace-commuted equations in Lemma 2.11 are collected in Section A.4. Low-order estimates (in particular estimates for $L = 2$) could often be improved if needed by more carefully avoiding curvature error terms, but we refrain from doing so where it is not necessary to keep estimates as unified as possible.

6.1. *Integral and energy estimates for the scalar field.*

6.1.1. *Scalar field energy estimates.* Over the following two lemmas, we prove the core energy estimates to control the matter variables, which are immediately prepared differently at base, intermediate and top order for the total energy estimates in Section 7.

Lemma 6.2 (even-order scalar field energy estimates). *Let $t \in (t_{\text{Boot}}, t_0]$. Then, one has*

$$\mathcal{E}^{(0)}(\phi, t) + \int_t^{t_0} \dot{a}(s) a(s)^3 \mathcal{E}^{(1)}(N, s) + \frac{\dot{a}(s)}{a(s)} \mathcal{E}^{(0)}(N, s) \, ds \lesssim \mathcal{E}^{(0)}(\phi, t_0) + \int_t^{t_0} \varepsilon a(s)^{-3} \mathcal{E}^{(0)}(\phi, s) + \varepsilon a^{-3} \mathcal{E}^{(0)}(\Sigma, s) \, ds. \quad (6\text{-}2)$$

Further, for any $L \in 2\mathbb{N},\ 2 \le L \le 18$, *the following estimate is satisfied*:

$$\begin{aligned}\mathcal{E}^{(L)}(\phi,t)+\int_t^{t_0}\dot{a}(s)a(s)^3\mathcal{E}^{(L+1)}(N,s)+\frac{\dot{a}(s)}{a(s)}\mathcal{E}^{(L)}(N,s)\,ds&\\ \lesssim \mathcal{E}^{(L)}(\phi,t_0)+\int_t^{t_0}(\varepsilon a(s)^{-3}+a(s)^{-1-c\sqrt{\varepsilon}})\,\mathcal{E}^{(L)}(\phi,s)\,ds&\\ +\int_t^{t_0}\varepsilon a(s)^{-3}\mathcal{E}^{(L)}(\Sigma,s)+\varepsilon^{3/2}a(s)^{-3}\mathcal{E}^{(L-2)}(\mathrm{Ric},s)\,ds&\\ +\int_t^{t_0}\sqrt{\varepsilon}a(s)^{-3-c\sqrt{\varepsilon}}\mathcal{E}^{(\le L-2)}(\phi,s)+\varepsilon a(s)^{-3-c\sqrt{\varepsilon}}\mathcal{E}^{(\le L-2)}(\Sigma,s)\,ds&\\ \underbrace{+\int_t^{t_0}\varepsilon^{3/2}a(s)^{-3-c\sqrt{\varepsilon}}\mathcal{E}^{(\le L-4)}(\mathrm{Ric},s)\,ds}_{\text{if } L\ge 4}.&\end{aligned}\tag{6-3}$$

Remark 6.3. This proof relies on two mechanisms: Firstly, we use the structure of the wave equation and integration by parts to cancel the highest-order scalar field derivative terms. Getting this cancellation is what necessitates scaling the potential term in the scalar field energy by a^4. Secondly, we deal with the highest-order lapse terms using the elliptic structure of the (Laplace-commuted) lapse equation — both in an indirect way by invoking the elliptic energy estimate in Corollary 5.5, as well as by directly inserting (2-37a) to cancel some ill-behaved terms. While the framework significantly differs from the scalar field energy estimates [Speck 2018], these two core mechanisms also appear there and play similarly crucial roles.

Proof. Since the arguments are essentially the same, we will only write down the proof, for $L \ge 2$, in full and make short comments throughout the argument which terms do not occur for $L = 0$.

We use the evolution equations (2-39a) and (2-39b) and Lemma 4.4 to compute, for $L \ge 2$,

$$\begin{aligned}-\partial_t\mathcal{E}^{(L)}(\phi,\cdot)=\int_M&-2\partial_t\Delta^{L/2}\Psi\cdot\Delta^{L/2}\Psi-2a^4\langle\partial_t\nabla\Delta^{L/2}\phi,\nabla\Delta^{L/2}\phi\rangle_G\\ &-a^4(\partial_tG^{-1})^{ij}\nabla_i\Delta^{L/2}\phi\nabla_j\Delta^{L/2}\phi\\ &-3N\frac{\dot{a}}{a}[|\Delta^{L/2}\Psi|^2+a^4|\nabla\Delta^{L/2}\phi|_G^2]-4\frac{\dot{a}}{a}\cdot a^4|\nabla\Delta^{L/2}\phi|_G^2\,\mathrm{vol}_G\\ =\int_M&\Big(-2a(N+1)\Delta^{L/2+1}\phi-2a\langle\nabla\Delta^{L/2}N,\nabla\phi\rangle_G+6C\frac{\dot{a}}{a}\Delta^{L/2}N\Big)\cdot(\Delta^{L/2}\Psi)&\text{(6-4a)}\\ &-2a(N+1)\langle\nabla\Delta^{L/2}\Psi,\nabla\Delta^{L/2}\phi\rangle_G-2Ca\langle\nabla\Delta^{L/2}N,\nabla\Delta^{L/2}\phi\rangle_G&\text{(6-4b)}\\ &-2(\mathfrak{P}_{L,\mathrm{Border}}+\mathfrak{P}_{L,\mathrm{Junk}})\cdot\Delta^{L/2}\Psi-2a^4\langle\mathfrak{Q}_{L,\mathrm{Border}}+\mathfrak{Q}_{L,\mathrm{Junk}},\nabla\Delta^{L/2}\phi\rangle_G&\text{(6-4c)}\\ &+2(N+1)a\cdot(\Sigma^\sharp)^{ij}\nabla_i\Delta^{L/2}\phi\nabla_j\Delta^{L/2}\phi-2N\frac{\dot{a}}{a}\cdot a^4|\nabla\Delta^{L/2}\phi|_G^2&\text{(6-4d)}\\ &-3N\frac{\dot{a}}{a}[|\Delta^{L/2}\Psi|^2+a^4|\nabla\Delta^{L/2}\phi|_G^2]-4\frac{\dot{a}}{a}\cdot a^4|\nabla\Delta^{L/2}\phi|_G^2\,\mathrm{vol}_G.&\text{(6-4e)}\end{aligned}$$

Note that, for $L = 0$, the equivalent equality holds where the borderline and junk terms are replaced by $-2a^4\Psi\langle\nabla N,\nabla\phi\rangle_G$ (to verify this, insert (2-32a) and (2-32b) instead of (2-39a) and (2-39b)). We now go through (6-4a)–(6-4e) term by term:

After integrating by parts, the first term in (6-4a) reads

$$\int_M 2a(N+1)\langle \nabla\Delta^{L/2}\phi, \nabla\Delta^{L/2}\Psi\rangle_G + 2a\langle \nabla N, \nabla\Delta^{L/2}\phi\rangle_G \cdot \Delta^{L/2}\Psi \operatorname{vol}_G. \tag{6-5}$$

The first term *precisely* cancels the first term in (6-4b), while we can use the bootstrap assumption (3-17h) to estimate the other term in (6-5) up to constant by

$$\varepsilon a^{3-c\sigma}\cdot a^2 \|\nabla\Delta^{L/2}\phi\|_{L^2_G}\|\Delta^{L/2}\Psi\|_{L^2_G} \lesssim \varepsilon a^{3-c\sigma}\mathcal{E}^{(L)}(\phi,\cdot).$$

For the second term in (6-4a), we use (4-4e) to estimate $\nabla\phi$ and Corollary 5.5 at order L to deal with the lapse, getting

$$\begin{aligned}\Big|\int_M 2a\langle \nabla\Delta^{L/2}N, \nabla\phi\rangle_G \cdot \Delta^{L/2}\Psi \operatorname{vol}_G\Big| &\lesssim \sqrt{\varepsilon}a^{-1-c\sqrt{\varepsilon}}\sqrt{a^4\mathcal{E}^{(L+1)}(N,\cdot)}\sqrt{\mathcal{E}^{(L)}(\phi,\cdot)}\\ &\lesssim \varepsilon a^{-1}\cdot a^4\mathcal{E}^{(L+1)}(N,\cdot) + a^{-1-c\sqrt{\varepsilon}}\mathcal{E}^{(L)}(\phi,\cdot)\\ &\lesssim \varepsilon^3 a^{-1}\mathcal{E}^{(L)}(\Sigma,\cdot) + a^{-1-c\sqrt{\varepsilon}}\mathcal{E}^{(L)}(\phi,\cdot)\\ &\quad + \varepsilon^2 a^{-1-c\sqrt{\varepsilon}}\mathcal{E}^{(\le L-2)}(\Sigma,\cdot) + \varepsilon^2 a^{-1-c\sqrt{\varepsilon}}\mathcal{E}^{(\le L-2)}(\phi,\cdot)\\ &\quad + \underbrace{\varepsilon^2 a^{-1-c\sqrt{\varepsilon}}\mathcal{E}^{(\le L-3)}(\mathrm{Ric},\cdot)}_{\text{not present for } L=2}.\end{aligned}$$

Repeating this argument for $L=0$, the last two lines do not appear.

To deal with the remaining term in (6-4a), we can insert the following zero on the right-hand side of the differential equality, where the equality (6-6) holds due to (2-37a):

$$\begin{aligned}0 &= -\frac{3}{8\pi}\dot{a}a^3\int_M \operatorname{div}_G(\nabla\Delta^{L/2}N\cdot\Delta^{L/2}N)\operatorname{vol}_G\\ &= -\frac{3}{8\pi}\dot{a}a^3\int_M \Delta^{L/2+1}N\cdot\Delta^{L/2}N + |\nabla\Delta^{L/2}N|^2_G \operatorname{vol}_G\\ &= \int_M -\frac{3}{8\pi}\dot{a}a^3|\nabla\Delta^{L/2}N|^2_G - \frac{3}{8\pi}\Big(\frac{\dot{a}a^3}{3}+12\pi C^2\frac{\dot{a}}{a}\Big)|\Delta^{L/2}N|^2\\ &\qquad - 6C\frac{\dot{a}}{a}\Delta^{L/2}N\cdot\Delta^{L/2}\Psi - \frac{3}{8\pi}\dot{a}a^3[\mathfrak{N}_{L,\mathrm{Border}}+\mathfrak{N}_{L,\mathrm{Junk}}]\cdot\Delta^{L/2}N\operatorname{vol}_G.\end{aligned}\tag{6-6}$$

Note that the first line has a negative sign, so (after absorbing a few terms into it without changing the sign, see namely lapse quantities in (6-7) and (6-8a)) we pull it to the left-hand side of the differential inequality. Further, the first term in the second line of (6-6) precisely cancels the third term in (6-4a). That leaves the borderline and junk terms in (6-6), for which we use (A-17b) and (A-19d) (along with $\dot{a}\simeq a^{-2}$ due to (2-3)) to get, for $L\ge 4$,

$$\begin{aligned}&\frac{3}{8\pi}\dot{a}a^3\Big|\int_M[\mathfrak{N}_{L,\mathrm{Border}}+\mathfrak{N}_{L,\mathrm{Junk}}]\cdot\Delta^{L/2}N\operatorname{vol}_G\Big|\\ &\lesssim \varepsilon a^{-3}[\mathcal{E}^{(L)}(\phi,\cdot)+\mathcal{E}^{(L)}(\Sigma,\cdot)+\mathcal{E}^{(L)}(N,\cdot)] + \varepsilon a^{-3-c\sqrt{\varepsilon}}[\mathcal{E}^{(\le L-2)}(\phi,\cdot)+\mathcal{E}^{(\le L-2)}(\Sigma,\cdot)+\mathcal{E}^{(\le L-2)}(N,\cdot)]\\ &\qquad + \underbrace{\varepsilon^3 a^{-3}\mathcal{E}^{(\le L-2)}(\mathrm{Ric},\cdot)+\varepsilon^3 a^{-3-c\sqrt{\varepsilon}}\mathcal{E}^{(\le L-3)}(\mathrm{Ric},\cdot)}_{\text{not present for } L=2}.\end{aligned}$$

Again, the same estimate holds for $L=0$ with the last two lines dropped.

From (6-4a)–(6-4b), only the term $-2Ca\langle\nabla\Delta^{L/2}N, \nabla\Delta^{L/2}\phi\rangle_G$ still needs to be handled: Using the inequality (2-8) arising from the Friedman equation, we can estimate this by

$$\int_M 2\sqrt{\tfrac{3}{4\pi}}\dot{a}a^3|\nabla\Delta^{L/2}N|_G|\nabla\Delta^{L/2}\phi|_G \operatorname{vol}_G \leq \int_M 4\dot{a}a^3|\nabla\Delta^{L/2}\phi|_G^2 + \tfrac{3}{16\pi}\dot{a}a^3|\nabla\Delta^{L/2}N|_G^2 \operatorname{vol}_G. \quad (6\text{-}7)$$

Note that the first term precisely cancels the final term in (6-4e), while the second term can be absorbed into the first term in (6-6) while preserving that term's sign.

To bound the error terms in (6-4c), we insert the borderline term estimates (A-17d) and (A-17e), as well as the junk term estimates (A-19f) and (A-19h), where (3-8) is used to estimate odd order by even-order energies where needed. Furthermore, observe that we can estimate the $\mathfrak{Q}_L$-terms as

$$(a^2\|\mathfrak{Q}_L\|_{L^2_G}) \cdot \sqrt{\mathcal{E}^{(L)}(\phi, \cdot)},$$

so all borderline and junk terms arising from it, beside the scalar field energies, are dominated by terms occurring elsewhere.

Finally, all terms that remain, namely (6-4d) and the first term in (6-4e), can be bounded by $\varepsilon a^{-3}\mathcal{E}^{(L)}(\phi, \cdot)$ due to the strong base level estimate (4-2b) and (3-17h). In summary, and always only keeping the worst terms for each energy and squared norm, this yields, for $L \geq 4$,

$$\begin{aligned}
&-\partial_t\mathcal{E}^{(L)}(\phi, \cdot) + \dot{a}a^3\mathcal{E}^{(L+1)}(N, \cdot) + \frac{\dot{a}}{a}\mathcal{E}^{(L)}(N, \cdot) \\
&\quad\lesssim (\varepsilon a^{-3} + a^{-1-c\sqrt{\varepsilon}})\mathcal{E}^{(L)}(\phi, \cdot) + (\varepsilon a^{-3} + \sqrt{\varepsilon}a^{-1-c\sqrt{\varepsilon}})(a^4\mathcal{E}^{(L+1)}(N, \cdot) + \mathcal{E}^{(L)}(N, \cdot)) && (6\text{-}8a)\\
&\qquad + \varepsilon a^{-3}\mathcal{E}^{(L)}(\Sigma, \cdot) + \varepsilon^{3/2}a^{-3}\mathcal{E}^{(L-2)}(\mathrm{Ric}, \cdot) + \varepsilon a^{-3-c\sqrt{\varepsilon}}\mathcal{E}^{(\leq L-2)}(\phi, \cdot) && (6\text{-}8b)\\
&\qquad\quad + \varepsilon a^{-3-c\sqrt{\varepsilon}}\mathcal{E}^{(\leq L-2)}(\Sigma, \cdot) + [\varepsilon a^{-3-c\sqrt{\varepsilon}} + \sqrt{\varepsilon}a^{-1-c\sqrt{\varepsilon}}]\mathcal{E}^{(\leq L-2)}(N, \cdot) && (6\text{-}8c)\\
&\qquad\quad \underbrace{+\varepsilon^{3/2}a^{-3-c\sqrt{\varepsilon}}\mathcal{E}^{(\leq L-4)}(\mathrm{Ric}, \cdot)}_{\text{not present for } L=2}. && (6\text{-}8d)
\end{aligned}$$

The lapse energies in (6-8a) can now also be absorbed into those on the left-hand side of the inequality by updating the implicit constant in "$\lesssim$". We can treat the lower-order lapse energies in (6-8c) with Corollary 5.5 and see that the resulting terms are all dominated by terms we already have on the right-hand side of the inequality above.

Inserting these estimates and integrating over $(t, t_0]$ then yields (6-3) for $L \geq 4$, and the statement for $L = 2$ is obtained completely analogously.

As mentioned earlier, (6-4c) is replaced by the following term for $L = 0$:

$$\begin{aligned}
\int_M -2a\Psi\langle\nabla N, \nabla\phi\rangle_G \operatorname{vol}_G &\lesssim \varepsilon a^{-3}\int_M a^2|\nabla N|_G \cdot a^2|\nabla\phi|_G \operatorname{vol}_G \\
&\lesssim \varepsilon\dot{a}a^3\mathcal{E}^{(1)}(N, \cdot) + \varepsilon a^{-3}\mathcal{E}^{(0)}(\phi, \cdot).
\end{aligned}$$

Here, we applied (4-2a) and (2-3). Both of these terms can be absorbed into terms that are already present, and (6-2) then follows by dealing with terms in $\partial_t\mathcal{E}^{(0)}(\phi, \cdot)$ as described and integrating. □

To close the argument, we will need a scaled scalar field energy estimate at the odd orders $L+1$, which is not covered by the previous lemma and we hence establish separately:

Lemma 6.4 (odd-order scalar field energy estimate). *For $L \in 2\mathbb{N}$, $2 \leq L \leq 18$, we have*

$$\begin{aligned}
&a(t)^4\mathcal{E}^{(L+1)}(\phi,t) + \int_t^{t_0}\{\dot{a}(s)a(s)^7\mathcal{E}^{(L+2)}(N,s) + \dot{a}(s)a(s)^3\mathcal{E}^{(L+1)}(N,s)\}\,ds \\
&\quad\lesssim a(t_0)^4\mathcal{E}^{(L+1)}(\phi,t_0) + \int_t^{t_0}(\varepsilon a(s)^{-3} + a(s)^{-1-c\sqrt{\varepsilon}})\cdot a(s)^4\mathcal{E}^{(L+1)}(\phi,s)\,ds \\
&\qquad + \int_t^{t_0}\Big\{\varepsilon a(s)^{-3}\cdot a(s)^4\mathcal{E}^{(L+1)}(\Sigma,s) + (\varepsilon a(s)^{-3} + a(s)^{-1-c\sqrt{\varepsilon}})\mathcal{E}^{(L)}(\phi,s) + \varepsilon a(s)^{-3}\mathcal{E}^{(L)}(\Sigma,s) \\
&\qquad\qquad + \varepsilon a(s)^{-1-c\sqrt{\varepsilon}}\cdot a(s)^4\mathcal{E}^{(L-1)}(\mathrm{Ric},s) + (\varepsilon^3 a^{-3} + \varepsilon a^{-1-c\sqrt{\varepsilon}})\mathcal{E}^{(L-2)}(\mathrm{Ric},s) \\
&\qquad\qquad + \varepsilon a(s)^{-3-c\sqrt{\varepsilon}}(\mathcal{E}^{(\leq L-2)}(\phi,s) + \mathcal{E}^{(\leq L-2)}(\Sigma,s)) \\
&\qquad\qquad \underbrace{+(\varepsilon^3 a(s)^{-3-c\sqrt{\varepsilon}} + \varepsilon^2 a(s)^{-1-c\sqrt{\varepsilon}})\mathcal{E}^{(\leq L-4)}(\mathrm{Ric},s)}_{\text{not present for } L=2}\Big\}\,ds\,. \qquad (6\text{-}9)
\end{aligned}$$

At order 1, *the analogous estimate holds where the last three lines of* (6-9) *are dropped.*

Proof. These estimates follow completely analogously to Lemma 6.2, with the exception that high-order lapse terms can now be estimated at order $L+2$ due to the scalar field energy being scaled by a^4. In particular, we note that to deal with the analogous term to (6-4a), one now inserts the following zero on the right and applies the commuted lapse equation (2-37b):

$$\begin{aligned}
0 &= -\frac{3}{8\pi}\dot{a}a^7\int_M \mathrm{div}_G(\nabla\Delta^{L/2}N\cdot\Delta^{L/2+1}N)\,\mathrm{vol}_G \\
&= \int_M\Big\{-\frac{3}{8\pi}\dot{a}a^7|\Delta^{L/2+1}N|^2 - \frac{3}{8\pi}\Big(\frac{1}{3}\dot{a}a^3 + 12\pi C^2\frac{\dot{a}}{a}\Big)\cdot a^4|\nabla\Delta^{L/2}N|_G^2 \\
&\qquad - 6C\frac{\dot{a}}{a}\cdot a^4\langle\nabla\Delta^{L/2}N, \nabla\Delta^{L/2}\Psi\rangle_G - \frac{3}{8\pi}\dot{a}a^7\langle\mathfrak{N}_{L+1,\mathrm{Border}} + \mathfrak{N}_{L+1,\mathrm{Junk}}, \nabla\Delta^{L/2}N\rangle_G\Big\}\,\mathrm{vol}_G.
\end{aligned}$$

For $L=0$, the argument is again the same as at higher orders with less complicated junk terms. We briefly highlight some specific junk terms: The term analogous to (6-4c) is now estimated as follows using (4-13a):

$$\begin{aligned}
a^4\cdot\int_M -2a\langle\nabla\Psi\nabla N, \nabla^2\phi\rangle_G &\lesssim \varepsilon\int_M a^{1/2}|\nabla N|_G\cdot a^{1/2-c\sqrt{\varepsilon}}\cdot a^4|\nabla\phi|_G \\
&\lesssim \varepsilon\dot{a}a^3\mathcal{E}^{(1)}(N,\cdot) + \varepsilon a^{1-c\sqrt{\varepsilon}}\cdot a^4\mathcal{E}^{(\leq 1)}(\phi,\cdot).
\end{aligned}$$

Further, note that, by the commutator formula (A-8c) and applying (4-4e), one has

$$\begin{aligned}
\Big|\int_M a^8[\partial_t,\Delta]\phi\cdot\Delta\phi\,\mathrm{vol}_G\Big| &\lesssim \varepsilon a^{5-c\sqrt{\varepsilon}}(\|\nabla\Sigma\|_{L^2_G} + \|\nabla N\|_{L^2_G})\|\Delta\phi\|_{L^2_G} \\
&\lesssim \varepsilon a^{-1-c\sqrt{\varepsilon}}(a^4\mathcal{E}^{(1)}(\phi,\cdot) + a^4\mathcal{E}^{(1)}(\Sigma,\cdot)) + \varepsilon a^{6-c\sigma}\cdot\dot{a}a^3\mathcal{E}^{(1)}(N,\cdot). \qquad \square
\end{aligned}$$

6.1.2. *Sobolev norm estimate for $\nabla\phi$.* To improve the bootstrap assumptions on $\nabla\phi$, we will need sharper bounds than those on $a^4\|\nabla\phi\|^2_{H^L}$ that will follow from bounds on $\mathcal{E}^{(L)}(\phi,\cdot)$:

Lemma 6.5. *Let $l \in (t_{\mathrm{Boot}}, t_0]$. Then, for $l \in \mathbb{Z}_+$, $l \leq 17$, the following estimate holds*:

$$\|\nabla\phi\|_{H^l_G(\Sigma_t)} \lesssim (1+\varepsilon a(t)^{-c\sqrt{\varepsilon}})\|\Sigma\|_{H^{l+1}_G(\Sigma_t)} + \varepsilon a(t)^{-c\sqrt{\varepsilon}}\|\Psi\|_{H^l_G(\Sigma_t)}.$$

Proof. By (4-2a), $\Psi + C > \frac{1}{2}C$ holds if ε is chosen small enough. Consequently, we can rearrange (2-29b) and apply the product rule to obtain

$$|\nabla^l \nabla\phi|_G = \frac{1}{8\pi}\left|\nabla^l\left(\frac{\operatorname{div}_G \Sigma}{\Psi + C}\right)\right|_G \lesssim \sum_{I_\Sigma + I_\Psi = l} |\nabla^{I_\Sigma+1}\Sigma|_G |\nabla^{I_\Psi}(\Psi + C)|_G.$$

The statement then follows by integrating over Σ_t and applying (4-2a) and (4-2b). □

6.2. ***Energy estimates for the Bel–Robinson variables.*** In this subsection, we collect the energy estimates for the Bel–Robinson variables:

Lemma 6.6 (Bel–Robinson energy estimates). *Let $t \in (t_{\mathrm{Boot}}, t_0]$. Then one has*

$$\begin{aligned}
&\mathcal{E}^{(0)}(W,t) + \int_t^{t_0}\int_M \Big[8\pi C^2 a(s)^{-3}(N+1)\langle \Sigma, \boldsymbol{E}\rangle_G + 6\frac{\dot a(s)}{a(s)}(N+1)|\boldsymbol{E}|_G^2\Big]\operatorname{vol}_G\, ds \\
&\quad\lesssim \mathcal{E}^{(0)}(W,t_0) + \int_t^{t_0}(\varepsilon a(s)^{-3} + a(s)^{-1-c\sqrt{\varepsilon}})\mathcal{E}^{(0)}(W,s) + \varepsilon^{-1/8}a(s)^{-3}\cdot a(s)^4\mathcal{E}^{(1)}(\phi,s)\,ds \\
&\qquad\qquad + \int_t^{t_0} a(s)^{-1-c\sqrt{\varepsilon}}\mathcal{E}^{(0)}(\phi,s) + \varepsilon a(s)^{-3}\mathcal{E}^{(0)}(\Sigma,s)\,ds, \qquad (6\text{-}10)
\end{aligned}$$

as well as, for $L \in 2\mathbb{N}$, $2 \le L \le 18$,

$$\begin{aligned}
&\mathcal{E}^{(L)}(W,t) + \int_t^{t_0}\int_M \Big[8\pi C^2 a(s)^{-3}(N+1)\langle \Delta^{L/2}\Sigma, \Delta^{L/2}\boldsymbol{E}\rangle_G + 6(N+1)\frac{\dot a(s)}{a(s)}|\Delta^{L/2}\boldsymbol{E}|_G^2\Big]\operatorname{vol}_G\, ds \\
&\lesssim \mathcal{E}^{(L)}(W,t_0) + \int_t^{t_0}(\varepsilon^{1/8}a(s)^{-3} + a(s)^{-1-c\sqrt{\varepsilon}})\mathcal{E}^{(L)}(W,s)\,ds \\
&\quad + \int_t^{t_0}\big\{\varepsilon^{-1/8}a(s)^{-3}\cdot a(s)^4\mathcal{E}^{(L+1)}(\phi,s) + (\varepsilon^{1/8}a(s)^{-3} + a(s)^{-1})\mathcal{E}^{(L)}(\phi,s) \\
&\qquad + \varepsilon a(s)^{-3}\mathcal{E}^{(L)}(\Sigma,s) + \varepsilon^{7/8}a(s)^{-3}\cdot a(s)^4\mathcal{E}^{(L-1)}(\mathrm{Ric},s) \\
&\qquad\quad + \varepsilon^{31/8}a(s)^{-3}\mathcal{E}^{(\le L-2)}(\mathrm{Ric},s) + (\varepsilon^{15/8}a(s)^{-3-c\sqrt{\varepsilon}} + a(s)^{-1-c\sqrt{\varepsilon}})\mathcal{E}^{(\le L-2)}(\phi,s) \\
&\qquad\quad + \varepsilon^{15/8}a(s)^{-3-c\sqrt{\varepsilon}}(\mathcal{E}^{(\le L-2)}(\Sigma,s) + \mathcal{E}^{(\le L-2)}(W,s)) \\
&\qquad\quad + \underbrace{\varepsilon^{15/8}a(s)^{-3-c\sqrt{\varepsilon}}\mathcal{E}^{(\le L-4)}(\mathrm{Ric},s)\,ds}_{\text{not present for } L=2}\big\}. \qquad (6\text{-}11)
\end{aligned}$$

Remark 6.7. We preemptively note that the error terms on the left-hand side, once combined with the similar terms on the left-hand side in Lemma 6.8 and given suitable weights, will turn out to have positive sign, even if they do not have definite sign in isolation.

The main idea in deriving this inequality is that we can use the algebraic identity (A-3d) and integration by parts to exploit the Maxwell system that lies at the core of the Bel–Robinson evolution equations. As a result, we avoid having higher-order energies of the Bel–Robinson variables on the right-hand side of the integral energy inequalities (which would break the bootstrap argument), then only having to deal with scalar field and Ricci energies at the next derivative level.

Proof. We first prove (6-11), and then explain how the same ideas lead to the simpler estimate (6-10). To this end, we start out by taking the time derivative of the energy as usual:

$$-\partial_t \mathcal{E}^{(L)}(W,\cdot) = \int_M -3N\frac{\dot a}{a}[|\Delta^{L/2}\boldsymbol{E}|_G^2 + |\Delta^{L/2}\boldsymbol{B}|_G^2] - 2(\langle \partial_t \Delta^{L/2}\boldsymbol{E}, \Delta^{L/2}\boldsymbol{E}\rangle_G + \langle \partial_t \Delta^{L/2}\boldsymbol{B}, \Delta^{L/2}\boldsymbol{B}\rangle_G) \\ - 2(\partial_t G^{-1})^{i_1 j_1}(G^{-1})^{i_2 j_2}[\Delta^{L/2}\boldsymbol{E}_{i_1 i_2}\Delta^{L/2}\boldsymbol{E}_{j_1 j_2} + \Delta^{L/2}\boldsymbol{B}_{i_1 i_2}\Delta^{L/2}\boldsymbol{B}_{j_1 j_2}]\,\mathrm{vol}_G.$$

$\boldsymbol{E}$ and $\boldsymbol{B}$ are symmetric and tracefree; thus symmetrizations become redundant, and any scalar product with a tensor that is pure trace or with an antisymmetric tensor can be dropped.[9] With this in hand, we get, inserting (2-38a) and (2-38b):

$$\begin{aligned}
&-\partial_t \mathcal{E}^{(L)}(W,\cdot)\\
&= \int_M \Big\{ \frac{\dot a}{a}(-6(N+1)+9N)(|\Delta^{L/2}\boldsymbol{E}|_G^2 + |\Delta^{L/2}\boldsymbol{B}|_G^2) && \text{(6-12a)}\\
&\quad + 2(N+1)a^{-1}\big(\langle \mathrm{curl}_G\,\Delta^{L/2}\boldsymbol{E}, \Delta^{L/2}\boldsymbol{B}\rangle_G - \langle \mathrm{curl}_G\,\Delta^{L/2}\boldsymbol{B}, \Delta^{L/2}\boldsymbol{E}\rangle_G\big) && \text{(6-12b)}\\
&\quad + 2a^{-1}\big(\langle \nabla\Delta^{L/2}N \wedge_G \boldsymbol{B}, \Delta^{L/2}\boldsymbol{E}\rangle_G - \langle \nabla\Delta^{L/2}N \wedge_G \boldsymbol{E}, \Delta^{L/2}\boldsymbol{B}\rangle_G\big) && \text{(6-12c)}\\
&\quad - 8\pi C^2 a^{-3}(N+1)\langle \Delta^{L/2}\Sigma, \Delta^{L/2}\boldsymbol{E}\rangle_G - 8\pi a(\Psi + C)\langle \nabla\Delta^{L/2}N\nabla\phi, \Delta^{L/2}\boldsymbol{E}\rangle_G && \text{(6-12d)}\\
&\quad - 8\pi a(\Psi + C)(N+1)\langle \nabla^2\Delta^{L/2}\phi, \Delta^{L/2}\boldsymbol{E}\rangle_G && \text{(6-12e)}\\
&\quad + 16\pi a(N+1)\langle \nabla\phi\nabla\Delta^{L/2}\Psi, \Delta^{L/2}\boldsymbol{E}\rangle_G && \text{(6-12f)}\\
&\quad + a^3\boldsymbol{\varepsilon}[G] * \nabla\phi * \nabla^2\Delta^{L/2}\phi * \Delta^{L/2}\boldsymbol{B} && \text{(6-12g)}\\
&\quad + (N+1)a^{-3}\Sigma * (\Delta^{L/2}\boldsymbol{E} * \Delta^{L/2}\boldsymbol{E} + \Delta^{L/2}\boldsymbol{B} * \Delta^{L/2}\boldsymbol{B}) && \text{(6-12h)}\\
&\quad - 2\langle \mathfrak{E}_{L,\mathrm{Border}} + \mathfrak{E}_{L,\mathrm{top}} + \mathfrak{E}^{\|}_{L,\mathrm{Junk}}, \Delta^{L/2}\boldsymbol{E}\rangle_G && \text{(6-12i)}\\
&\quad - 2\langle \mathfrak{B}_{L,\mathrm{Border}} + \mathfrak{B}_{L,\mathrm{top}} + \mathfrak{B}^{\|}_{L,\mathrm{Junk}}, \Delta^{L/2}\boldsymbol{B}\rangle_G \Big\}\,\mathrm{vol}_G. && \text{(6-12j)}
\end{aligned}$$

For (6-12a), we pull $6(N+1)\dot a a^{-1}|\Delta^{L/2}\boldsymbol{E}|_G^2$ to the left. This leaves

$$\int_M -6\frac{\dot a}{a}|\Delta^{L/2}\boldsymbol{B}|_G^2 + 3N\frac{\dot a}{a}|\Delta^{L/2}\boldsymbol{B}|_G^2 + 9N\frac{\dot a}{a}|\Delta^{L/2}\boldsymbol{E}|_G^2\,\mathrm{vol}_G,$$

where we can estimate the last two terms up to constant by $\varepsilon a^{1-c\sigma}\mathcal{E}^{(L)}(W,\cdot)$ by (3-17h) and can drop the first term since it is nonpositive.

Regarding (6-12b), note that we have

$$a^{-1}\big(\langle \mathrm{curl}_G\,\Delta^{L/2}\boldsymbol{E}, \Delta^{L/2}\boldsymbol{B}\rangle_G - \langle \mathrm{curl}_G\,\Delta^{L/2}\boldsymbol{B}, \Delta^{L/2}\boldsymbol{E}\rangle_G\big) = -a^{-1}\,\mathrm{div}_G(\Delta^{L/2}\boldsymbol{E} \wedge_G \Delta^{L/2}\boldsymbol{B}).$$

Hence, the absolute value of (6-12b), using (A-4c) for the wedge product and (3-17h), can be bounded by

$$\begin{aligned}
\Big|\int_M 2a^{-1}(N+1)\,\mathrm{div}_G(\Delta^{L/2}\boldsymbol{E} \wedge_G \Delta^{L/2}\boldsymbol{B})\,\mathrm{vol}_G\Big| &= \Big|\int_M 2a^{-1}\langle \nabla N, \Delta^{L/2}\boldsymbol{E} \wedge_G \Delta^{L/2}\boldsymbol{B}\rangle_G\,\mathrm{vol}_G\Big|\\
&\lesssim \int_M a^{-1}|\nabla N|_G|\Delta^{L/2}\boldsymbol{E}|_G|\Delta^{L/2}\boldsymbol{B}|_G\,\mathrm{vol}_G\\
&\lesssim \varepsilon a^{3-c\sigma}\mathcal{E}^{(L)}(W,\cdot).
\end{aligned}$$

[9] Recall the superscript "$\|$" notation for error terms; see Remark 2.12.

For (6-12c), we use the pointwise wedge product estimate (A-4d) and a priori estimates (4-2c) and (4-4g) to bound it as follows:

$$\begin{aligned}|\text{(6-12c)}| &\le 2a^{-1}|\nabla\Delta^{L/2}N|_G(|\boldsymbol{B}|_G\cdot|\Delta^{L/2}\boldsymbol{E}|_G+|\boldsymbol{E}|_G\cdot|\Delta^{L/2}\boldsymbol{B}|_G)\\ &\lesssim \varepsilon a^{-3}\sqrt{a^4\mathcal{E}^{(L+1)}(N,\cdot)}\sqrt{\mathcal{E}^{(L)}(W,\cdot)}\\ &\lesssim \varepsilon a^{-3}(\mathcal{E}^{(L)}(W,\cdot)+a^4\mathcal{E}^{(L+1)}(N,\cdot)).\end{aligned}$$

We pull the first term of (6-12d) to the left as well, and estimate the second using the strong C_G-norm estimates (4-2a) and (4-4e) by

$$\sqrt{\varepsilon}a^{-1-c\sqrt{\varepsilon}}\sqrt{a^4\mathcal{E}^{(L+1)}(N,\cdot)}\sqrt{\mathcal{E}^{(L)}(W,\cdot)}\lesssim a^{-1-c\sqrt{\varepsilon}}\mathcal{E}^{(L)}(W,\cdot)+\varepsilon a^{-1}\cdot a^4\mathcal{E}^{(L+1)}(N,\cdot).$$

Moving on to (6-12e)–(6-12g), we see (using (4-2a), (4-4e), (4-13a) with $\zeta=\Delta^{L/2}\phi$ and (3-8))

$$\begin{aligned}|\text{(6-12e)}| &\lesssim \big(a^{-3}\sqrt{a^4\mathcal{E}^{(L+1)}(\phi,\cdot)}+a^{-1}\sqrt{\mathcal{E}^{(L)}(\phi,\cdot)}+a^{-1-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(L-2)}(\phi,\cdot)}\big)\sqrt{\mathcal{E}^{(L)}(W,\cdot)}\\ &\lesssim (\varepsilon^{1/8}a^{-3}+a^{-1-c\sqrt{\varepsilon}})\mathcal{E}^{(L)}(W,\cdot)+\varepsilon^{-1/8}a^{-3}\cdot a^4\mathcal{E}^{(L+1)}(\phi,\cdot)+a^{-1}\mathcal{E}^{(\le L)}(\phi,\cdot),\\ |\text{(6-12f)}| &\lesssim \sqrt{\varepsilon}a^{1-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(L+1)}(\phi,\cdot)}\sqrt{\mathcal{E}^{(L)}(W,\cdot)}\\ &\lesssim a^{1-c\sqrt{\varepsilon}}\mathcal{E}^{(L)}(W,\cdot)+\varepsilon a^{1-c\sqrt{\varepsilon}}\mathcal{E}^{(L+1)}(\phi,\cdot),\\ |\text{(6-12g)}| &\lesssim \sqrt{\varepsilon}a^{1-c\sqrt{\varepsilon}}\cdot a^2\|\nabla^2\Delta^{L/2}\phi\|_{L^2_G}\cdot\sqrt{\mathcal{E}^{(L)}(W,\cdot)}\\ &\lesssim \sqrt{\varepsilon}a^{1-c\sqrt{\varepsilon}}\big(\sqrt{\mathcal{E}^{(L+1)}(\phi,\cdot)}+a^{-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(L-1)}(\phi,\cdot)}\big)\cdot\sqrt{\mathcal{E}^{(L)}(W,\cdot)}\\ &\lesssim a^{1-c\sqrt{\varepsilon}}\mathcal{E}^{(L)}(W,\cdot)+\varepsilon a^{1-c\sqrt{\varepsilon}}[\mathcal{E}^{(L+1)}(\phi,\cdot)+\mathcal{E}^{(L)}(\phi,\cdot)+\mathcal{E}^{(\le L-2)}(\phi,\cdot)].\end{aligned}$$

We can estimate (6-12h) by $\varepsilon a^{-3}\mathcal{E}^{(L)}(W,\cdot)$ as usual, and obtain the following in summary:

$$\begin{aligned}&-\partial_t\mathcal{E}^{(L)}(W,\cdot)+8\pi C^2a^{-3}\int_M(N+1)\langle\Delta^{L/2}\Sigma,\Delta^{L/2}\boldsymbol{E}\rangle_G\,\mathrm{vol}_G+6\frac{\dot a}{a}\int_M(N+1)|\Delta^{L/2}\boldsymbol{E}|^2_G\,\mathrm{vol}_G\\ &\lesssim(\varepsilon a^{-3}+a^{-1-c\sqrt{\varepsilon}})\mathcal{E}^{(L)}(W,\cdot)+a^{-1}\mathcal{E}^{(L+1)}(\phi,\cdot)+a^{-1}\mathcal{E}^{(L)}(\phi,\cdot)+\varepsilon a^{-3}\cdot a^4\mathcal{E}^{(L+1)}(N,\cdot)\\ &\quad+a^{-1}\mathcal{E}^{(\le L-2)}(\phi,\cdot)\big[\|\mathfrak{E}_{L,\mathrm{Border}}\|_{L^2_G}+\|\mathfrak{E}_{L,\mathrm{top}}\|_{L^2_G}+\|\mathfrak{E}^{\|}_{L,\mathrm{Junk}}\|_{L^2_G}\\ &\qquad\qquad+\|\mathfrak{B}_{L,\mathrm{Border}}\|_{L^2_G}+\|\mathfrak{B}_{L,\mathrm{top}}\|_{L^2_G}+\|\mathfrak{B}^{\|}_{L,\mathrm{Junk}}\|_{L^2_G}\big]\sqrt{\mathcal{E}^{(L)}(W,\cdot)}.\end{aligned}$$

We can now apply Corollary 5.5 for $2l=L$ to estimate the lapse energy in the second line (leading to borderline scalar field energy and Σ-energy contributions, as well as junk terms), and insert the borderline (see (A-17k)), top (see (A-18a) and (A-18b)) and junk estimates (see (A-19n)), dealing with the lapse energies there analogously. In particular, the top-order curvature terms arise as follows:

$$\begin{aligned}\|\mathfrak{E}_{L,\mathrm{top}}\|_{L^2_G}\sqrt{\mathcal{E}^{(L)}(W,\cdot)} &\lesssim \sqrt{\varepsilon}a^{-1-c\sqrt{\varepsilon}}\sqrt{a^4\mathcal{E}^{(L-1)}(\mathrm{Ric},\cdot)}\sqrt{\mathcal{E}^{(L)}(W,\cdot)}\\ &\lesssim \varepsilon^{1/8}a^{-1-c\sqrt{\varepsilon}}\mathcal{E}^{(L)}(W,\cdot)+\varepsilon^{7/8}a^{-1}\cdot a^4\mathcal{E}^{(L-1)}(\mathrm{Ric},\cdot),\\ \|\mathfrak{B}_{L,\mathrm{top}}\|\sqrt{\mathcal{E}^{(L)}(W,\cdot)} &\lesssim \varepsilon a^{-3}\sqrt{a^4\mathcal{E}^{(L-1)}(\mathrm{Ric},\cdot)}\sqrt{\mathcal{E}^{(L)}(W,\cdot)}\\ &\lesssim \varepsilon^{1/8}a^{-3}\mathcal{E}^{(L)}(W,\cdot)+\varepsilon^{15/8}a^{-3}\cdot a^4\mathcal{E}^{(L-1)}(\mathrm{Ric},\cdot).\end{aligned}$$

Hence, both top-order curvature terms can be bounded by $\varepsilon^{7/8}a^{-3}\cdot a^4\mathcal{E}^{(L-1)}(\mathrm{Ric},\cdot)$.

Integrating the inequality yields (6-11).

For (6-10), we get applying (2-31a) and (2-31b) and again using that $\boldsymbol{E}$ and $\boldsymbol{B}$ are symmetric and tracefree,

$$\begin{aligned}-\partial_t\mathcal{E}^{(0)}(W,\cdot) = \int_M \Big\{ &\frac{\dot a}{a}(-6(N+1)+9N)(|\boldsymbol{E}|_G^2+|\boldsymbol{B}|_G^2)+2(N+1)(\langle \operatorname{curl}\boldsymbol{E},\boldsymbol{B}\rangle_G-\langle\operatorname{curl}\boldsymbol{B},\boldsymbol{E}\rangle_G)\\ &+2\big(\langle\nabla N\wedge\boldsymbol{B},\boldsymbol{E}\rangle_G-\langle\nabla N\wedge\boldsymbol{E},\boldsymbol{B}\rangle_G\big)+(N+1)a^{-3}\Sigma*(\boldsymbol{E}*\boldsymbol{E}+\boldsymbol{B}*\boldsymbol{B})\\ &-8\pi a^{-3}(N+1)(\Psi+C)^2\langle\Sigma,\boldsymbol{E}\rangle_G+[\dot a a^3\nabla\phi*\nabla\phi+a(\Psi+C)\cdot\nabla N*\nabla\phi]*\boldsymbol{E}\\ &+a(N+1)[\nabla\phi*\nabla\Psi+\Sigma*\nabla\phi*\nabla\phi+(\Psi+C)\nabla^2\phi]*\boldsymbol{E}\\ &+(N+1)\boldsymbol{\varepsilon}[G]*\big(a^3\nabla^2\phi*\nabla\phi+a^{-1}(\Psi+C)\Sigma*\nabla\phi\big)*\boldsymbol{B}\Big\}\operatorname{vol}_G.\end{aligned}$$

The first two lines are treated as in the general case. For the third line, we get $\varepsilon a^{3-c\sigma}\mathcal{E}^{(0)}(W,\cdot)$ with (A-4d) and (3-17h), while the fourth term is bounded by $\varepsilon a^{-3}\mathcal{E}^{(0)}(W,\cdot)$ with (4-2b). This leaves the surviving matter terms in the final four lines.

We pull $\int_M 8\pi a^{-3}(N+1)C^2\langle\Sigma,\boldsymbol{E}\rangle_G\operatorname{vol}_G$ to the left as before. For the remaining terms, we can apply a priori estimates (4-2a), (4-4a) and (4-4e), the bootstrap assumption (3-17h) and Lemma 4.1 for N, which yields the following bound up to constant remaining terms in the last three lines:

$$\sqrt{\mathcal{E}^{(0)}(W,\cdot)}\cdot\big[a^{-1}\cdot a^2\|\nabla^2\phi\|_{L^2_G}+\sqrt{\varepsilon}a^{-1-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(0)}(\phi,\cdot)}+(\varepsilon a^{-3}+\sqrt{\varepsilon}a^{-1-c\sqrt{\varepsilon}})\sqrt{\mathcal{E}^{(0)}(\Sigma,\cdot)}\big].$$

Applying (4-13a) to the scalar field norm and then (3-8), this leads to (6-10) along with the previous observations. □

6.3. ***Energy estimates for the second fundamental form.*** For the energy estimates for Σ, we again first derive even-order integral estimates:

Lemma 6.8 (energy estimates for the second fundamental form for even orders). *Let $t\in(t_{\mathrm{Boot}},t_0]$. Then, one has*

$$\begin{aligned}\mathcal{E}^{(0)}(\Sigma,t)+2\int_t^{t_0}\int_M\Big[a(s)^{-3}(N+1)\langle\boldsymbol{E},\Sigma\rangle_G+\frac{\dot a(s)}{a(s)}(N+1)|\Delta^{L/2}\Sigma|_G^2\Big]\operatorname{vol}_G\,ds\\ \lesssim\mathcal{E}^{(0)}(\Sigma,t_0)+\int_t^{t_0}\varepsilon^{1/8}a(s)^{-3}\mathcal{E}^{(0)}(\Sigma,s)\,ds+\int_t^{t_0}\varepsilon^{-1/8}a(s)^{-3}\mathcal{E}^{(0)}(\phi,s)\,ds.\end{aligned}\tag{6-13}$$

For $L\in2\mathbb{N}$, $L\le18$,

$$\begin{aligned}&\mathcal{E}^{(L)}(\Sigma,t)+2\int_t^{t_0}\int_M\Big[a(s)^{-3}(N+1)\langle\Delta^{L/2}\boldsymbol{E},\Delta^{L/2}\Sigma\rangle_G+\frac{\dot a(s)}{a(s)}(N+1)|\Delta^{L/2}\Sigma|_G^2\Big]\operatorname{vol}_G\,ds\\ &\lesssim\mathcal{E}^{(L)}(\Sigma,t_0)+\int_t^{t_0}\varepsilon^{1/8}a(s)^{-3}\mathcal{E}^{(L)}(\Sigma,s)\,ds\\ &\quad+\int_t^{t_0}\big\{\varepsilon^{-1/8}a(s)^{-3}\mathcal{E}^{(L)}(\phi,s)+\varepsilon^{15/8}a(s)^{5-c\sigma}\mathcal{E}^{(L-1)}(\mathrm{Ric},s)+\varepsilon^2a(s)^{-3}\mathcal{E}^{(L-2)}(\mathrm{Ric},s)\\ &\qquad+\varepsilon^{15/8}a(s)^{-3-c\sqrt{\varepsilon}}\mathcal{E}^{(\le L-2)}(\Sigma,s)+(\varepsilon^{15/8}a(s)^{-3-c\sqrt{\varepsilon}}+\varepsilon a(s)^{-1-c\sqrt{\varepsilon}})\mathcal{E}^{(\le L-2)}(\phi,s)\\ &\qquad+\underbrace{\varepsilon^2a(s)^{-3-c\sqrt{\varepsilon}}\mathcal{E}^{(\le L-4)}(\mathrm{Ric},s)}_{\text{not present for } L=2}\big\}\,ds.\end{aligned}\tag{6-14}$$

Remark 6.9. The main hurdle of dealing with the second fundamental form is that a high-order curvature term occurs in the evolution equation. It is to precisely this end that the Bel–Robinson variables needed to be introduced, since (2-36d) is what facilitates controlling said term without having to use $\mathcal{E}^{(L)}(\mathrm{Ric},\cdot)$ or similar high-order metric energies. Again, the resulting leading terms will turn out to have definite sign when combined with the Bel–Robinson energy estimates above.

Proof. Here we omit the proof for the inequality at order zero since is completely analogous in structure to the one for orders 2 and higher; the only differences that arise are that lower-order error terms do not occur.

Once again, we start out by differentiating $-\mathcal{E}^{(L)}(\Sigma,\cdot)$ and insert (2-35):

$$\begin{aligned}
-\partial_t\mathcal{E}^{(L)}(\Sigma,\cdot) &= \int_M -2\langle\partial_t\Delta^{L/2}\Sigma,\Delta^{L/2}\Sigma\rangle_G + (\partial_t G^{-1})*G^{-1}*\Delta^{L/2}\Sigma*\Delta^{L/2}\Sigma - 3N\frac{\dot a}{a}|\Delta^{L/2}\Sigma|_G^2\,\mathrm{vol}_G\\
&= \int_M \big\{2a\langle\nabla^2\Delta^{L/2}N,\Delta^{L/2}\Sigma\rangle_G - 2a(N+1)\langle\Delta^{L/2}\,\mathrm{Ric}[G],\Delta^{L/2}\Sigma\rangle_G\\
&\qquad + (\partial_t G^{-1})*G^{-1}*\Delta^{L/2}\Sigma*\Delta^{L/2}\Sigma - 3N\frac{\dot a}{a}|\Delta^{L/2}\Sigma|_G^2\\
&\qquad - 2\langle\mathfrak{S}_{L,\mathrm{Border}},\Delta^{L/2}\Sigma\rangle_G - 2\langle\mathfrak{S}^{\parallel}_{L,\mathrm{Junk}},\Delta^{L/2}\Sigma\rangle_G\big\}\,\mathrm{vol}_G.
\end{aligned}$$

For the first term, one can use (4-13a), Corollary 5.5 at order L and (3-8) to bound its absolute value by the following:

$$\begin{aligned}
&\lesssim a\|\Delta^{L/2}N\|_{\dot H^2_G}\sqrt{\mathcal{E}^{(L)}(\Sigma,\cdot)}\\
&\lesssim [a^{-3}\sqrt{a^8\mathcal{E}^{(L+2)}(N,\cdot)} + a^{1-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(L)}(N,\cdot)}]\sqrt{\mathcal{E}^{(L)}(\Sigma,\cdot)}\\
&\lesssim \big[\varepsilon a^{-3}\sqrt{\mathcal{E}^{(L)}(\Sigma,\cdot)} + a^{-3}\mathcal{E}^{(L)}(\phi,\cdot) + \varepsilon a^{-3-c\sqrt{\varepsilon}}\big(\sqrt{\mathcal{E}^{(\le L-2)}(\Sigma,\cdot)} + \sqrt{\mathcal{E}^{(\le L-2)}(\phi,\cdot)}\big)\\
&\qquad + \underbrace{(\varepsilon^2 a^{-3-c\sqrt{\varepsilon}} + \varepsilon a^{1-c\sigma})\sqrt{\mathcal{E}^{(\le L-3)}(\mathrm{Ric},\cdot)}}_{\text{not present for } L=2}\big]\sqrt{\mathcal{E}^{(L)}(\Sigma,\cdot)}\\
&\lesssim (\varepsilon^{1/8}a^{-3} + a^{1-c\sigma})\mathcal{E}^{(L)}(\Sigma,\cdot) + \varepsilon^{-1/8}a^{-3}\mathcal{E}^{(L)}(\phi,\cdot) + \varepsilon a^{-3-c\sqrt{\varepsilon}}[\mathcal{E}^{(\le L-2)}(\Sigma,\cdot) + \mathcal{E}^{(\le L-2)}(\phi,\cdot)]\\
&\qquad + \underbrace{(\varepsilon^{31/8}a^{-3} + \varepsilon^2 a^{1-c\sigma})\mathcal{E}^{(L-2)}(\mathrm{Ric},\cdot) + (\varepsilon^{31/8}a^{-3-c\sqrt{\varepsilon}} + \varepsilon^2 a^{1-c\sigma})\mathcal{E}^{(\le L-4)}(\mathrm{Ric},\cdot)}_{\text{not present for } L=2}.
\end{aligned}$$

Next, we replace the high-order curvature term as follows, using the commuted rescaled Hamiltonian constraint equation (2-36d) that $\Delta^{L/2}\Sigma$ is tracefree and symmetric:

$$\begin{aligned}
&\int_M -2a(N+1)\langle\Delta^{L/2}\,\mathrm{Ric}[G],\Delta^{L/2}\Sigma\rangle_G\,\mathrm{vol}_G\\
&= \int_M -2(N+1)a^{-3}\langle\Delta^{L/2}\boldsymbol{E},\Delta^{L/2}\Sigma\rangle_G - 2(N+1)\frac{\dot a}{a}|\Delta^{L/2}\Sigma|_G^2 + \langle\mathfrak{H}_{L,\mathrm{Border}} + \mathfrak{H}^{\parallel}_{L,\mathrm{Junk}},\Delta^{L/2}\Sigma\rangle_G\,\mathrm{vol}_G.
\end{aligned}$$

We pull the first two terms to left, only keeping the error terms on the right. After inserting the borderline and junk term estimates for the Hamiltonian constraint equations ((A-17a) and (A-19c)) and the evolution equation itself ((A-17h) and (A-19k)), as well as bounding $|\partial_t G^{-1}| \lesssim \varepsilon a^{-3}$ and inserting (3-17h) as usual, we obtain (6-14) by integrating. □

Additionally, we can exploit the structure of the momentum constraint equations to gain an elliptic estimate for $\mathcal{E}^{(L+1)}(\Sigma,\cdot)$. Crucially, the upper bound only depends on Σ-, scalar field and Bel–Robinson

energies up to order L, and appropriately small and time-scaled curvature contributions up to order $L-1$. This will allow us to close the argument since we do not need to consider the Bel–Robinson energy at order $L+1$ to control Σ at that order, which would require higher-order scalar field and curvature energies.

Lemma 6.10 (odd-order energy estimate for the second fundamental form). *For any $L \in 2\mathbb{Z}_+$, $2 \le L \le 18$, we have*

$$a^4\mathcal{E}^{(L+1)}(\Sigma,\cdot) \lesssim (a^{4-c\sqrt{\varepsilon}}+\varepsilon a^{2-c\sqrt{\varepsilon}})\mathcal{E}^{(L)}(\Sigma,\cdot)+\mathcal{E}^{(L)}(\phi,\cdot)+\mathcal{E}^{(L)}(W,\cdot)+\varepsilon^2a^4\mathcal{E}^{(L-1)}(\mathrm{Ric},\cdot)$$
$$+\varepsilon a^{-c\sqrt{\varepsilon}}\mathcal{E}^{(\le L-2)}(\phi,\cdot)+a^{2-c\sqrt{\varepsilon}}\mathcal{E}^{(\le L-2)}(\Sigma,\cdot)+\varepsilon a^{2-c\sqrt{\varepsilon}}\mathcal{E}^{(\le L-2)}(\mathrm{Ric},\cdot). \quad (6\text{-}15)$$

For $L=0$, one analogously has

$$a^4\mathcal{E}^{(1)}(\Sigma,\cdot) \lesssim (a^{4-c\sqrt{\varepsilon}}+\varepsilon a^{2-c\sqrt{\varepsilon}})\mathcal{E}^{(0)}(\Sigma,\cdot)+\mathcal{E}^{(0)}(\phi,\cdot)+\mathcal{E}^{(0)}(W,\cdot). \quad (6\text{-}16)$$

Proof. We prove the statement for $L \ge 2$, since the proof of (6-16) is entirely analogous.

By [Andersson and Moncrief 2004, (A.22)], since (Σ_t, g) is a three-dimensional compact Riemannian manifold for any $t \in (t_{\mathrm{Boot}}, t_0]$, any tracefree $(0,2)$ tensor U_{ij} on (Σ_t, g) satisfies

$$\int_{\Sigma_t} |\nabla U|_g^2 + 3\,\mathrm{Ric}[g]\cdot U\cdot U - \tfrac{1}{2}R[g]\,|U|_g^2\,\mathrm{vol}_g = \int_{\Sigma_t} |\mathrm{curl}\,U|_g^2 + \tfrac{3}{2}|\mathrm{div}_g\,U|_g^2\,\mathrm{vol}_g. \quad (6\text{-}17)$$

In particular, for $U = \Delta^{L/2}\Sigma$ and after rescaling, this reads

$$\int_M |\nabla\Delta^{L/2}\Sigma|_G^2 + 3(\mathrm{Ric}[G]^\sharp)^i{}_j(\Delta^{L/2}\Sigma^\sharp)^j{}_l(\Delta^{L/2}\Sigma^\sharp)^l{}_i - \tfrac{1}{2}R[G]|\Delta^{L/2}\Sigma|_G^2\,\mathrm{vol}_G$$
$$= \int_M \tfrac{3}{2}|\mathrm{div}_G\,\Delta^{L/2}\Sigma|_G^2 + a^2|\mathrm{curl}\,\Delta^{L/2}\Sigma|_G^2\,\mathrm{vol}_G.$$

The last two terms on the left-hand side can be estimated by $(1+\sqrt{\varepsilon}a^{-c\sqrt{\varepsilon}})\mathcal{E}^{(L)}(\Sigma,\cdot)$ in absolute value using the strong C_G-norm estimate (4-4f). Thus, inserting the Laplace-commuted rescaled momentum constraint equations (2-36a) and (2-36b), we obtain for a suitable constant $K>0$

$$\mathcal{E}^{(L+1)}(\Sigma,\cdot) - K(1+\sqrt{\varepsilon}a^{-c\sqrt{\varepsilon}})\mathcal{E}^{(L)}(\Sigma,\cdot)$$
$$\lesssim \int_M \big\{|\Psi+C|^2|\nabla\Delta^{L/2}\phi|_G^2 + |\nabla\phi|_G^2|\Delta^{L/2-1}\,\mathrm{Ric}[G]|_G^2 + |\Sigma|_G^2|\nabla\Delta^{L/2-1}\,\mathrm{Ric}[G]|_G^2 + |\mathfrak{M}_{L,\mathrm{Junk}}|_G^2$$
$$+a^{-4}|\Delta^{L/2}\boldsymbol{B}|_G^2 + |\Sigma|_G^2|\nabla\Delta^{L/2-1}\,\mathrm{Ric}[G]|_G^2 + |\nabla\Sigma|_G^2|\nabla^2\Delta^{L/2-2}\,\mathrm{Ric}[G]|_G^2 + |\widetilde{\mathfrak{M}}_{L,\mathrm{Junk}}|_G^2\big\},\,\mathrm{vol}_G.$$

After rearranging, using the strong C_G-norm estimates (4-2a), (4-4e), (4-2b) and (4-4b) and multiplying by a^4 on both sides, we get

$$a^4\mathcal{E}^{(L+1)}(\Sigma,\cdot) \lesssim (1+\sqrt{\varepsilon}a^{-c\sqrt{\varepsilon}})a^4\mathcal{E}^{(L)}(\Sigma,\cdot)+\mathcal{E}^{(L)}(\phi,\cdot)+\mathcal{E}^{(L)}(W,\cdot)+\varepsilon^2a^4\mathcal{E}^{(L-1)}(\mathrm{Ric},\cdot)$$
$$+\varepsilon^2a^{4-c\sqrt{\varepsilon}}\mathcal{E}^{(L-2)}(\mathrm{Ric},\cdot)+a^4\|\mathfrak{M}_{L,\mathrm{Junk}}\|_{L^2_G}^2+a^4\|\widetilde{\mathfrak{M}}_{L,\mathrm{Junk}}\|_{L^2_G}^2.$$

The statement follows inserting the estimates (A-19a) and (A-19b). □

6.4. ***Energy estimates for the curvature.*** To control commutator errors, we will also need some additional estimates on curvature energies. Unlike the other energies, these inequalities do not rely on any delicate structure within the equations and instead just rely on pointwise estimates, the Young inequality and

near-coercivity of energies in the sense of Lemma 4.5. For the sake of convenience, we phrase these estimates for $\mathcal{E}^{(L-2)}(\mathrm{Ric},\cdot)$ since this is the order needed when improving behaviour of the total energy at order L.

Lemma 6.11 (curvature energy estimates at even orders). *Let $L \in 2\mathbb{Z}$, $4 \le L \le 16$, and $t \in (t_{\mathrm{Boot}}, t_0]$. Then, one has*

$$\begin{aligned}\mathcal{E}^{(L-2)}(\mathrm{Ric},t) \lesssim \mathcal{E}^{(L-2)}(\mathrm{Ric},t_0) &+ \int_t^{t_0} (\varepsilon^{1/8} a(s)^{-3} + a(s)^{8-c\sigma})\,\mathcal{E}^{(L-2)}(\mathrm{Ric},s)\,ds \\ &+ \int_t^{t_0} \Big\{ \varepsilon^{-1/8} a(s)^{-3} (\mathcal{E}^{(L)}(\phi,s) + \mathcal{E}^{(L)}(\Sigma,s)) \\ &\qquad + \varepsilon^{-1/8} a(s)^{-3-c\sqrt{\varepsilon}} (\mathcal{E}^{(\le L-2)}(\phi,s) + \mathcal{E}^{(\le L-2)}(\Sigma,s)) \\ &\qquad + \varepsilon^{7/8} a(s)^{-3-c\sqrt{\varepsilon}} \mathcal{E}^{(\le L-4)}(\mathrm{Ric},s) \Big\}\,ds. \quad (6\text{-}18)\end{aligned}$$

Additionally,

$$\begin{aligned}\mathcal{E}^{(0)}(\mathrm{Ric},t) \lesssim \mathcal{E}^{(0)}(\mathrm{Ric},t_0) &+ \int_t^{t_0} \varepsilon^{1/8} a(s)^{-3} \mathcal{E}^{(0)}(\mathrm{Ric},s)\,ds \\ &+ \int_t^{t_0} \varepsilon^{-1/8} a(s)^{-3} (\mathcal{E}^{(0)}(\phi,s) + \mathcal{E}^{(0)}(\Sigma,s))\,ds. \quad (6\text{-}19)\end{aligned}$$

Proof. First, we note that

$$\|\mathrm{div}_G^\sharp \nabla \Delta^{L/2-1}\Sigma\|_{L^2_G} \lesssim \|\nabla^2 \Delta^{L/2-1}\Sigma\|_{L^2_G} \lesssim \|\Delta^{L/2}\Sigma\|_{L^2_G} + a^{-c\sqrt{\varepsilon}} \sqrt{\mathcal{E}^{(L-2)}(\Sigma,\cdot)}$$

holds using the low-order version of (4-14a) with $\mathfrak{T} = \Delta^{L/2-1}\Sigma$ for $l = 2$, and similarly

$$\|\nabla^2 \Delta^{L/2-1} N\|_{L^2_G} \lesssim \|\Delta^{L/2} N\|_{L^2_G} + a^{-c\sqrt{\varepsilon}} \sqrt{\mathcal{E}^{(L-2)}(N,\cdot)}$$

using (4-13b) at order 2. Now, using $\Delta^{L/2-1} G = 0$ for $L \ge 4$, we continue as usual by applying (2-40a) to the expression below:

$$\begin{aligned}-\partial_t \mathcal{E}^{(L-2)}(\mathrm{Ric},\cdot) \lesssim \int_M \Big\{ & a^{-3}(|\Delta^{L/2}\Sigma|_G + |\nabla^2\Delta^{L/2-1}\Sigma|_G)\,|\Delta^{L/2-1}\,\mathrm{Ric}[G]|_G \\ &+ \frac{\dot a}{a}(|\nabla^2\Delta^{L/2-1}N|_G + |\Delta^{L/2}N|_G)\,|\Delta^{L/2-1}\,\mathrm{Ric}[G]|_G \\ &+ (|\mathfrak{R}_{L-2,\mathrm{Border}}|_G + |\mathfrak{R}_{L-2,\mathrm{Junk}}|_G)\cdot|\Delta^{L/2-1}\,\mathrm{Ric}[G]|_G \\ &+ a^{-3}\Sigma * \Delta^{L/2-1}\,\mathrm{Ric}[G] * \Delta^{L/2-1}\,\mathrm{Ric}[G] + N\frac{\dot a}{a}|\Delta^{L/2-1}\,\mathrm{Ric}[G]|_G^2 \Big\}\,\mathrm{vol}_G.\end{aligned}$$

Due to the estimates above as well as (4-2b) and (3-17h), this implies

$$\begin{aligned}-\partial_t \mathcal{E}^{(L-2)}(\mathrm{Ric},\cdot) \lesssim\ & a^{-3}[\sqrt{\mathcal{E}^{(L)}(\Sigma,\cdot)} + \sqrt{\mathcal{E}^{(L)}(N,\cdot)}]\sqrt{\mathcal{E}^{(L-2)}(\mathrm{Ric},\cdot)} \\ &+ a^{-3-c\sqrt{\varepsilon}}[\sqrt{\mathcal{E}^{(\le L-2)}(\Sigma,\cdot)} + \sqrt{\mathcal{E}^{(\le L-2)}(N,\cdot)}]\sqrt{\mathcal{E}^{(L-2)}(\mathrm{Ric},\cdot)} \\ &+ (\|\mathfrak{R}_{L-2,\mathrm{Border}}\|_{L^2_G} + \|\mathfrak{R}_{L-2,\mathrm{Junk}}\|_{L^2_G})\sqrt{\mathcal{E}^{(L-2)}(\mathrm{Ric},\cdot)} + \varepsilon a^{-3}\mathcal{E}^{(L-2)}(\mathrm{Ric},\cdot).\end{aligned}$$

Using Corollary 5.5 at order L and distributing terms containing $\mathcal{E}^{(L-3)}(\mathrm{Ric},\cdot)$ with (3-8) as usual, we get

$$\begin{aligned}-\partial_t\mathcal{E}^{(L-2)}(\mathrm{Ric},\cdot)\lesssim{}&[\varepsilon^{1/8}a^{-3}+a^{8-c\sigma}]\mathcal{E}^{(L-2)}(\mathrm{Ric},\cdot)+\varepsilon^{-1/8}a^{-3}[\mathcal{E}^{(L)}(\Sigma,\cdot)+\mathcal{E}^{(L)}(\phi,\cdot)]\\&+[\varepsilon^{31/8}a^{-3-c\sqrt{\varepsilon}}+\varepsilon^2a^{8-c\sigma}]\mathcal{E}^{(\leq L-4)}(\mathrm{Ric},\cdot)\\&+\varepsilon^{-1/8}a^{-3-c\sqrt{\varepsilon}}[\mathcal{E}^{(\leq L-2)}(\Sigma,\cdot)+\mathcal{E}^{(\leq L-2)}(\phi,\cdot)]\\&+(\|\mathfrak{R}_{L-2,\mathrm{Border}}\|_{L^2_G}+\|\mathfrak{R}_{L-2,\mathrm{Junk}}\|_{L^2_G})\sqrt{\mathcal{E}^{(L-2)}(\mathrm{Ric},\cdot)}.\end{aligned}$$

Equation (6-18) now follows inserting the borderline and junk term estimates (A-17i) and (A-19l) and applying the lapse energy estimates from Corollary 5.5.

Equation (6-19) follows almost identically by inserting (2-33) instead of (2-40a), as well as (2-28a) for the additional $\partial_t G*\big(\mathrm{Ric}[G]+\frac{2}{9}G\big)$-terms. These can be estimated as

$$\begin{aligned}&\lesssim\int_M a^{-3}\Sigma*\big(\mathrm{Ric}[G]+\tfrac{2}{9}G\big)+\frac{\dot a}{a}N\cdot G*\big(\mathrm{Ric}[G]+\tfrac{2}{9}G\big)\,\mathrm{vol}_G\\&\lesssim a^{-3}\big(\sqrt{\mathcal{E}^{(0)}(\Sigma,\cdot)}+\sqrt{\mathcal{E}^{(0)}(N,\cdot)}\big)\sqrt{\mathcal{E}^{(0)}(\mathrm{Ric},\cdot)},\end{aligned}$$

which can be treated as at higher orders. □

Lemma 6.12 (odd-order curvature energy estimate). *For $L\in2\mathbb{N}$, $4\leq L\leq 18$ and $t\in(t_{\mathrm{Boot}},t_0]$,*

$$\begin{aligned}&a(t)^4\mathcal{E}^{(L-1)}(\mathrm{Ric},t)\\&\lesssim a(t_0)^4\mathcal{E}^{(L-1)}(\mathrm{Ric},t_0)+\int_t^{t_0}(\varepsilon^{1/8}a(s)^{-3}+a(s)^{-1-c\sqrt{\varepsilon}})(a(s)^4\mathcal{E}^{(L-1)}(\mathrm{Ric},s))\,ds\\&\quad+\int_t^{t_0}\varepsilon^{-1/8}a(s)^{-3}\cdot a(s)^4\mathcal{E}^{(L+1)}(\Sigma,s)\,ds\\&\quad+\int_t^{t_0}\big\{\varepsilon^{-1/8}a(s)^{-3}\mathcal{E}^{(L)}(\phi,s)+(\varepsilon^{15/8}a(s)^{-3}+a(s)^{-1-c\sqrt{\varepsilon}})\mathcal{E}^{(L)}(\Sigma,s)\\&\qquad+(\varepsilon^{15/8}a(s)^{-3-c\sqrt{\varepsilon}}+a(s)^{-1-c\sqrt{\varepsilon}})(\mathcal{E}^{(\leq L-2)}(\phi,s)+\mathcal{E}^{(\leq L-2)}(\Sigma,s))\\&\qquad+\varepsilon^{15/8}a(s)^{-3}\mathcal{E}^{(L-2)}(\mathrm{Ric},\cdot)+\varepsilon^{15/8}a^{-3-c\sqrt{\varepsilon}}\mathcal{E}^{(\leq L-4)}(\mathrm{Ric},s)\big\}\,ds.\end{aligned}\tag{6-20}$$

Proof. The proof is very similar to that of Lemma 6.11 since we did not exploit any structure within (2-40a) that does not equally occur in (2-40b), and thus we omit the details. As in the proof of Lemma 6.4, we note that the differences within the estimate come from how top-order lapse terms are treated: the scaling of the top-order energy allows one to estimate $a^4\mathcal{E}^{(L+1)}(N,\cdot)$ by scalar field energies and Σ-energies of up to order L and curvature energies up to order $L-3$. □

6.5. ***Sobolev norm estimates for metric objects.*** To close the bootstrap argument, we need to improve the behaviour of metric quantities in addition to the energy formalism, both to capture the intrinsic behaviour of the metric and to relate energies to supremum norms.

Lemma 6.13 (Sobolev norm estimates for Christoffel symbols). *Let U be a coordinate neighbourhood on M, viewed as a coordinate neighbourhood on Σ_t for $t\in(t_{\mathrm{Boot}},t_0]$. For any $l\in\mathbb{N}$, $l\leq 17$, the following Sobolev estimate then holds*:

$$\|\Gamma-\widehat{\Gamma}\|^2_{H^l_G(U)}\lesssim a^{-c\varepsilon^{1/8}}\big(\varepsilon^4+\varepsilon^{-1/4}\sup_{s\in(t,t_0)}(\|N\|^2_{H^{l+1}_G(\Sigma_s)}+\|\Sigma\|^2_{H^{l+1}_G(\Sigma_s)})\big).\tag{6-21}$$

Proof. Commuting the evolution equation (2-34) with ∇^J, we get for $J \in \mathbb{N},\ J \leq 17$,

$$\begin{aligned}-\partial_t \|\Gamma - \widehat{\Gamma}\|^2_{\dot{H}^J_G} = \int_U \Big[&(\partial_t G^{-1}) * G^{-1} * \cdots * G^{-1} * G * \nabla^J(\Gamma - \widehat{\Gamma}) * \nabla^J(\Gamma - \widehat{\Gamma}) \\ &+ (G^{-1}) * \cdots * (G^{-1}) * \partial_t G * \nabla^J(\Gamma - \widehat{\Gamma}) * \nabla^J(\Gamma - \widehat{\Gamma}) \\ &+ \Big(a^{-3} \sum_{I_N + I_\Sigma = J+1} \nabla^{I_N}(N+1) * \nabla^{I_\Sigma}\Sigma + \frac{\dot{a}}{a}\nabla^{J+1}N \Big) * \nabla^J(\Gamma - \widehat{\Gamma}) \\ &+ 2\langle [\partial_t, \nabla^J](\Gamma - \widehat{\Gamma}), \nabla^J(\Gamma - \widehat{\Gamma})\rangle_G - 3N\frac{\dot{a}}{a}|\nabla^J(\Gamma - \widehat{\Gamma})|^2_G \Big] \mathrm{vol}_G.\end{aligned}$$

We recall that (4-4c) implies

$$\|\Gamma - \widehat{\Gamma}\|_{C^{11}_G} \lesssim \sqrt{\varepsilon} a^{-c\sqrt{\varepsilon}} \tag{6-22}$$

by (3-5). It follows from inserting this in (A-10b) along with (4-2b), (4-4b) and (3-17h) that

$$\|[\partial_t, \nabla^J](\Gamma - \widehat{\Gamma})\|_{L^2_G} \lesssim \sqrt{\varepsilon} a^{-3-c\sqrt{\varepsilon}} \|\Sigma\|_{H^J_G} + \sqrt{\varepsilon} a^{-3-c\sqrt{\varepsilon}} \|N\|_{H^J_G} + \varepsilon a^{-3} \|\Gamma - \widehat{\Gamma}\|_{H^{J-1}_G}$$

is satisfied. Consequently and using the same strong C_G-norm bounds along with Lemma 4.1, the differential inequality becomes

$$\begin{aligned}-\partial_t \|\Gamma - \widehat{\Gamma}\|^2_{\dot{H}^J_G} \lesssim\ & \varepsilon^{1/8} a^{-3} \|\Gamma - \widehat{\Gamma}\|^2_{\dot{H}^J_G} + \varepsilon^{-1/8} a^{-3} (\|N\|^2_{H^{J+1}_G} + \|\Sigma\|^2_{H^{J+1}_G}) \\ &+ \varepsilon^{7/8} a^{-3-c\sqrt{\varepsilon}} \|\Sigma\|^2_{H^J_G} + \varepsilon^{7/8} a^{-3-c\sqrt{\varepsilon}} \|N\|^2_{H^J_G} \\ &+ \varepsilon^{15/8} a^{-3-c\sqrt{\varepsilon}} \|\Gamma - \widehat{\Gamma}\|^2_{H^{J-1}_G}.\end{aligned}$$

Further, we analogously get

$$-\partial_t \|\Gamma - \widehat{\Gamma}\|^2_{L^2_G} \lesssim \varepsilon^{1/8} a^{-3} \|\Gamma - \widehat{\Gamma}\|^2_{L^2_G} + \varepsilon^{-1/8} a^{-3} (\|\Sigma\|^2_{H^1_G} + \|N\|^2_{H^1_G}),$$

and thus, with the Gronwall lemma and (2-7),

$$\begin{aligned}\|\Gamma - \widehat{\Gamma}\|^2_{L^2_G(\Sigma_t)} &\lesssim a^{-c\varepsilon^{1/8}} \Big(\varepsilon^4 + \int_t^{t_0} \varepsilon^{-1/8} a(s)^{-3} (\|\Sigma\|^2_{H^1_G(\Sigma_s)} + \|N\|^2_{H^1_G(\Sigma_s)})\, ds \Big) \\ &\lesssim a^{-c\varepsilon^{1/8}} \big(\varepsilon^4 + \varepsilon^{-1/4} \sup_{s \in (t, t_0)} (\|\Sigma\|^2_{H^1_G(\Sigma_s)} + \|N\|^2_{H^1_G(\Sigma_s)}) \big).\end{aligned}$$

This proves (6-21) for $l = 0$, and we assume for an iterative argument that the statement has been proved for $l = J - 1$. Then, we obtain (estimating the error terms in Σ and N by their supremum immediately)

$$\begin{aligned}-\partial_t \|\Gamma - \widehat{\Gamma}\|^2_{\dot{H}^J_G} \lesssim\ & \varepsilon^{1/8} a^{-3} \|\Gamma - \widehat{\Gamma}\|^2_{\dot{H}^J_G} + \varepsilon^{-1/8} a^{-3} (\|N\|^2_{H^J_G} + \|\Sigma\|^2_{H^J_G}) \\ &+ \varepsilon^{7/8+4} a^{-3-c\varepsilon^{1/8}} + \varepsilon^{7/8} a^{-3-c\varepsilon^{1/8}} \sup_{s \in (\cdot, t_0)} (\|N\|^2_{H^{J-1}(\Sigma_s)} + \|\Sigma\|^2_{H^{J-1}(\Sigma_s)}).\end{aligned}$$

After integrating, applying the Gronwall lemma and dealing with the first line as before, we get

$$\begin{aligned}\|\Gamma - \widehat{\Gamma}\|^2_{\dot{H}^J_G} \lesssim\ & \varepsilon^4 a^{-c\varepsilon^{1/8}} + \varepsilon^{-1/4} a^{-c\varepsilon^{1/8}} \sup_{s \in (\cdot, t_0]} (\|N\|^2_{H^J_G(\Sigma_s)} + \|\Sigma\|^2_{H^J_G(\Sigma_s)}) \\ &+ \varepsilon^{4+6/8} a^{-c\varepsilon^{1/8}} + \varepsilon^{6/8} a^{-c\varepsilon^{1/8}} \sup_{s \in (\cdot, t_0)} (\|N\|^2_{H^{J-1}(\Sigma_s)} + \|\Sigma\|^2_{H^{J-1}(\Sigma_s)}),\end{aligned}$$

where the second line can obviously be absorbed into the first up to constant. Combining this with the assumption yields (6-21) for $l = J$ and thus iteratively for all $l \leq 17$. □

Lemma 6.14 (Sobolev norm estimates for the metric). *For any $t \in (t_{\mathrm{Boot}}, t_0]$ and any $l \in \mathbb{N}$, $l \leq 18$, we have*

$$\|G-\gamma\|^2_{H^l_G(\Sigma_t)} \lesssim a^{-c\varepsilon^{1/8}}\big(\varepsilon^4 + \varepsilon^{-1/4} \sup_{s\in(t,t_0)} (\|N\|^2_{H^l_G(\Sigma_s)} + \|\Sigma\|^2_{H^l_G(\Sigma_s)})\big). \tag{6-23}$$

Proof. For $l = 0$, we compute the following using (2-28a) and (2-28b):

$$\begin{aligned}-\partial_t\|G-\gamma\|^2_{L^2_G} &= \int_M \Big\{-2(\partial_t G^{-1})^{i_1 j_1}(G^{-1})^{i_2 j_2}(G-\gamma)_{i_1 i_2}(G-\gamma)_{j_1 j_2} - 2\langle \partial_t G, G-\gamma\rangle_G - 3N\frac{\dot a}{a}|G-\gamma|^2_G\Big\}\,\mathrm{vol}_G \\ &= \int_M \Big\{(N+1)a^{-3}[\Sigma * (G-\gamma) + \Sigma] * (G-\gamma) + N\frac{\dot a}{a}|G-\gamma|^2_G - 4N\frac{\dot a}{a}\langle G, G-\gamma\rangle_G\Big\}\,\mathrm{vol}_G.\end{aligned}$$

We apply (4-2b) and (3-17h) and get

$$\begin{aligned}-\partial_t\|G-\gamma\|^2_{L^2_G} &\lesssim \varepsilon a^{-3}\|G-\gamma\|^2_{L^2_G} + a^{-3}(\|\Sigma\|_{L^2_G} + \|N\|_{L^2_G})\|G-\gamma\|_{L^2_G} \\ &\lesssim \varepsilon^{1/8}a^{-3}\|G-\gamma\|^2_{L^2_G} + \varepsilon^{-1/8}a^{-3}(\|\Sigma\|^2_{L^2_G} + \|N\|^2_{L^2_G}).\end{aligned}$$

After integrating and applying the Gronwall lemma (as well as the initial data assumption), we obtain

$$\|G-\gamma\|^2_{L^2_G(\Sigma_t)} \lesssim a^{-c\varepsilon^{1/8}}\Big(\varepsilon^4 + \varepsilon^{-1/8}\int_t^{t_0} a(s)^{-3}(\|\Sigma\|^2_{L^2_G(\Sigma_s)} + \|N\|^2_{L^2_G(\Sigma_s)})\,ds\Big).$$

The statement for $l = 0$ now follows taking the supremum over the norms under the integral and applying (2-7). This extends to higher orders via the same iteration argument as in Lemma 6.13. □

7. Big bang stability: improving the bootstrap assumptions

In this section, we combine the energy estimates obtained in the last two sections to improve the boostrap assumptions for the energies themselves, and then show how this improves the behaviour of the solution norms. For an outline of the energy improvement arguments that we perform in Section 7.1, we refer back to Remark 6.1.

Before carrying out the improvements themselves, we quickly collect an estimate that shows that combining Lemmas 6.6 and 6.8 yields sufficient control on the energies themselves:

Lemma 7.1. *Let $L \in 2\mathbb{N}$. Then, the following estimate is satisfied*:

$$16\pi C^2 a^{-3}(N+1)\langle \Delta^{L/2}\boldsymbol{E}, \Delta^{L/2}\Sigma\rangle_G + 8\pi C^2\frac{\dot a}{a}(N+1)|\Delta^{L/2}\Sigma|^2_G + 6\frac{\dot a}{a}(N+1)|\Delta^{L/2}\boldsymbol{E}|^2_G \geq 0. \tag{7-1}$$

Proof. First, we recall that $N + 1 > 0$ holds by Lemma 4.1. Additionally, we can apply (2-8) and the Young inequality and get

$$\begin{aligned}|16\pi C^2 a^{-3}(N+1)\langle \Delta^{L/2}\boldsymbol{E}, \Delta^{L/2}\Sigma\rangle_G| &\leq 16\pi C^2 \cdot \sqrt{\frac{3}{4\pi C^2}}\frac{\dot a}{a}\cdot(N+1)|\Delta^{L/2}\boldsymbol{E}|_G|\Delta^{L/2}\Sigma|_G \\ &\leq 4(N+1)\frac{\dot a}{a}\cdot(\sqrt{3}\cdot|\Delta^{L/2}\boldsymbol{E}|_G)\cdot(\sqrt{4\pi C^2}|\Delta^{L/2}\Sigma|_G) \\ &\leq 6\frac{\dot a}{a}(N+1)|\Delta^{L/2}\boldsymbol{E}|^2_G + 8\pi C^2\frac{\dot a}{a}(N+1)|\Delta^{L/2}\Sigma|^2_G. \qquad \square\end{aligned}$$

This shows that $\mathcal{E}^{(L)}(W,\cdot) + 4\pi C^2\mathcal{E}^{(L)}(\Sigma,\cdot)$ is controlled by the sum of the left-hand sides of the inequalities in Lemmas 6.6 and 6.8 for $L \in 2\mathbb{N}$, $0 \leq L \leq 18$.

7.1. *Improving energy bounds.*

Proposition 7.2 (improved energy bounds). *Under the bootstrap assumptions* (*see Assumption 3.16*) *and the initial data assumptions in Assumption 3.10, the following improved estimates hold on* $(t_{\mathrm{Boot}}, t_0]$:

$$\mathcal{E}^{(\leq 18)}(\phi, \cdot) \lesssim \varepsilon^4 a^{-c\varepsilon^{1/8}}, \tag{7-2a}$$

$$\mathcal{E}^{(\leq 18)}(\Sigma, \cdot) \lesssim \varepsilon^{15/4} a^{-c\varepsilon^{1/8}}, \tag{7-2b}$$

$$\mathcal{E}^{(\leq 18)}(W, \cdot) \lesssim \varepsilon^{15/4} a^{-c\varepsilon^{1/8}}, \tag{7-2c}$$

$$\mathcal{E}^{(\leq 16)}(\mathrm{Ric}, \cdot) \lesssim \varepsilon^{7/2} a^{-c\varepsilon^{1/8}}, \tag{7-2d}$$

$$\mathcal{E}^{(\leq 16)}(N, \cdot) + a^4 \mathcal{E}^{(17)}(N, \cdot) + a^8 \mathcal{E}^{(18)}(N, \cdot) \lesssim \varepsilon^{7/2} a^{8-c\varepsilon^{1/8}}. \tag{7-2e}$$

Proof. We prove this estimate by performing an induction over even energy orders. Starting at order 0, we first observe that by Lemma 7.1, we can bound the (base level) total energy

$$\mathcal{E}^{(0)}_{\mathrm{total}} := \mathcal{E}^{(0)}(\phi, \cdot) + \varepsilon^{1/4}\big(\mathcal{E}^{(0)}(W, \cdot) + 4\pi C^2 \mathcal{E}^{(0)}(\Sigma, \cdot)\big) + a^4 \mathcal{E}^{(1)}(\phi, \cdot) + \varepsilon^{1/2} \mathcal{E}^{(1)}(\Sigma, \cdot)$$

by the sum of the left-hand side of (6-2), the left-hand side of (6-10) weighted by $\varepsilon^{1/4}$ and the left-hand side of (6-13) weighted by $4\pi C^2 \cdot \varepsilon^{1/4}$, and the left-hand sides of (6-9) and $\varepsilon^{1/2} \cdot$ (6-15) extended to $L = 0$.[10] Combining said estimates and inserting the initial data assumption from (3-11), the following holds in total:

$$\mathcal{E}^{(0)}_{\mathrm{total}}(t) \lesssim \varepsilon^4 + \int_t^{t_0} (\varepsilon^{1/8} a(s)^{-3} + a(s)^{-1-c\sqrt{\varepsilon}}) \mathcal{E}^{(0)}_{\mathrm{total}}(s)\, ds. \tag{7-3}$$

Applying the Gronwall lemma (see Lemma A.1) to (7-3), we get for some suitable constant $c' > 0$

$$\mathcal{E}^{(0)}_{\mathrm{total}}(t) \lesssim \varepsilon^4 \exp\Big(c' \int_t^{t_0} \varepsilon^{1/8} a(s)^{-3} + a(s)^{-1-c\sqrt{\varepsilon}}\, ds\Big) \lesssim \varepsilon^4 a^{-c'\varepsilon^{1/8}}.$$

Now assume that, for $L \in 2\mathbb{N},\ 2 \leq L \leq 18$, we have already shown

$$\mathcal{E}^{(\leq L-2)}(\phi, \cdot) + \varepsilon^{1/4}(\mathcal{E}^{(\leq L-2)}(\Sigma, \cdot) + \mathcal{E}^{(\leq L-2)}(W, \cdot)) \lesssim \varepsilon^4 a^{-c\varepsilon^{1/8}} \tag{7-4a}$$

on $(t_{\mathrm{Boot}}, t_0]$. Note that (7-1) means this holds true for $L = 2$ after updating $c > 0$. Further, if $L \geq 4$ holds, we assume

$$\mathcal{E}^{(\leq L-4)}(\mathrm{Ric}, \cdot) \lesssim \varepsilon^{7/2} a^{-c\varepsilon^{1/8}}. \tag{7-4b}$$

We will show that these assumptions hold at $L = 4$ after having shown the induction step for $L = 2$. We define, for $2 \leq L \leq 18$,

$$\mathcal{E}^{(L)}_{\mathrm{total}} := \mathcal{E}^{(L)}(\phi, \cdot) + \varepsilon^{1/4}(\mathcal{E}^{(L)}(W, \cdot) + 4\pi C^2 \mathcal{E}^{(L)}(\Sigma, \cdot)) + a^4 \mathcal{E}^{(L+1)}(\phi, \cdot) + \varepsilon^{1/2} a^4 \mathcal{E}^{(L+1)}(\Sigma, \cdot)$$
$$+ \varepsilon^{1/2} \mathcal{E}^{(L-2)}(\mathrm{Ric}, \cdot) + \varepsilon^{3/4} a^4 \mathcal{E}^{(L-1)}(\mathrm{Ric}, \cdot).$$

[10] We need to weight $\mathcal{E}^{(0)}(\Sigma, \cdot)$ in the total energy by $K \cdot \varepsilon^{1/4}$ for some $K > 0$ to balance out the $\varepsilon^{-1/8}$-weight from the scalar field energy on the right-hand side of (6-13). The weight on the Bel–Robinson energy is then needed to obtain the cancellation in Lemma 7.1. The additional weight on $a^4 \mathcal{E}^{(1)}(\Sigma, \cdot)$ is needed so that the div-curl-estimates only generates a term of size $\varepsilon^{1/4} \mathcal{E}^{(L)}_{\mathrm{total}}$ that can be absorbed later in the argument.

We combine the respective energy estimates with the appropriate scalings,[11] namely (in the listed order) (6-3), (6-11), (6-14), (6-9), (6-15), (6-18) and (6-20). Observe that the sum of these scaled left-hand sides controls $\mathcal{E}^{(L)}_{\text{total}}$ by Lemma 7.1. Combining all of these estimates and inserting the initial data assumption (3-11), we get the following estimate:

$$\mathcal{E}^{(L)}_{\text{total}}(t) \lesssim \varepsilon^4 + \int_t^{t_0} (\varepsilon^{1/8} a(s)^{-3} + a(s)^{-1-c\sqrt{\varepsilon}})\, \mathcal{E}^{(L)}_{\text{total}}(s)\, ds \tag{7-5a}$$

$$+ \int_t^{t_0} \big\{ \varepsilon^{3/8} a(s)^{-3-c\sqrt{\varepsilon}} [\mathcal{E}^{(\leq L-2)}(\phi, s) + \mathcal{E}^{(\leq L-2)}(\Sigma, s)] \tag{7-5b}$$

$$+ \varepsilon^{17/8} a(s)^{-3-c\sqrt{\varepsilon}} \mathcal{E}^{(\leq L-2)}(W, s) + \underbrace{\varepsilon^{11/8} \mathcal{E}^{(\leq L-4)}(\text{Ric}, s)}_{\text{if } L=4} \big\}\, ds \tag{7-5c}$$

$$+ \varepsilon^{1/4} (a(t)^{4-c\sqrt{\varepsilon}} + \varepsilon a(t)^{2-c\sqrt{\varepsilon}}) \cdot \varepsilon^{1/2} \mathcal{E}^{(L)}(\Sigma, t) + \varepsilon^{1/2} \mathcal{E}^{(L)}(\phi, t) + \varepsilon^{1/4} \cdot \varepsilon^{1/4} \mathcal{E}^{(L)}(W, t) \tag{7-5d}$$

$$+ \varepsilon^{7/4} \cdot \varepsilon^{3/4} a^4 \mathcal{E}^{(L-1)}(\text{Ric}, t) + \varepsilon^{5/2} a^{-c\sqrt{\varepsilon}} \mathcal{E}^{(\leq L-2)}(\phi, t) + \varepsilon^{1/2} a^{2-c\sqrt{\varepsilon}} \mathcal{E}^{(\leq L-2)}(\Sigma, t) \tag{7-5e}$$

$$+ \varepsilon^{3/2} a^{2-c\sqrt{\varepsilon}} \mathcal{E}^{(\leq L-2)}(\text{Ric}, t). \tag{7-5f}$$

We briefly summarize which inequalities contain the listed error term bounds as explicit terms: The first two terms in (7-5b) come from (6-18) and the latter from (6-11), those in (7-5c) from (6-3) and (6-18), and finally the last three lines are precisely the scaled right-hand side of (6-15). Regarding the curvature energies in the various individual energy estimates, any summand with $\mathcal{E}^{(L-3)}(\text{Ric}, \cdot)$ can be split using (3-8), the resulting summands containing $\mathcal{E}^{(L-2)}(\text{Ric}, \cdot)$ can always be absorbed into the total energy term in the first line, and anything with $\mathcal{E}^{(\leq L-4)}(\text{Ric}, \cdot)$ is tracked in $\langle \text{Err} \rangle_L$ for $L \geq 4$.

Inserting (7-4a)–(7-4b), (7-5b)–(7-5c) can be bounded up to constant by $\varepsilon^{33/8} a^{-3-c\varepsilon^{1/8}}$. Here, the error term dominating all others arises from

$$\varepsilon^{3/8} a(s)^{-3-c\sqrt{\varepsilon}} \mathcal{E}^{(\leq L-2)}(\Sigma, s).$$

Regarding (7-5d)–(7-5f), notice that the first four summands can be bounded from above by $\varepsilon^{1/4} \mathcal{E}^{(L)}_{\text{total}}(t)$ up to constant. For the remaining three terms, we can again insert the induction assumptions (7-4a)–(7-4b), bounding them by $\varepsilon^{17/4} a(t)^{-c\varepsilon^{1/8}}$.

In summary and after rearranging, for some constant $K > 0$, (7-5a)–(7-5f) becomes

$$\begin{aligned}
&(1 - K\varepsilon^{1/4})\, \mathcal{E}^{(L)}_{\text{total}}(t) \\
&\quad \lesssim \varepsilon^4 + \int_t^{t_0} (\varepsilon^{1/8} a(s)^{-3} + a(s)^{-1-c\sqrt{\varepsilon}})\, \mathcal{E}^{(L)}_{\text{total}}(s)\, ds + \int_t^{t_0} \varepsilon^{33/8} a(s)^{-3-c\varepsilon^{1/8}}\, ds + \varepsilon^{17/4} a(t)^{-c\varepsilon^{1/8}} \\
&\quad \lesssim \varepsilon^4 a(t)^{-c\varepsilon^{1/8}} + \int_t^{t_0} (\varepsilon^{1/8} a(s)^{-3} + a(s)^{-1-c\sqrt{\varepsilon}})\, \mathcal{E}^{(L)}_{\text{total}}(s)\, ds.
\end{aligned}$$

The prefactor on the left-hand side is positive for small enough $\varepsilon > 0$, and the Gronwall lemma then yields

$$\mathcal{E}^{(L)}_{\text{total}}(t) \lesssim \varepsilon^4 a^{-c\varepsilon^{1/8}}. \tag{7-6}$$

[11] The weights on all terms beside the curvature energies are necessary for the same reasons as at order 0. We need to scale the curvature energy at order L by $\varepsilon^{1/2}$ to account for $\varepsilon^{-1/8} a^{-3} \mathcal{E}^{(L)}(\Sigma, \cdot)$ in (6-18), and the weight on the $(L+1)$-order curvature energy again needs to be chosen according to that on $\mathcal{E}^{(L+1)}(\Sigma, \cdot)$.

In particular, this directly implies that the induction assumptions (7-4a) and (7-4b), using (3-9) to cover the skipped odd order, hold at order L, completing the induction step, and clearly also that (7-4b) holds for $L-2=2$ using (7-6) at order 2. This completes the induction argument, proving (7-2a)–(7-2d). Finally, applying the obtained improved estimates for $\nabla\phi$ and $\mathrm{Ric}[G]$ to Corollary 5.8, we also get (7-2e). □

7.2. ***Improving solution norm control.*** To close the bootstrap argument, it now remains to show that the improved energy bounds also imply improved bounds for $\mathcal{H}$ and $\mathcal{C}$. The former follows almost directly using Lemma 4.5:

Corollary 7.3 (improved Sobolev norm bounds). *On* $(t_{\mathrm{Boot}}, t_0]$, *the following estimates hold*:

$$\mathcal{H} \lesssim \varepsilon^{7/4} a^{-c\varepsilon^{1/8}}, \tag{7-7a}$$

$$\|\Sigma\|^2_{H^{18}_G} \lesssim \varepsilon^{15/4} a^{-c\varepsilon^{1/8}}, \tag{7-7b}$$

$$\|N\|^2_{H^{18}_G} \lesssim \varepsilon^{4} a^{-c\varepsilon^{1/8}}. \tag{7-7c}$$

Proof. First, we apply the improved energy estimates from Proposition 7.2 as well as the strong C_G-norm bounds from Lemma 4.3 to the near-coercivity estimates in Lemma 4.5. With this, we directly obtain the following Sobolev norm estimates (updating c):

$$\begin{aligned}
\|\Psi\|^2_{H^{18}_G} &\lesssim \varepsilon^4 a^{-c\varepsilon^{1/8}} + \varepsilon a^{-c\varepsilon^{1/8}} \cdot \varepsilon^{15/4} a^{-c\varepsilon^{1/8}} \lesssim \varepsilon^4 a^{-c\varepsilon^{1/8}},\\
\|\Sigma\|^2_{H^{18}_G} &\lesssim \varepsilon^{15/4} a^{-c\varepsilon^{1/8}},\\
\big\|\mathrm{Ric}[G] + \tfrac{2}{9} G\big\|^2_{H^{16}_G} &\lesssim \varepsilon^{7/2} a^{-c\varepsilon^{1/8}},\\
\|\boldsymbol{E}\|^2_{H^{18}_G} + \|\boldsymbol{B}\|^2_{H^{18}_G} &\lesssim \varepsilon^{15/4} a^{-c\varepsilon^{1/8}}.
\end{aligned}$$

By Lemma 6.5, we also have

$$\|\nabla\phi\|_{H^{17}_G} \lesssim (1+\varepsilon a^{-c\sqrt{\varepsilon}})\|\Sigma\|_{H^{18}_G} + \varepsilon\|\Psi\|_{H^{18}_G} \lesssim \varepsilon^{15/4} a^{-c\varepsilon^{1/8}}. \tag{7-8}$$

We take particular care in showing that the improved bound holds for $a^2\|\nabla\phi\|_{\dot{H}^{18}_G}$: First, note that (7-2d) implies $\mathcal{E}^{(\le 17)}(\mathrm{Ric},\cdot) \lesssim \varepsilon^{7/2} a^{-c\varepsilon^{1/8}}$. Applying this along with (4-4e) to (4-13d), as well as (4-13a) in the second line and (7-2a) as well as (7-8) in the final step, we obtain

$$\begin{aligned}
a^4\|\nabla\phi\|^2_{H^{18}_G} &\lesssim a^4\|\nabla\Delta^9\phi\|^2_{L^2_G} + a^{4-c\sqrt{\varepsilon}} \sum_{m=0}^{8} \|\nabla\Delta^m\phi\|^2_{L^2_G} + \varepsilon a^{4-c\sqrt{\varepsilon}} \cdot \mathcal{E}^{(\le 16)}(\mathrm{Ric},\cdot)\\
&\lesssim a^{4-c\sqrt{\varepsilon}}(\|\nabla\Delta^9\phi\|^2_{L^2_G} + \|\nabla\phi\|^2_{H^{17}_G}) + \varepsilon^{9/2} a^{-c\varepsilon^{1/8}}\\
&\lesssim a^{-c\sqrt{\varepsilon}} \cdot \mathcal{E}^{(\le 18)}(\phi,\cdot) + a^{4-c\sqrt{\varepsilon}}\|\nabla\phi\|^2_{H^{17}_G} + \varepsilon^{9/2} a^{-c\varepsilon^{1/8}}\\
&\lesssim \varepsilon^{15/4} a^{-c\varepsilon^{1/8}}.
\end{aligned}$$

Further, inserting (7-2a), (7-2b) and (7-2d) into Corollary 5.5 implies

$$a^8\|\Delta^{10}N\|^2_{L^2_G} + a^4\|\nabla\Delta^9 N\|^2_{L^2_G} + \sum_{m=0}^{9} \|\Delta^m N\|^2_{L^2_G} \lesssim \varepsilon^{11/4} a^{-c\varepsilon^{1/8}}$$

and subsequently, applying Lemma 4.5 as before,

$$a^8\|N\|^2_{\dot{H}^{20}_G}+a^4\|N\|^2_{\dot{H}^{19}_G}+\|N\|^2_{H^{18}_G}\lesssim\varepsilon^4a^{-c\varepsilon^{1/8}}.$$

Finally, having now shown (7-7b) and (7-7c), we can apply these to (6-23) to get

$$\|G-\gamma\|^2_{H^{18}_G}\lesssim a^{-c\varepsilon^{1/8}}(\varepsilon^4+\varepsilon^{-1/4+15/4}+\varepsilon^{-1/4+4})\lesssim\varepsilon^{7/2}a^{-c\varepsilon^{1/8}},$$

proving (7-7a). □

Intuitively, the bounds on $\mathcal{C}$ should now follow from $\mathcal{H}$ by the standard Sobolev embedding. However, since both of these norms are with respect to G, the embedding constant may be time-dependent. To circumvent this issue, we need to switch between norms with respect to G and γ and then apply the embedding with respect to C_γ and H_γ. The following lemma ensures that we still obtain bootstrap improvements after performing these norm switches:

Lemma 7.4 (moving between norms). *Let $l\in\mathbb{N}$, $l\leq 18$, ζ be a scalar field, $\mathfrak{T}$ be an arbitrary Σ_t-tangent tensor. Then, for some multivariate polynomial P_l with $P_l(0,0)=0$, we have*

$$\|\zeta\|_{C^l_G(U)}\lesssim a^{-c\sqrt{\varepsilon}}\|\zeta\|_{C^l_\gamma(M)}+a^{-c\sqrt{\varepsilon}}\|\zeta\|_{C^{\max\{0,\lfloor(l-1)/2\rfloor\}}_\gamma(M)}\cdot P_l(\|G-\gamma\|_{C^{l-1}_\gamma(M)},\|G^{-1}-\gamma^{-1}\|_{C^{l-1}_\gamma(M)}),\quad(7\text{-}9a)$$

$$\|\mathfrak{T}\|_{C^l_G(M)}\lesssim a^{-c\sqrt{\varepsilon}}\|\mathfrak{T}\|_{C^l_\gamma(M)}+a^{-c\sqrt{\varepsilon}}\|\mathfrak{T}\|_{C^{\max\{0,\lfloor(l-1)/2\rfloor\}}_\gamma(M)}\cdot P_l(\|G-\gamma\|_{C^l_\gamma(M)},\|G^{-1}-\gamma^{-1}\|_{C^l_\gamma(M)}),\quad(7\text{-}9b)$$

as well as the same inequalities with the roles of G and γ reversed. For $l\leq 12$, this reduces to

$$a^{c\sqrt{\varepsilon}}\|\zeta\|_{C^l_\gamma(M)}\lesssim\|\zeta\|_{C^l_G(M)}\lesssim a^{-c\sqrt{\varepsilon}}\|\zeta\|_{C^l_\gamma(M)},\quad(7\text{-}10a)$$

$$a^{c\sqrt{\varepsilon}}\|\mathfrak{T}\|_{C^l_\gamma(M)}\lesssim\|\mathfrak{T}\|_{C^l_G(M)}\lesssim a^{-c\sqrt{\varepsilon}}\|\mathfrak{T}\|_{C^l_\gamma(M)}.\quad(7\text{-}10b)$$

Further, one has

$$\|\zeta\|^2_{H^l_\gamma(M)}\lesssim a^{-c\sqrt{\varepsilon}}\|\zeta\|^2_{H^l_G(M)}+a^{-c\varepsilon^{1/8}}\|\zeta\|^2_{C^{\lceil(l-1)/2\rceil}_G(\Sigma_t)}\big(\varepsilon^4+\varepsilon^{-1/4}\sup_{s\in(t,t_0)}(\|N\|^2_{H^{l-1}_G(\Sigma_s)}+\|\Sigma\|^2_{H^{l-1}_G(\Sigma_s)})\big),\quad(7\text{-}11a)$$

$$\|\mathfrak{T}\|^2_{H^l_\gamma(M)}\lesssim a^{-c\sqrt{\varepsilon}}\|\mathfrak{T}\|^2_{H^l_G(M)}+a^{-c\varepsilon^{1/8}}\|\mathfrak{T}\|^2_{C^{\lceil(l-1)/2\rceil}_G(\Sigma_t)}\big(\varepsilon^4+\varepsilon^{-1/4}\sup_{s\in(t,t_0)}(\|N\|^2_{H^l_G(\Sigma_s)}+\|\Sigma\|^2_{H^l_G(\Sigma_s)})\big).\quad(7\text{-}11b)$$

Remark 7.5. While we only need the tensorial inequalities for gradient vector fields and (0, 2)-tensors when applied to norms in $\mathcal{H}$ and $\mathcal{C}$, the proof is simpler when considering tensors of arbitrary rank.

Proof. We restrict ourselves to proving the tensorial statements; the scalar field analogues follow analagously except for the fact that, since $\nabla_i\zeta=\hat{\nabla}_i\zeta=\partial_i\zeta$, error terms caused by Christoffel symbols always enter at one order less. Thus, it remains to show (7-9b), (7-10b) and (7-11b) by iterating over derivative order.

Starting with the base level estimates, we have if $\mathfrak{T}$ is of rank (r,s)

$$|\mathfrak{T}|^2_G-|\mathfrak{T}|^2_\gamma=[G_{i_1j_1}\cdots G_{i_rj_r}(G^{-1})^{p_1q_1}\cdots(G^{-1})^{p_sq_s}-\gamma_{i_1j_1}\cdots\gamma_{i_rj_r}(\gamma^{-1})^{p_1q_1}\cdots(\gamma^{-1})^{p_sq_s}]\cdot\mathfrak{T}^{i_1\cdots i_r}{}_{p_1\cdots p_s}\mathfrak{T}^{j_1\cdots j_r}{}_{q_1\cdots q_s}.$$

We successively replace $G^{\pm 1}$ by $(G^{\pm 1} - \gamma^{\pm 1}) + \gamma^{\pm 1}$, take the $|\cdot|_\gamma$-norm of each factor and use (4-4c)–(4-4d). This yields

$$\big||\mathfrak{T}|_G^2 - |\mathfrak{T}|_\gamma^2\big| \lesssim \sqrt{\varepsilon} a^{-c\sqrt{\varepsilon}} |\mathfrak{T}|_\gamma^2,$$

implying (7-9b) (and (7-10b)) for $l = 0$ after rearranging and taking supremums suitably.

To show (7-11b) at base level, consider

$$\int_M |\mathfrak{T}|_G^2 \operatorname{vol}_G - \int_M |\mathfrak{T}|_\gamma^2 \operatorname{vol}_\gamma = \int_M (|\mathfrak{T}|_G^2 - |\mathfrak{T}|_\gamma^2) \operatorname{vol}_G + \int_M |\mathfrak{T}|_\gamma^2 \frac{\mu_G - \mu_\gamma}{\mu_\gamma} \operatorname{vol}_\gamma.$$

We can control the first summand on the right-hand side as before, while we have $|\mu_G - \mu_\gamma| \lesssim \varepsilon$ by (4-11). Hence,

$$(1 - K\varepsilon)\|\mathfrak{T}\|_{L^2_\gamma}^2 \lesssim (1 + \sqrt{\varepsilon} a^{-c\sqrt{\varepsilon}})\|\mathfrak{T}\|_{L^2_G}^2$$

follows for a suitable constant $K > 0$, implying the statement for small enough $\varepsilon > 0$.

Next, we perform the iteration for (7-9b), assuming the statement and the analogue with γ and G reversed to hold up to order $l - 1$. As above, note that

$$\big||\nabla^J \mathfrak{T}|_G^2 - |\hat{\nabla}^J \mathfrak{T}|_\gamma^2\big| \lesssim \sqrt{\varepsilon} a^{-c\sqrt{\varepsilon}} |\hat{\nabla}^J \mathfrak{T}|_\gamma^2 + (1 + \sqrt{\varepsilon} a^{-c\sqrt{\varepsilon}})\big||\nabla^J \mathfrak{T}|_\gamma^2 - |\hat{\nabla}^J \mathfrak{T}|_\gamma^2\big|,$$

where we can rewrite the second term as

$$\big|2\langle \hat{\nabla}^J \mathfrak{T} - \nabla^J \mathfrak{T}, \hat{\nabla} \mathfrak{T}\rangle_\gamma - |\nabla^J \mathfrak{T} - \hat{\nabla}^J \mathfrak{T}|_\gamma^2\big|$$

and hence obtain (moving between pointwise norms as before)

$$\big||\nabla^J \mathfrak{T}|_G^2 - |\hat{\nabla}^J \mathfrak{T}|_\gamma^2\big| \lesssim a^{-c\sqrt{\varepsilon}} |\hat{\nabla}^J \mathfrak{T}|_\gamma^2 + a^{-c\sqrt{\varepsilon}} |\nabla^J \mathfrak{T} - \hat{\nabla}^J \mathfrak{T}|_\gamma^2.$$

Regarding $\nabla^J \mathfrak{T} - \hat{\nabla}^J \mathfrak{T}$, we have the following schematic decomposition:

$$\nabla^J \mathfrak{T} - \hat{\nabla}^J \mathfrak{T} = \sum_{I=0}^{J-1} \hat{\nabla}^{J-I-1}(\Gamma - \widehat{\Gamma}) *_\gamma (\nabla^I \mathfrak{T} + \hat{\nabla}^I \mathfrak{T}) + \langle \text{at least cubic nonlinear terms} \rangle. \tag{7-12}$$

Here, $*_\gamma$ encodes the analogous schematic product notation with regard to γ (see Section 2.1.8). Regarding the Christoffel symbols, notice (7-9b) with roles of γ and G reversed holding up to $l - 1$ implies that, for any $m \in \{0, \dots, l-1\}$ and some multivariate polynomial $\widetilde{P}_m$, we have

$$\|\Gamma - \widehat{\Gamma}\|_{C^m_\gamma(M)} \lesssim a^{-c\sqrt{\varepsilon}} \widetilde{P}_m(\|\Gamma - \widehat{\Gamma}\|_{C^m_G(M)}, \|G - \gamma\|_{C^m_G(M)}, \|G^{-1} - \gamma^{-1}\|_{C^m_G(M)}).$$

As explained in Remark 3.7, we can bound $\|\Gamma - \widehat{\Gamma}\|_{C^m_G(M)}$ by a polynomial in $\|G - \gamma\|_{C^{m+1}_G(M)}$. Hence, we can apply (4-4c) to obtain

$$\|\Gamma - \widehat{\Gamma}\|_{C^{l-1}_\gamma(M)} \lesssim \sqrt{\varepsilon} a^{-c\sqrt{\varepsilon}}. \tag{7-13}$$

Moving back to (7-12) and just considering the first line for now, this implies

$$\begin{aligned}\big|\|\mathfrak{T}\|^2_{\dot{C}^l_G(M)} - \|\mathfrak{T}\|^2_{\dot{C}^l_\gamma(M)}\big| \lesssim a^{-c\sqrt{\varepsilon}}\Big(\|\mathfrak{T}\|^2_{C^l_\gamma(M)} + \sum_{m=0}^{l-1}\|\nabla^m\mathfrak{T}\|^2_{C^0_\gamma(M)}\Big)\\ + \Big(\|\mathfrak{T}\|^2_{C^{\lceil (l-1)/2\rceil}_\gamma(M)} + \sum_{m=0}^{\lceil (l-1)/2\rceil}\|\nabla^m\mathfrak{T}\|^2_{C^0_\gamma(M)}\Big)\|\Gamma-\widehat{\Gamma}\|^2_{C^{l-1}_\gamma(M)}\\ + \langle\text{at least cubic nonlinear terms}\rangle. \qquad (7\text{-}14)\end{aligned}$$

We can rewrite $\nabla^m\mathfrak{T}$-norms in C_γ as ones in C_G up to $a^{-c\sqrt{\varepsilon}}$ as before. Then, we can apply the already-obtained estimates up to order $l-1$ show that the first two lines of the right-hand side can be estimated by the right-hand side of (7-9b). The highly nonlinear terms can be dealt with similarly, closing the induction over admissible l. The estimate (7-10b) immediately follows by applying (4-4c)–(4-4d) and (7-13).

Now, assume (7-11b) to be proven up to order $J-1$. By analogous arguments as at order zero, we get, after rearranging,

$$\int_M|\hat{\nabla}^J\mathfrak{T}|^2_\gamma\,\mathrm{vol}_\gamma \lesssim \Big|\int_M(|\nabla^J\mathfrak{T}|^2_G - |\hat{\nabla}^J\mathfrak{T}|^2_\gamma)\,\mathrm{vol}_G\Big| + \int_M|\nabla^J\mathfrak{T}|^2_G\,\mathrm{vol}_G,$$

so we only need to concern ourselves with the first summand. Reversing the roles of G and γ compared to the proof of (7-9b), we get

$$\big|\,|\nabla^J\mathfrak{T}|^2_G - |\hat{\nabla}^J\mathfrak{T}|^2_\gamma\big| \lesssim \sqrt{\varepsilon}a^{-c\sqrt{\varepsilon}}|\nabla^J\mathfrak{T}|^2_G + a^{-c\sqrt{\varepsilon}}\big|2\langle\nabla^J\mathfrak{T}-\hat{\nabla}^J\mathfrak{T},\nabla\mathfrak{T}\rangle_G - |\nabla^J\mathfrak{T}-\hat{\nabla}^J\mathfrak{T}|^2_G\big|,$$

and have the following, applying Lemma 6.13 immediately to estimate $\|\Gamma-\widehat{\Gamma}\|_{H^{l-1}_G}$:

$$\begin{aligned}&\Big|\int_V\{2\langle\nabla^l\mathfrak{T}-\hat{\nabla}^l\mathfrak{T},\nabla\mathfrak{T}\rangle_G - |\nabla^l\mathfrak{T}-\hat{\nabla}^l\mathfrak{T}|^2_G\}\,\mathrm{vol}_G\Big|\\ &\lesssim a^{-c\sqrt{\varepsilon}}(\|\mathfrak{T}\|^2_{H^l_G(M)} + \|\hat{\nabla}^{\le l-1}\mathfrak{T}\|^2_{H^0_G(M)})\\ &\quad + (\|\mathfrak{T}\|^2_{C^{\lceil (l-1)/2\rceil}_G(M)} + \|\hat{\nabla}^{\le\lceil (l-1)/2\rceil}\mathfrak{T}\|^2_{C^0_G(M)})\cdot a^{-c\varepsilon^{1/8}}\big(\varepsilon^4+\varepsilon^{-1/4}\sup_{s\in(\cdot,t_0)}(\|N\|^2_{H^l_G(\Sigma_s)}+\|\Sigma\|^2_{H^l_G(\Sigma_s)})\big).\end{aligned}$$

By the same arguments as earlier, we have $\|\hat{\nabla}^{\le l-1}\mathfrak{T}\|_{H^{l-1}_G(M)} \lesssim a^{-c\sqrt{\varepsilon}}\|\mathfrak{T}\|_{H^{l-1}_\gamma(M)}$ and can then apply the induction hypothesis. This proves (7-11b). □

Corollary 7.6 (improved C-norm bounds). *On* $(t_{\mathrm{Boot}}, t_0]$, *the following estimate is satisfied*:

$$\mathcal{C}+\mathcal{C}_\gamma \lesssim \varepsilon^{7/4}a^{-c\varepsilon^{1/8}}. \qquad (7\text{-}15)$$

Proof. We first apply the Sobolev norm estimates in Lemma 7.4 to (7-7a), to then control $\mathcal{C}_\gamma$ via the standard Sobolev embedding $H^{l+2}_\gamma(M)\hookrightarrow C^l(M)$, and finally control $\mathcal{C}$ with (7-9a)–(7-9b).

Note that by Lemma 4.3, we can control the C_G-norm up to order 10 of every quantity occurring in $\mathcal{H}$ beside the lapse by at worst $\sqrt{\varepsilon}a^{-c\sqrt{\varepsilon}}$, while the bootstrap assumption already implies better behaviour for the lapse. Thus, we can apply (7-11a)–(7-11b) to every norm appearing in $\mathcal{H}$, and obtain by applying (7-7b) and (7-7c) in the second line

$$\begin{aligned}\mathcal{C}^2_\gamma &\lesssim a^{-c\sqrt{\varepsilon}}\cdot\mathcal{H}^2 + \varepsilon a^{-c\sqrt{\varepsilon}}\cdot a^{-c\varepsilon^{1/8}}\big(\varepsilon^4+\varepsilon^{-1/4}\sup_{s\in(\cdot,t_0)}(\|N\|^2_{H^{18}_G(\Sigma_s)}+\|\Sigma\|^2_{H^{18}_G(\Sigma_s)})\big)\\ &\lesssim \varepsilon^{7/2}a^{-c\varepsilon^{1/8}} + \varepsilon a^{-c\varepsilon^{1/8}}(\varepsilon^4+\varepsilon^{7/2})\\ &\lesssim \varepsilon^{7/2}\cdot a^{-c\varepsilon^{1/8}}.\end{aligned}$$

In particular, we can update c such that

$$|P(\|G-\gamma\|_{C^{16}_{\gamma}(\Sigma_t)}, \|G-\gamma\|_{C^{16}_{\gamma}(\Sigma_t)})| \lesssim \varepsilon^{7/2} a^{-c\varepsilon^{1/8}}$$

holds for any multivariate polynomial P that appears when applying (7-9a)–(7-9b). Again using the strong C_G-norm estimates from Lemma 4.3, this then implies $\mathcal{C} \lesssim \varepsilon^{7/4} a^{-c\varepsilon^{1/8}}$. □

8. Big bang stability: the main theorem

In this section, we provide the proof of the first main result, Theorem 1.1, which we state in more detail in Theorem 8.2 below. As in [Rodnianski and Speck 2018b; Speck 2018], most of the work has already been done by establishing the necessary bounds on solution norms.

Remark 8.1 (existence of a CMC hypersurface). As mentioned in Section 1.2.1, it may seem that the generality of the results in Theorem 8.2 is restricted by taking the initial data on Σ_{t_0} to be CMC. However, as long as one remains close enough to a constant-time hypersurface of the FLRW reference metric (which is CMC), one can locally evolve the perturbed data in harmonic gauge to a nearby hypersurface that is CMC and remains close to the FLRW reference solution. To make this a bit more precise, and also since this is a little less involved than the arguments in [Rodnianski and Speck 2018b], we will briefly sketch how the arguments from [Fajman and Kröncke 2020, Section 2.5] extend to our setting.

First, we once again assume without loss of generality that our initial data is sufficiently regular. Note that we can locally evolve our data within harmonic gauge to get a C^{17}-regular family of metrics with near-FLRW initial data (for well-posedness, consider the analogue of [Rodnianski and Speck 2018b, Proposition 14.1]). Consider the Banach manifold $\mathcal{M}^{17}$ formed by the set of C^{17} Lorentz metrics on $I \times \overline{M}$ for an open interval I around t_0 such that the surfaces of constant time are Riemannian, endowed with the norm

$$\|\tilde{g}\| = \|\tilde{n}^2\|_{C^{17}_{dt^2+\gamma}(I\times M)} + \|\widetilde{X}\|_{C^{17}_{dt^2+\gamma}(I\times M)} + \|\tilde{g}_t\|_{C^{17}_{dt^2+\gamma}(I\times M)},$$

where $\tilde{g} \in \mathcal{M}^{17}$ has lapse $\tilde{n}$, shift $\widetilde{X}$ and spatial metrics $(\tilde{g}_t)_{t\in I}$. Further, for any $f \in C^{17}(M, I)$, we define the embedding $\iota_f : M \hookrightarrow \overline{M}$ by $x \mapsto (f(x), x)$, and subsequently define the smooth map

$$H_0 : \mathcal{D} := \{(\tilde{g}, f) \in \mathcal{M}^{17} \times C^{17}(M, I) | \iota_f^* \tilde{g} \text{ is Riemannian}\} \to C^{16}(M),$$
$$(\tilde{g}, f) \mapsto \text{mean curvature of } (M, \tilde{g}_t) \text{ embedded along } \iota_f.$$

One easily checks that $(\bar{g}_{\mathrm{FLRW}}, t_0)$ is a regular point of H_0. By the implicit function theorem for Banach manifolds, this means there is a (unique) smooth function F that maps an open neighbourhood of $\bar{g}_{\mathrm{FLRW}}$ in $\mathcal{M}^{17}$ to an open neighbourbood of the constant function $x \mapsto t_0$ in $C^{17}(M, I)$ such that $H_0(\cdot, F(\cdot)) = \tau(t_0)$ holds in that neighbourhood.

Thus, we can choose a surface Σ' with mean curvature $\tau(t_0)$ near the original Σ_{t_0}. Furthermore, for small enough $\varepsilon > 0$, the initial data on Σ' remains close to the FLRW initial data in the sense of Assumption 3.10, using similar arguments to control Sobolev norms. Thus, we can replace Σ_{t_0} by Σ' without loss of generality, proving that the CMC assumption (2-10) is not a true restriction.

Theorem 8.2 (stability of big bang formation). *Let $(M, \mathring{g}, \mathring{k}, \mathring{\pi}, \mathring{\psi})$ be initial data to the Einstein scalar-field system as discussed in Section 1.2.1. Further, let the data be embedded into a time-oriented 4-manifold such that it induces initial data for the rescaled solution variables (see Definition 2.9) at the initial hypersurface Σ_{t_0}. We also assume this rescaled initial data is close to that of the FLRW reference solution (see (2-1) and (2-2)) in the sense that*

$$\mathcal{H}(t_0) + \mathcal{H}_{\mathrm{top}}(t_0) + \mathcal{C}(t_0) \leq \varepsilon^2 \tag{8-1}$$

is satisfied (with $\mathcal{H}$ and $\mathcal{C}$ as in Definition 3.6).[12]

Then, the past maximal globally hyperbolic development $((0, t_0] \times M, \bar{g}, \phi)$ of this data within the Einstein scalar-field system (1-1a)–(1-1c) in CMC gauge (2-10) with zero shift is foliated by the CMC hypersurfaces $\Sigma_s = t^{-1}(\{s\})$, and one has

$$\mathcal{H}(t) + \mathcal{C}(t) + \mathcal{C}_\gamma(t) \lesssim \varepsilon^{7/4} a(t)^{-c\varepsilon^{1/8}} \tag{8-2}$$

for some $c > 0$ and any $t \in (0, t_0]$. In particular, this implies the following statements:

***Asymptotic behaviour of solution variables*:** *We denote the solution metric by $\bar{g} = -n^2\, dt^2 + g$, the second fundamental form (viewed as a (1, 1)-tensor) with respect to Σ_t by k, and the volume form with regard to g on Σ_t by vol_g. There exist a smooth function $\Psi_{\mathrm{Bang}} \in C^{15}_\gamma(M)$, a (1, 1)-tensor field $K_{\mathrm{Bang}} \in C^{15}_\gamma(M)$ and a volume form $\mathrm{vol}_{\mathrm{Bang}} \in C^{15}_\gamma(M)$ such that the following estimates hold for any $t \in (0, t_0]$:*

$$\|n - 1\|_{C^l_\gamma(\Sigma_t)} \lesssim \begin{cases} \varepsilon a(t)^{4-c\varepsilon^{1/8}}, & l \leq 14, \\ \varepsilon a(t)^{2-c\varepsilon^{1/8}}, & l = 15, \end{cases} \tag{8-3a}$$

$$\|a^{-3}\mathrm{vol}_g - \mathrm{vol}_{\mathrm{Bang}}\|_{C^l_\gamma(\Sigma_t)} \lesssim \begin{cases} \varepsilon a(t)^{4-c\varepsilon^{1/8}}, & l \leq 14, \\ \varepsilon a(t)^{2-c\varepsilon^{1/8}}, & l = 15, \end{cases} \tag{8-3b}$$

$$\|a^3 \partial_t \phi - (\Psi_{\mathrm{Bang}} + C)\|_{C^l_\gamma(\Sigma_t)} \lesssim \begin{cases} \varepsilon a(t)^{4-c\varepsilon^{1/8}}, & l \leq 14, \\ \varepsilon a(t)^{2-c\varepsilon^{1/8}}, & l = 15, \end{cases} \tag{8-3c}$$

$$\left\| \phi(t, \cdot) - \phi(t_0, \cdot) + \int_t^{t_0} a(s)^{-3}\, ds \cdot (\Psi_{\mathrm{Bang}} + C) \right\|_{\dot{C}^l_\gamma(\Sigma_t)} \lesssim \begin{cases} \varepsilon a(t)^{4-c\varepsilon^{1/8}}, & 1 \leq l \leq 14, \\ \varepsilon a(t)^{2-c\varepsilon^{1/8}}, & l = 15, \end{cases} \tag{8-3d}$$

$$\|a^3 k - K_{\mathrm{Bang}}\|_{C^l_\gamma(\Sigma_t)} \lesssim \begin{cases} \varepsilon a(t)^{4-c\varepsilon^{1/8}}, & l \leq 14, \\ \varepsilon a(t)^{2-c\varepsilon^{1/8}}, & l = 15. \end{cases} \tag{8-3e}$$

Further, these footprint states satisfy the equations[13]

$$(K_{\mathrm{Bang}})^a{}_a = -\sqrt{12\pi}\, C, \tag{8-4a}$$

$$8\pi(\Psi_{\mathrm{Bang}} + C)^2 + (K_{\mathrm{Bang}})^a{}_b (K_{\mathrm{Bang}})^b{}_a = 12\pi C^2 \tag{8-4b}$$

and remain close to the data of the reference solution in the following sense, where $\mathbb{I}$ denotes the Kronecker symbol:

$$\|\mathrm{vol}_\gamma - \mathrm{vol}_{\mathrm{Bang}}\|_{C^{15}_\gamma(M)} \lesssim \varepsilon, \tag{8-5a}$$

[12] Essentially, this translates to smallness in H^{19}_γ and C^{17}_γ. For $\varepsilon = 0$, the solution is the FLRW reference solution.

[13] These are precisely the (generalized) Kasner relations; see Section 1.2.3.

$$\|\Psi_{\mathrm{Bang}}\|_{C^{15}_{\gamma}(M)} \lesssim \varepsilon, \tag{8-5b}$$

$$\left\| K_{\mathrm{Bang}} + \sqrt{\tfrac{4\pi}{3}} C \mathbb{I} \right\|_{C^{15}_{\gamma}(M)} \lesssim \varepsilon. \tag{8-5c}$$

Additionally, there exists a $(0,2)$*-tensor field* $M_{\mathrm{Bang}} \in C^{15}_{\gamma}(M)$ *satisfying*

$$\|M_{\mathrm{Bang}} - \gamma\|_{C^{15}_{\gamma}(M)} \lesssim \varepsilon \tag{8-6}$$

and, with $\odot$ *and* exp *meant in the matrix product and exponential sense respectively, one has*

$$\left\| g \odot \exp\left[\left(-2 \int_t^{t_0} a(s)^{-3}\, ds \right) \cdot K_{\mathrm{Bang}} \right] - M_{\mathrm{Bang}} \right\|_{C^{l}_{\gamma}(\Sigma_t)} \lesssim \begin{cases} \varepsilon a(t)^{4-c\varepsilon^{1/8}}, & l \leq 14, \\ \varepsilon a(t)^{2-c\varepsilon^{1/8}}, & l = 15. \end{cases} \tag{8-7}$$

Moreover, the Bel–Robinson variables E *and* B *satisfy the estimates*

$$\|E\|_{C^{16}_{\gamma}(\Sigma_t)} \lesssim \varepsilon a^{-4-c\varepsilon^{1/8}}, \tag{8-8a}$$

$$\|B\|_{C^{l}_{\gamma}(\Sigma_t)} \lesssim \begin{cases} \varepsilon a^{-2-c\varepsilon^{1/8}}, & l \leq 15, \\ \varepsilon a^{-4-c\varepsilon^{1/8}}, & l \leq 16. \end{cases} \tag{8-8b}$$

***Causal disconnectedness*:** *Let* $\boldsymbol{\alpha}$ *be a past-directed causal curve on* $((0,t] \times M, \bar{g})$ *for* $t \leq t_0$ *with domain* $[s_1, s_{\max})$ *such that* $\boldsymbol{\alpha}(s_1) \in \Sigma_t$ *and* $s_{\max}$ *is maximal. Then, there exists a constant* $\mathcal{K} > 0$ *that does not depend on* $\boldsymbol{\alpha}$ *such that one has*

$$L[\boldsymbol{\alpha}] = \int_{s_1}^{s_{\max}} \sqrt{(\gamma_{ab})_{\boldsymbol{\alpha}(s)} \dot{\boldsymbol{\alpha}}^a(s) \dot{\boldsymbol{\alpha}}^b(s)}\, ds \leq \mathcal{K} a(t)^{2-c\varepsilon^{1/8}}, \tag{8-9}$$

where γ *is the negative Einstein spatial reference metric on* M (*see Definition 2.1*). *Hence, for points* $p, q \in \Sigma_t$ *with* $\mathrm{dist}_{\gamma}(p,q) > 2\mathcal{K} a(t)^{2-c\varepsilon^{1/8}}$, *the causal pasts of* p *and* q *cannot intersect.*

***Geodesic incompleteness*:** *Let* $\boldsymbol{\alpha}(\mathcal{A})$ *be a past-directed, affinely parametrized causal geodesic emanating from* Σ_{t_0}, *where* $\mathcal{A} : (0, t_0] \to [0, \infty)$ *denotes the parameter time that is normalized to* $\mathcal{A}(t_0) = 0$. *Then,*

$$\mathcal{A}(0) \leq \mathcal{K}_1 \cdot |\mathcal{A}'(t_0)| \cdot a(t_0)^{1+K_2\varepsilon} \int_0^{t_0} a(s)^{-1-\mathcal{K}_2\varepsilon}\, ds < \infty \tag{8-10}$$

holds for suitable constants $\mathcal{K}_1, \mathcal{K}_2 > 0$ *that are independent of* $\boldsymbol{\alpha}$, *and thus any such geodesic crashes into the big bang hypersurface in finite affine parameter time.*

***Blow-up*:** *The norm* $|k|_g$ *behaves toward the big bang hypersurface as follows*:

$$\|a^6 |k|_g^2 - (K_{\mathrm{Bang}})^i{}_j (K_{\mathrm{Bang}})^j{}_i\|_{C^0_{\gamma}(\Sigma_t)} \lesssim \varepsilon a^{4-c\varepsilon^{1/8}}. \tag{8-11a}$$

Further, with $W[\bar{g}]$ *denoting the Weyl curvature and* $P[\bar{g}] = \mathrm{Riem}[\bar{g}] - W[\bar{g}]$,

$$\left\| a^{12} P_{\alpha\beta\gamma\delta} P^{\alpha\beta\gamma\delta} - \tfrac{5}{3} \cdot (8\pi)^2 (\Psi_{\mathrm{Bang}} + C)^4 \right\|_{C^0_{\gamma}(M)} \lesssim \varepsilon a^{4-c\varepsilon^{1/8}} \tag{8-11b}$$

is satisfied, whereas there exists a scalar footprint $W_{\mathrm{Bang}} \in C^{15}_{\gamma}(M)$ *such that one has*

$$\|a^{12} W_{\alpha\beta\gamma\delta} W^{\alpha\beta\gamma\delta} - W_{\mathrm{Bang}}\|_{C^0(M)} \lesssim \varepsilon a^{2-c\varepsilon^{1/8}}. \tag{8-11c}$$

Here, W_{Bang} is a fourth-order polynomial in $\widehat{K}_{\mathrm{Bang}} = K_{\mathrm{Bang}} + \sqrt{\frac{4\pi}{3}} C\mathbb{I}$ and Ψ_{Bang} and satisfies

$$\| W_{\mathrm{Bang}} \|_{C^{15}_{\gamma}(M)} \lesssim \varepsilon. \tag{8-11d}$$

Finally, the scalar curvature $R[\bar{g}]$ and the Ricci curvature invariant $\mathrm{Ric}[\bar{g}]_{\alpha\beta}\,\mathrm{Ric}[\bar{g}]^{\alpha\beta}$ blow up with the asymptotics

$$\| a^6 R[\bar{g}] - 8\pi (\Psi_{\mathrm{Bang}} + C)^2 \|_{C^0(M)} \lesssim \varepsilon a^{4 - c\varepsilon^{1/8}} \tag{8-11e}$$

and

$$\| a^{12}\,\mathrm{Ric}[\bar{g}]_{\alpha\beta}\,\mathrm{Ric}[\bar{g}]^{\alpha\beta} - (8\pi)^2 (\Psi_{\mathrm{Bang}} + C)^4 \|_{C^0(M)} \lesssim \varepsilon a^{4 - c\varepsilon^{1/8}}, \tag{8-11f}$$

and the Kretschmann scalar $\mathcal{K} = \mathrm{Riem}[\bar{g}]_{\alpha\beta\gamma\delta}\,\mathrm{Riem}[\bar{g}]^{\alpha\beta\gamma\delta}$ exhibits stable blow-up in the following sense:

$$\left\| a^{12}\mathcal{K} - \tfrac{5}{3} \cdot (8\pi)^2 (\Psi_{\mathrm{Bang}} + C)^4 - W_{\mathrm{Bang}} \right\|_{C^0(M)} \lesssim \varepsilon a^{2 - c\varepsilon^{1/8}}. \tag{8-11g}$$

Remark 8.3 (The solution variables exhibit AVTD behaviour.). The estimates (8-3a)–(8-3e) and (8-7) imply that the solution is asymptotically velocity term dominated (AVTD) in the sense that, toward the big bang singularity, they behave at leading order like solutions to the (formal) velocity term dominated equations. These arise by dropping any terms containing spatial derivatives in the decomposed Einstein system, i.e., in (2-15a), (2-15b), (2-17a) and (2-18).

Proof. As argued at the end of Section 3.4, we can assume without loss of generality that our initial data is sufficiently regular. Hence, the local existence statement in Lemma 3.14 and the initial data requirements (8-1) ensure that there exists a local solution to the Einstein scalar-field system on $[t_1, t_0] \times M$ and that the bootstrap assumption (see Assumption 3.16) holds on $[t_1, t_0] \times M$ with $t_1 \in (0, t_0)$ and $\sigma = \varepsilon^{1/16}$. Let $\mathfrak{t} \in (0, t_0)$ be such that $(\mathfrak{t}, t_0] \times M$ is the maximal domain on which the solution variables exist and satisfy the bootstrap assumptions. For contradiction, we now assume that $\mathfrak{t} > 0$ were to hold.

Due to Corollary 7.6, there exist (summarizing all updates) constants $c_1, K_1 > 0$ such that, for any $t \in (\mathfrak{t}, t_0]$,

$$\mathcal{C}(t) \leq K_1 \varepsilon^{7/4} a(t)^{-c_1 \varepsilon^{1/8}}. \tag{8-12}$$

If ε is small enough such that $K_1 \varepsilon^{1/8} < K_0$ and $c_1 \varepsilon^{1/8} < c_0 \sigma$ hold, this is a strict improvement of the bootstrap assumption. Furthermore, argued exactly as in the proof of [Rodnianski and Speck 2018b, Theorem 15.1], above improvement ensures none of the blow-up criteria of Lemma 3.14 are satisfied if $\mathfrak{t} > 0$ were to hold, essentially as a direct consequence of (8-12). Hence, the solution could be classically extended to a CMC hypersurface $\Sigma_{\mathfrak{t}}$ diffeomorphic to M, while satisfying the improved estimates by continuity, and further to an interval $(\mathfrak{t}', t_0]$ for some $0 < \mathfrak{t}' < \mathfrak{t}$ on which the bootstrap assumptions must then be satisfied, also by continuity. This contradicts the maximality of $(\mathfrak{t}, t_0]$.

Thus, the rescaled solution variables induce a unique solution to the Einstein scalar-field system on $(0, t_0] \times M$ such that (8-12) is satisfied for any $t \in (0, t_0]$. The core estimate (8-2) follows since Corollaries 7.3 and 7.6 now hold on $(0, t_0]$.

From (8-2), the asymptotic behaviour in (8-3a)–(8-3e) and (8-7) is established as in [Rodnianski and Speck 2018b, Theorem 15.1], which we briefly outline: First, we note that (8-3a) follows directly

from (8-2). For the remaining estimates, the arguments are similar, so consider for example $\partial_t\phi$: By the rescaled wave equation (2-32a) and (8-3a), we have that

$$\|\partial_t\Psi\|_{C^l_\gamma(\Sigma_t)} \lesssim \begin{cases} \varepsilon a^{1-c\varepsilon^{1/8}}, & l \leq 14, \\ \varepsilon a^{-1-c\varepsilon^{1/8}}, & l = 15. \end{cases}$$

Hence, for an arbitrary decreasing sequence $(t_m)_{m\in\mathbb{N}}$, on $(0, t_0]$ that converges to zero, we have

$$\|\Psi(t_{m_1},\cdot) - \Psi(t_{m_2},\cdot)\|_{C^l_\gamma(M)} \lesssim \begin{cases} \varepsilon a(t_{m_1})^{4-c\varepsilon^{1/8}}, & l \leq 14, \\ \varepsilon a(t_{m_1})^{2-c\varepsilon^{1/8}}, & l = 15, \end{cases}$$

for any $m_1, m_2 \in \mathbb{N}$, $m_1 < m_2$ by (2-6). This shows that $\Psi(t_{m_1},\cdot)$ is a Cauchy sequence in $C^{15}_\gamma(M)$ and hence there exists a limit function $\Psi_{\text{Bang}} \in C^{15}_\gamma(M)$ that satisfies

$$\|\Psi(t,\cdot) - \Psi_{\text{Bang}}\|_{C^l_\gamma(M)} \lesssim \begin{cases} \varepsilon a(t)^{4-c\varepsilon^{1/8}}, & l \leq 14, \\ \varepsilon a(t)^{2-c\varepsilon^{1/8}}, & l = 15, \end{cases}$$

for any $t \in (0, t_0]$. Since $\Psi = a^3 n^{-1}\partial_t\phi - C$ holds by definition, (8-3c) now follows by examining the Taylor expansion of $n^{-1} - 1$ at 0 using (8-3a).

The identity (8-4a) follows directly from the CMC condition (2-10), the asymptotic behaviour (8-3e) of $a^3 k$ and the Friedman equation (2-3), while (8-4b) follows from the asymptotic limit of the Hamiltonian constraint (2-16a), with (2-3), (8-3a), (8-3c) and (8-3e), as well as (8-2) for lower-order terms. The asymptotics in (8-7) follow exactly as in [Rodnianski and Speck 2018b, Theorem 15.1], and (8-5a)–(8-6) are a direct result of the initial data assumptions and applying the respective asymptotic estimates to $t = t_0$.

For the first estimate in (8-8b), we apply the momentum constraint (2-29d) to get

$$|\nabla^J B|_G = a^{-4}|\nabla^J \boldsymbol{B}|_G = a^{-2}|\nabla^J \operatorname{curl}_G \Sigma|_G \lesssim a^{-2}|\nabla^{J+1}\Sigma|_G,$$

and consequently, with Lemma 7.4, as well as (4-4g) and (8-2),

$$\begin{aligned} \|B\|_{C^{15}_\gamma(\Sigma_t)} &\lesssim a^{-c\sqrt{\varepsilon}}\|B\|_{C^{15}_G(\Sigma_t)} + \varepsilon a^{-2-c\sqrt{\varepsilon}} \cdot P_{15}(\|G-\gamma\|_{C^{15}_\gamma(\Sigma_t)}) \\ &\lesssim a^{-2-c\sqrt{\varepsilon}}\|\Sigma\|_{C^{16}_G(\Sigma_t)} + \varepsilon a^{-2-c\sqrt{\varepsilon}} \cdot P_{15}(\|G-\gamma\|_{C^{15}_\gamma(\Sigma_t)}) \\ &\lesssim \varepsilon a^{-2-c\varepsilon^{1/8}}. \end{aligned}$$

The remaining estimates in (8-8a) and (8-8b) are contained in (8-2). The results (8-9) and (8-10) follow as in the proofs of (15.6) and (15.7) in [Rodnianski and Speck 2018b, Theorem 15.1] from the asymptotic behaviour of the solution variables in (8-3a)–(8-3e) and (8-7). We briefly sketch the proof of (8-10): Consider a geodesic $\boldsymbol{\alpha}$ affinely parametrized by $\mathcal{A}$ as in the statement. The geodesic equations then lead to the following estimate for some suitable $\mathcal{K} > 0$:

$$|\mathcal{A}''| \leq \frac{\dot{a}}{a}|\mathcal{A}'| + \mathcal{K}\Big[\frac{\dot{a}}{a}|N| + n^{-1}|\partial_t N| + n^{-1}|\nabla N|_g + n|\hat{k}|_g\Big]|\mathcal{A}'|.$$

The leading term is hereby arises from the mean curvature condition. Arguing as with the elliptic estimates in Section 5, one can show that $|\partial_t N| \lesssim \varepsilon a^{-1-c\varepsilon^{1/8}}$. Thus, along with the other pointwise bounds on n,

g and $\hat{k}$, one obtains

$$|\mathcal{A}''| \leq \frac{\dot{a}}{a}(1+c\varepsilon)|\mathcal{A}'|$$

and consequently

$$|\mathcal{A}'(t)| \leq |\mathcal{A}'(t_0)|a(t)^{-1-c\varepsilon}$$

by the Gronwall lemma. Equation (8-10) follows by integrating.

Turning to the blow-up behaviour of geometric invariants, observe (8-11a) is a direct consequence of (8-3e). Regarding (8-11c), we first compute using (2-19) and standard algebraic manipulations that

$$a^{12}W_{\alpha\beta\gamma\delta}W^{\alpha\beta\gamma\delta} = a^{12}(8|E|_g^2 + 8|B|_g^2) = 8|\boldsymbol{E}|_G^2 + 8|\boldsymbol{B}|_G^2.$$

By the rescaled constraint equation (2-29c), we have

$$\boldsymbol{E}_{ij} = -\dot{a}a^2\Sigma_{ij} + (\Sigma \odot \Sigma)_{ij} - \left[\tfrac{8\pi}{3}\Psi^2 + \tfrac{16\pi}{3}C\Psi\right]G_{ij} + \mathcal{O}(\varepsilon a^{4-c\varepsilon^{1/8}})$$

for $t \downarrow 0$. Further, by expanding (2-3) around $a = 0$, we have $\dot{a}a^2 = \sqrt{\frac{4\pi}{3}}C + \mathcal{O}(a^2)$. Since $\Sigma^\sharp$ and Ψ converge to footprint states $\widehat{K}_{\mathrm{Bang}} = K_{\mathrm{Bang}} + \sqrt{\frac{4\pi}{3}}C\mathbb{I}$ and Ψ_{Bang} in $C^{15}_\gamma(M)$ respectively, this shows that $8|\boldsymbol{E}|_G^2$ converges to some $W_{\mathrm{Bang}} \in C^{15}_\gamma(M)$ that can be expressed as a fourth-order polynomial in $\widehat{K}_{\mathrm{Bang}}$ and Ψ_{Bang} and satisfies

$$\left\| |\boldsymbol{E}|_G^2 - \tfrac{1}{8}W_{\mathrm{Bang}} \right\|_{C^0(M)} \lesssim \varepsilon a^{2-c\varepsilon^{1/8}},$$

as well as (8-11d). Due to (8-8b), the $|\boldsymbol{B}|_G^2$-term in the Weyl curvature scalar is negligible in comparison, and thus (8-11c) immediately follows.

Furthermore, one has

$$P_{\alpha\beta\gamma\delta}P^{\alpha\beta\gamma\delta} = 2\,\mathrm{Ric}[\bar{g}]_{\alpha\beta}\,\mathrm{Ric}[\bar{g}]^{\alpha\beta} - \tfrac{2}{9}R[\bar{g}]^2,$$

and (8-11b) is a direct consequence of (8-11e)–(8-11f), which follow once more with (8-3c) and (8-3a) as well as (8-2) for error terms. Finally, (8-11g) is obtained from (8-11b)–(8-11c). □

9. Future stability

The goal of this section is to show the following theorem:

Theorem 9.1 (future stability of Milne spacetime). *Let the rescaled initial data* $(\boldsymbol{g}, \boldsymbol{k}, \nabla\phi, \phi')$ *on* M *be sufficiently close to* $\left(\gamma, \frac{1}{3}\gamma, 0, 0\right)$ *in* $H^5 \times H^4 \times H^4 \times H^4$ *on some initial hypersurface* $\Sigma_{\tau=\tau_0}$ (*see Definition 9.4 and Assumption 9.7*). *Then, its maximal globally hyperbolic development* $(\overline{M}, \bar{g}, \phi)$ *within the Einstein scalar-field system in CMCSH gauge is foliated by the CMC Cauchy hypersurfaces* $(\Sigma_\tau)_{\tau\in[\tau_0,0)}$, *is future* (*causally*) *complete and exhibits the following asymptotic behaviour*:

$$(\boldsymbol{g}, \boldsymbol{k}, \phi', \nabla\phi)(\tau) \to \left(\gamma, \tfrac{1}{3}\gamma, 0, 0\right) \quad \text{as } \tau \uparrow 0.$$

Since the control of geometric perturbations uses the same arguments as in [Andersson and Fajman 2020], the focus in this section will lie on dealing with the scalar field. The key idea is controlling decay of the scalar field using an indefinite corrective term on top of the canonical energy (see Definition 9.6).

9.1. *Preliminaries.*

9.1.1. *Notation, gauge and spatial reference geometry.* Within this section, we will decompose the Lorentzian metric as

$$\bar{g} = -n^2\,dt^2 + g_{ab}(dx^a + X^a)(dx^b + X^b\,dt). \tag{9-1a}$$

We impose CMCSH gauge (see [Andersson and Moncrief 2004]) via

$$t = \tau, \quad g^{ij}(\Gamma^a_{ij} - \widehat{\Gamma}^a_{ij}) = 0, \tag{9-1b}$$

where $\widehat{\Gamma}$ refers to the Christoffel symbols with regard to the spatial reference metric γ.

We extend the notation from the big bang stability analysis regarding foliations, derivatives, indices and schematic term notation to this setting (see Section 2.1). In particular, Σ_T and Σ_τ will refer to spatial hypersurfaces along which the logarithmic time T (see (9-2c)) and the mean curvature τ are constant (see (9-2c) on why these are interchangeable), and we will write for example $\Sigma_{T=0}$ when inserting a specific value to avoid potential ambiguity. We use similar notation for scalar functions and tensors that depend on T or, respectively, τ.

For the extent of the future stability analysis, we have to introduce an additional condition for the spatial geometry beyond Definition 2.1:

Definition 9.2 (spectral condition for the Laplacian of the spatial reference manifold). Let $\mu_0(\gamma)$ to be the smallest positive eigenvalue of the Laplace operator $-\Delta_\gamma = -(\gamma^{-1})^{ab}\hat{\nabla}_a\hat{\nabla}_b$ acting on scalar functions, where (M, γ) is as in Definition 2.1. (M, γ) additionally is assumed to satisfy

$$\mu_0(\gamma) > \tfrac{1}{9}.$$

Remark 9.3 (manifolds that satisfy Definition 9.2). The available literature on spectra of $-\Delta_\gamma$ usually focuses on hyperbolic manifolds with sectional curvature $\kappa = -1$. Thus, one needs to check that μ_0 is strictly greater than 1 to verify the analogue of Definition 9.2 after rescaling.

Numerical works, e.g., [Cornish and Spergel 1999; Inoue 2001], provide evidence for over 250 compact hyperbolic 3-manifolds to satisfy this spectral bound, many of which are closed. In particular, both [Cornish and Spergel 1999][14] and [Inoue 2001] consider the smallest closed orientable hyperbolic 3-manifold, the Weeks space m003$(-3, 1)$, and compute that it falls under Definition 9.2 with $\mu_0 \approx 27.8$ in [Cornish and Spergel 1999, Table IV] and $26 \lessapprox \mu_0 \lessapprox 27.8$ in [Inoue 2001, Table 2]. Moreover, as demonstrated in [Inoue 2001, Figure 6], many manifolds with small enough diameter d satisfy this condition. In fact, the analytical bound

$$\mu_0 \geq \max\Big\{\frac{\pi^2}{2d^2} - \frac{1}{2}, \sqrt{\frac{\pi^4}{d^4} + \frac{1}{4}} - \frac{3}{2}, \frac{\pi^2}{d^2}e^{-d}\Big\}$$

(see [Cheng and Zhou 1995, Theorem 1.1–1.2] with $L = 2$) implies that $\mu_0 > 10$ holds for Weeks space, which has diameter $d \approx 0,843$ (see [Cornish and Spergel 1999, Table V]). Furthermore, [Inoue 2001] finds no closed hyperbolic manifolds that violate this bound. More recently, the Selberg trace formula

[14] These results have to be interpreted cautiously since the numerical method cannot detect eigenvalues below 1.

has been used in [Lin and Lipnowski 2022; 2024; Bonifacio et al. 2025] to compute candidates for eigenvalues of $-\Delta_\gamma$ and related operators, based on an optimization approach originating in [Booker and Strömbergsson 2007]. In particular, the calculations visualized in [Bonifacio et al. 2025, Figure 3] demonstrate that one must have $\mu_0 \geq 27,6$ on the Weeks manifold.

We also note that it is conjectured that one at least has $\mu_0 \geq 1$ for any arithmetic hyperbolic 3-manifold (see [Bergeron 2003, Conjecture 2.3]). In fact, this is tied to the Ramanujan conjecture for automorphic forms. Finally, one can construct compact manifolds with boundary and with constant sectional curvature -1 where μ_0 becomes arbitrarily small; see [Callahan 1994, Corollary 4.4].

9.1.2. *Rescaled variables and Einstein equations.* We will use the standard rescaling of the solution variables by τ:

Definition 9.4 (rescaled variables for future stability).

$$\boldsymbol{g}_{ij} = \tau^2 g_{ij}, \quad (\boldsymbol{g}^{-1})^{ij} = \tau^{-2} g^{ij}, \quad \boldsymbol{\Sigma}_{ij} = \tau \hat{k}_{ij}, \tag{9-2a}$$

$$\boldsymbol{n} = \tau^2 n, \quad \hat{\boldsymbol{n}} = \tfrac{1}{3}\boldsymbol{n} - 1, \quad \boldsymbol{X}^a = \tau X^a. \tag{9-2b}$$

Furthermore, we introduce the logarithmic time

$$T = -\log\Big(\frac{\tau}{\tau_0}\Big) \quad \Longleftrightarrow \quad \tau = \tau_0 e^{-T}, \tag{9-2c}$$

which satisfies $\partial_T = -\tau\partial_\tau$. Toward the future, τ increases from τ_0 to 0, and thus T increases from 0 to ∞. We additionally introduce

$$\tilde{\partial}_0 = \partial_T + \mathcal{L}_{\boldsymbol{X}} = -\tau(\partial_\tau - \mathcal{L}_X), \tag{9-2d}$$

$$\phi' = \boldsymbol{n}^{-1}\tilde{\partial}_0\phi = n^{-1}(-\tau)^{-1}(\partial_\tau - \mathcal{L}_X)\phi. \tag{9-2e}$$

Moreover, for any scalar function ζ, we denote by $\bar{\zeta}$ the mean integral with respect to $(\Sigma_T, \boldsymbol{g}_T)$.

For symmetric $(0,2)$-tensors h, we define the perturbed Lichnerowicz Laplacian

$$\mathcal{L}_{\boldsymbol{g},\gamma} h_{ab} = -\frac{1}{\mu_{\boldsymbol{g}}}\hat{\nabla}_k((\boldsymbol{g}^{-1})^{kl}\mu_{\boldsymbol{g}}\hat{\nabla}_l h_{ab}) - 2\,\mathrm{Riem}[\gamma]_{akbl}(\boldsymbol{g}^{-1})^{kk'}(\boldsymbol{g}^{-1})^{ll'}h_{k'l'}. \tag{9-3}$$

This operator satisfies

$$(\mathrm{Ric}[\boldsymbol{g}] - \mathrm{Ric}[\gamma])_{ij} = \tfrac{1}{2}\mathcal{L}_{\boldsymbol{g},\gamma}(\boldsymbol{g}-\gamma)_{ij} + J_{ij}, \quad \|J\|_{H^{l-1}} \lesssim \|\boldsymbol{g}-\gamma\|_{H^l}; \tag{9-4}$$

see [Andersson and Moncrief 2003, Proof of Theorem 3.1]. Under our conditions for the reference geometry, [Kröncke 2015] implies that the smallest positive eigenvalue of $\mathcal{L}_{\gamma,\gamma}$, denoted by λ_0, satisfies $\lambda_0 \geq \frac{1}{9}$, and that $\mathcal{L}_{\gamma,\gamma}$ has trivial kernel. The spectral condition in Definition 9.2 is not necessary for this to hold true.

We now collect the (3+1)-decomposition of the Einstein scalar-field equations in CMCSH gauge with the help of [Andersson and Fajman 2020, (2.13)–(2.18)]:

Lemma 9.5 (rescaled CMCSH equations). *The rescaled CMCSH Einstein scalar-field equations take the following form*: *The constraint equations*

$$R[\boldsymbol{g}] - |\boldsymbol{\Sigma}|^2_{\boldsymbol{g}} - \tfrac{2}{3} = 8\pi[|\phi'|^2 + |\nabla\phi|^2_{\boldsymbol{g}}], \quad \operatorname{div}_{\boldsymbol{g}} \boldsymbol{\Sigma}_b = 8\pi\tau^3\phi'\nabla_b\phi, \tag{9-5a}$$

the elliptic lapse and shift equations

$$\big(\Delta_{\boldsymbol{g}} - \tfrac{1}{3}\big)\boldsymbol{n} = \boldsymbol{n}(|\boldsymbol{\Sigma}|^2_{\boldsymbol{g}} + 4\pi[|\phi'|^2 + |\nabla\phi|^2_{\boldsymbol{g}}]) - 1, \tag{9-5b}$$

$$\begin{aligned}\Delta_{\boldsymbol{g}} \boldsymbol{X}^a + (\boldsymbol{g}^{-1})^{ab}\operatorname{Ric}[\boldsymbol{g}]_{bm}\boldsymbol{X}^m &= 2(\boldsymbol{g}^{-1})^{am}(\boldsymbol{g}^{-1})^{bn}\nabla_b\boldsymbol{n}\cdot\boldsymbol{\Sigma}_{mn} - (\boldsymbol{g}^{-1})^{ab}\nabla_b\hat{\boldsymbol{n}} + 8\pi\boldsymbol{n}\tau^3\phi'\nabla_b\phi \\ &\quad - 2(\boldsymbol{g}^{-1})^{bk}((\boldsymbol{g}^{-1})^{cl}\boldsymbol{n}\cdot\boldsymbol{\Sigma}_{bc} - \nabla_b\boldsymbol{X}^l)(\Gamma^a_{kl} - \widehat{\Gamma}^a_{kl}),\end{aligned} \tag{9-5c}$$

the geometric evolution equations

$$\tilde{\partial}_0\boldsymbol{g}_{ab} = 2\boldsymbol{n}\boldsymbol{\Sigma}_{ab} + 2\hat{\boldsymbol{n}}\boldsymbol{g}_{ab}, \tag{9-5d}$$

$$\tilde{\partial}_0(\boldsymbol{g}^{-1})^{ab} = -2\boldsymbol{n}(\boldsymbol{g}^{-1})^{ac}(\boldsymbol{g}^{-1})^{bd}\boldsymbol{\Sigma}_{cd} - 2\hat{\boldsymbol{n}}(\boldsymbol{g}^{-1})^{ab}, \tag{9-5e}$$

$$\begin{aligned}\tilde{\partial}_0\boldsymbol{\Sigma}_{ab} &= -2\boldsymbol{\Sigma}_{ab} - \boldsymbol{n}\big(\operatorname{Ric}[\boldsymbol{g}]_{ab} + \tfrac{2}{9}\boldsymbol{g}_{ab}\big) + \nabla_a\nabla_b\boldsymbol{n} \\ &\quad + 2\boldsymbol{n}\cdot(\boldsymbol{g}^{-1})^{mn}\boldsymbol{\Sigma}_{am}\boldsymbol{\Sigma}_{bn} - \tfrac{1}{3}\hat{\boldsymbol{n}}\boldsymbol{g}_{ab} - \hat{\boldsymbol{n}}\Sigma_{ab} - 8\pi\boldsymbol{n}\nabla_a\phi\nabla_b\phi\end{aligned} \tag{9-5f}$$

and the wave equation

$$\tilde{\partial}_0\phi' = \langle\nabla\boldsymbol{n}, \nabla\phi\rangle_{\boldsymbol{g}} + \boldsymbol{n}\Delta_{\boldsymbol{g}}\phi + (1-\boldsymbol{n})\phi'. \tag{9-5g}$$

9.1.3. *Energies and data assumptions.* The proof will rely on the following corrected energy quantities:

Definition 9.6 (energies for future stability).

$$\mathbb{E}^{(l)}_{\mathrm{SF}} = (-1)^l\int_M [\phi'\Delta^l_{\boldsymbol{g}}\phi' - \phi\Delta^{l+1}_{\boldsymbol{g}}\phi]\operatorname{vol}_{\boldsymbol{g}}, \qquad \mathcal{C}^{(l)}_{\mathrm{SF}} = (-1)^l\int_M (\phi - \bar{\phi})\Delta^l_{\boldsymbol{g}}\phi'\operatorname{vol}_{\boldsymbol{g}}, \tag{9-6a}$$

$$E_{\mathrm{SF}} = \sum_{m=0}^{4}\big(\mathbb{E}^{(m)}_{\mathrm{SF}} + \tfrac{2}{3}\mathcal{C}^{(m)}_{\mathrm{SF}}\big), \tag{9-6b}$$

$$\begin{aligned}E_{\mathrm{geom}} = \sum_{m=1}^{5}\Big(&\frac{9}{2}\int_M\langle\boldsymbol{g}-\gamma, \mathcal{L}^m_{\boldsymbol{g},\gamma}(\boldsymbol{g}-\gamma)\rangle_{\boldsymbol{g}}\operatorname{vol}_{\boldsymbol{g}} \\ &+\frac{1}{2}\int_M\langle 6\boldsymbol{\Sigma}, \mathcal{L}^{m-1}_{\boldsymbol{g},\gamma}(6\boldsymbol{\Sigma})\rangle_{\boldsymbol{g}}\operatorname{vol}_{\boldsymbol{g}} + c_E\int_M\langle 6\boldsymbol{\Sigma}, \mathcal{L}^{m-1}_{\boldsymbol{g},\gamma}(\boldsymbol{g}-\gamma)\rangle_{\boldsymbol{g}}\operatorname{vol}_{\boldsymbol{g}}\Big).\end{aligned} \tag{9-6c}$$

The constant c_E is given by

$$c_E = \begin{cases}1, & \lambda_0 > \frac{1}{9}, \\ 9(\lambda_0 - \delta'), & \lambda_0 = \frac{1}{9},\end{cases} \tag{9-7}$$

where $\delta' > 0$ is chosen to be small enough within the argument.

The Sobolev norms $H^l_{\boldsymbol{g}}$ and $C^l_{\boldsymbol{g}}$ are defined analogously to Definitions 3.3 and 3.4, with similar conventions on suppressing time-dependence in notation whereever possible. Since norms with respect to $\boldsymbol{g}$ and γ are equivalent under the bootstrap assumption (and consequently throughout the entire argument), we will simply denote the norms by H^l and C^l throughout unless the specific metric is crucial.

Assumption 9.7 (initial data assumption). *The initial data on the spatial hypersurface $\Sigma_{T=0}$ is assumed to be small in the following sense*:

$$\|\boldsymbol{g}-\gamma\|_{C^3}+\|\boldsymbol{\Sigma}\|_{C^2}+\|\hat{\boldsymbol{n}}\|_{C^4}+\|\boldsymbol{X}\|_{C^4}+\|\phi'\|_{C^2}+\|\nabla\phi\|_{C^2}\\ +\|\boldsymbol{g}-\gamma\|_{H^5}+\|\boldsymbol{\Sigma}\|_{H^4}+\|\hat{\boldsymbol{n}}\|_{H^6}+\|\boldsymbol{X}\|_{H^6}+\|\phi'\|_{H^4}+\|\nabla\phi\|_{H^4}\leq\delta^2. \quad (9\text{-}8)$$

Remark 9.8 (local well-posedness toward the future). Under the above initial data assumption, local well-posedness is satisfied by analogizing the arguments for local well-posedness in the vacuum setting (see [Andersson and Moncrief 2003, Theorem 3.1]) with the matter coupling added. Since this only consists of adding another wave equation to the hyperbolic system, the argument is structurally unchanged given appropriate smallness assumptions on ϕ' and $\nabla\phi$ (where ϕ itself does not enter into the Einstein system). As before, we can without loss of generality assume that the initial is sufficiently regular to ensure that E_{geom}, $\mathcal{E}^{(l)}_{\mathrm{SF}}$ and $\mathcal{C}^{(l)}_{\mathrm{SF}}$ initially are continuously differentiable (in time) for any $l\leq 4$.

Assumption 9.9 (bootstrap assumption). *On the bootstrap interval $T\in[0,T_{\mathrm{Boot}})$, we assume one has*

$$\|\boldsymbol{g}-\gamma\|_{C^3}+\|\boldsymbol{\Sigma}\|_{C^2}+\|\hat{\boldsymbol{n}}\|_{C^4}+\|\boldsymbol{X}\|_{C^4}+\|\phi'\|_{C^2}+\|\nabla\phi\|_{C^2}\\ +\|\boldsymbol{g}-\gamma\|_{H^5}+\|\boldsymbol{\Sigma}\|_{H^4}+\|\hat{\boldsymbol{n}}\|_{H^6}+\|\boldsymbol{X}\|_{H^6}+\|\phi'\|_{H^4}+\|\nabla\phi\|_{H^4}\leq\delta e^{-T/2}. \quad (9\text{-}9)$$

We only choose not to use "$\lesssim$"-notation in the above assumptions for notational convenience in some technical computations. As before, δ can be chosen to be sufficiently small for the following estimates to hold and for the decay estimates we derive from the bootstrap assumptions to be strict improvements. Moreover, note that (9-9) is satisfied since all of the norms are continuous in time (see Remark 9.8).

Before moving on to the energy estimates, we quickly collect the following immediate consequence of the bootstrap assumptions:

Lemma 9.10 (Sobolev estimate for the curvature). *The following estimate holds for any $l\in\mathbb{N}_0$*:

$$\left\|\mathrm{Ric}[\boldsymbol{g}]+\tfrac{2}{9}\boldsymbol{g}\right\|_{H^l}\lesssim\|\boldsymbol{g}-\gamma\|_{H^{l+2}}+\|\boldsymbol{g}-\gamma\|^2_{H^{l+1}}. \quad (9\text{-}10a)$$

Under the bootstrap assumptions, this implies

$$\left\|\mathrm{Ric}[\boldsymbol{g}]+\tfrac{2}{9}\boldsymbol{g}\right\|_{C^1}+\left\|\mathrm{Ric}[\boldsymbol{g}]+\tfrac{2}{9}\boldsymbol{g}\right\|_{H^3}\lesssim\delta e^{-T/2}. \quad (9\text{-}10b)$$

Proof. By (9-4), one has

$$\left\|\mathrm{Ric}[\boldsymbol{g}]+\tfrac{2}{9}\boldsymbol{g}\right\|_{H^l}\leq\tfrac{1}{2}\|\mathcal{L}_{\boldsymbol{g},\gamma}(\boldsymbol{g}-\gamma)\|_{H^l}+K\|\boldsymbol{g}-\gamma\|^2_{H^{l+1}}$$

for some suitably large $K>0$, along with the fact that $\mathcal{L}_{\boldsymbol{g},\gamma}$ is elliptic. This implies the first inequality, while the latter follows from directly from the bootstrap assumption (9-9) and by applying the standard Sobolev embedding. □

9.2. ***Elliptic estimates.*** We briefly collect the elliptic estimates for lapse and shift:

Lemma 9.11 (elliptic estimates for lapse and shift). *Let* $l \in \{3, 4, 5, 6\}$. *Then, one has* $\boldsymbol{n} \in (0, 3)$ (*thus* $\hat{\boldsymbol{n}} \in (-1, 0)$) *and the following estimates hold*:

$$\|\hat{\boldsymbol{n}}\|_{H^l} \lesssim \delta e^{-T/2}\|\boldsymbol{\Sigma}\|_{H^{l-2}} + \delta^2 e^{-T}\|\boldsymbol{g}-\gamma\|_{H^{l-2}} + \delta e^{-T/2}[\|\phi'\|_{H^{l-2}} + \|\nabla\phi\|_{H^{l-2}}], \tag{9-11a}$$

$$\|\boldsymbol{X}\|_{H^l} \lesssim \delta e^{-T/2}\|\boldsymbol{\Sigma}\|_{H^{l-2}} + \delta e^{-T/2}\|\boldsymbol{g}-\gamma\|_{H^{l-1}} + \delta e^{-T/2}[\|\phi'\|_{H^{l-2}} + \|\nabla\phi\|_{H^{l-2}}]. \tag{9-11b}$$

Proof. The pointwise bounds on $\boldsymbol{n}$ follow via (9-5b) and the maximum principle as in Lemma 4.1. For the remaining estimates, applying elliptic regularity theory to (9-5b) and (9-5c) implies

$$\begin{aligned}\|\hat{\boldsymbol{n}}\|_{H^l} \lesssim \|\boldsymbol{\Sigma}\|_{C^{\lfloor (l-2)/2\rfloor}}\|\boldsymbol{\Sigma}\|_{H^{l-2}} + \|\nabla\phi\|_{C^2}^2\|\boldsymbol{g}-\gamma\|_{H^{l-2}} \\ + [\|\nabla\phi\|_{C^2}(1+\|\boldsymbol{g}-\gamma\|_{C^2}) + \|\phi'\|_{C^2}][\|\phi'\|_{H^{l-2}} + \|\nabla\phi\|_{H^{l-2}}],\\ \|\boldsymbol{X}\|_{H^l} \lesssim \|\boldsymbol{\Sigma}\|_{C^{\lfloor (l-2)/2\rfloor}}\|\boldsymbol{\Sigma}\|_{H^{l-2}} + \|\boldsymbol{g}-\gamma\|_{H^{l-1}}^2 + \|\nabla\phi\|_{C^1}\|\boldsymbol{g}-\gamma\|_{H^{l-3}} \\ + [\|\nabla\phi\|_{C^2}^2(1+\|\boldsymbol{g}-\gamma\|_{C^2}) + \|\phi'\|_{C^2}][1+\|\hat{\boldsymbol{n}}\|_{C^2}][\|\phi'\|_{H^{l-2}} + \|\nabla\phi\|_{H^{l-2}}].\end{aligned}$$

The statement then follows by inserting (9-9). □

9.3. *Scalar field energy estimates.*

9.3.1. *Near-coercivity of* E_{SF}. We will be able to prove a decay estimate via a Gronwall argument only for the corrected energy E_{SF}. Hence, we first need to verify that this energy controls the solution norms, for which we first show that it controls the "canonical" scalar field energies:

Lemma 9.12 (positivity of corrected scalar field energies). *Let*

$$Q = \frac{\sqrt{1+9q}-1}{\sqrt{1+9q}}, \quad \text{with } q = \tfrac{1}{2}\big(\mu_0(\gamma) - \tfrac{1}{9}\big).$$

Then, for any $l \in \{0, 1, 2, 3, 4\}$ *and* $\delta > 0$ *small enough, one has*

$$Q\mathbb{E}_{\mathrm{SF}}^{(l)} \le \mathbb{E}_{\mathrm{SF}}^{(l)} + \tfrac{2}{3}\mathcal{C}_{\mathrm{SF}}^{(l)}, \quad \text{hence} \quad Q\sum_{m=0}^{4}\mathbb{E}_{\mathrm{SF}}^{(l)} \le E_{\mathrm{SF}}. \tag{9-12}$$

Proof. We denote the smallest positive eigenvalue of $-\Delta_{\boldsymbol{g}}$ acting on scalar functions on Σ_T by $\mu_0(\boldsymbol{g}_T)$. By the bootstrap assumption (9-9) and since μ_0 depends continuously on the metric, we obtain the following for small enough $\delta > 0$:

$$\mu_0(\boldsymbol{g}_T) \ge \mu_0(\gamma) - \tfrac{1}{2}\big(\mu_0(\gamma) - \tfrac{1}{9}\big) \ge \tfrac{1}{9} + q.$$

By the Poincaré inequality applied on $(\Sigma_T, \boldsymbol{g}_T)$ (see [Choquet-Bruhat and Moncrief 2001, p. 1037]), the above spectral bound implies the following for any $\zeta \in H^1(\Sigma_T)$:

$$\|\zeta - \bar{\zeta}\|_{L^2_{\boldsymbol{g}}(\Sigma_T)}^2 \le \mu_0(\boldsymbol{g}_T)^{-1}\|\nabla\zeta\|_{L^2_{\boldsymbol{g}}(\Sigma_T)}^2 \le \big(\tfrac{1}{9}+q\big)^{-1}\|\nabla\zeta\|_{L^2_{\boldsymbol{g}}(\Sigma_T)}^2. \tag{9-13}$$

For $l = 0$, this means

$$\begin{aligned}\mathbb{E}_{\mathrm{SF}}^{(0)} + \tfrac{2}{3}\mathcal{C}_{\mathrm{SF}}^{(0)} &\ge \|\phi'\|_{L^2_{\boldsymbol{g}}}^2 + \|\nabla\phi\|_{L^2_{\boldsymbol{g}}}^2 - \tfrac{2}{3}\|\phi-\bar{\phi}\|_{L^2_{\boldsymbol{g}}}\|\phi'\|_{L^2_{\boldsymbol{g}}} \\ &\ge \|\phi'\|_{L^2_{\boldsymbol{g}}}^2 + \|\nabla\phi\|_{L^2_{\boldsymbol{g}}}^2 - 2(1+9q)^{-1/2}\|\nabla\phi\|_{L^2_{\boldsymbol{g}}}\|\phi'\|_{L^2_{\boldsymbol{g}}} \ge \frac{\sqrt{1+9q}-1}{\sqrt{1+9q}}\mathbb{E}_{\mathrm{SF}}^{(0)}.\end{aligned}$$

For $l = 1$, notice that we can rewrite $\mathcal{C}^{(1)}_{\mathrm{SF}}$ as

$$\mathcal{C}^{(1)}_{\mathrm{SF}} = \int_M \langle \nabla\phi, \nabla\phi' \rangle_{\boldsymbol{g}} \operatorname{vol}_{\boldsymbol{g}} = \int_M \langle \nabla\phi, \nabla(\phi' - \overline{\phi'}) \rangle_{\boldsymbol{g}} \operatorname{vol}_{\boldsymbol{g}} = -\int_M (\phi' - \overline{\phi'}) \Delta_{\boldsymbol{g}} \phi \operatorname{vol}_{\boldsymbol{g}}.$$

Hence, applying (9-13) to $\zeta = \phi'$ yields

$$\mathbb{E}^{(1)}_{\mathrm{SF}} + \tfrac{2}{3}\mathcal{C}^{(1)}_{\mathrm{SF}} \geq \mathbb{E}^{(1)}_{\mathrm{SF}} - 2(1+9q)^{-1/2} \|\nabla\phi'\|_{L^2_{\boldsymbol{g}}} \|\Delta_{\boldsymbol{g}}\phi\|_{L^2_{\boldsymbol{g}}} \geq \frac{\sqrt{1+9q}-1}{\sqrt{1+9q}} \mathbb{E}^{(1)}_{\mathrm{SF}}.$$

For $l = 2, 3, 4$, notice $\overline{\Delta_{\boldsymbol{g}}\phi} = \overline{\Delta_{\boldsymbol{g}}\phi'} = \overline{\Delta^2_{\boldsymbol{g}}\phi} = 0$ holds due to the divergence theorem; hence the argument proceeds as in $l = 0, 1$. □

Lemma 9.13 (near-coercivity of corrected scalar field energy). *For any scalar function ζ and $k \in \{1, 2\}$, one has the following under the bootstrap assumptions*:

$$\begin{aligned}
\int_M |\nabla^2\zeta|^2_{\boldsymbol{g}} \operatorname{vol}_{\boldsymbol{g}} &\lesssim \int_M |\Delta_{\boldsymbol{g}}\zeta|^2_{\boldsymbol{g}} + |\nabla\zeta|^2_{\boldsymbol{g}} \operatorname{vol}_{\boldsymbol{g}}, \\
\|\zeta\|^2_{\dot{H}^{2k}} &\lesssim \|\Delta^k_{\boldsymbol{g}}\zeta\|^2_{L^2} + (\|\zeta\|^2_{\dot{H}^{2k-1}} + \|\zeta\|^2_{\dot{H}^{2k-2}}) + \|\nabla\zeta\|^2_{C^1} \big\|\mathrm{Ric}[\boldsymbol{g}] + \tfrac{2}{9}\boldsymbol{g}\big\|^2_{H^{2k-2}}, \\
\|\nabla\zeta\|^2_{\dot{H}^{2k}} &\lesssim \|\nabla\Delta^k_{\boldsymbol{g}}\zeta\|^2_{L^2} + (\|\nabla\zeta\|^2_{\dot{H}^{2k-1}} + \|\nabla\zeta\|^2_{\dot{H}^{2k-2}}) + \|\nabla\zeta\|^2_{C^2} \big\|\mathrm{Ric}[\boldsymbol{g}] + \tfrac{2}{9}\boldsymbol{g}\big\|^2_{H^{2k-2}}.
\end{aligned}$$

Consequently, the following estimate holds:

$$\|\phi'\|^2_{H^4} + \|\nabla\phi\|^2_{H^4} \lesssim E^{(4)}_{\mathrm{SF}} + (\|\phi'\|^2_{C^2} + \|\nabla\phi\|^2_{C^2}) \big\|\mathrm{Ric}[\boldsymbol{g}] + \tfrac{2}{9}\boldsymbol{g}\big\|^2_{H^2}. \tag{9-14}$$

Proof. The inequalities for ζ follow from the same arguments as Lemma 4.5, except that we have $\|\mathrm{Ric}[\boldsymbol{g}]\|_{C^1_{\boldsymbol{g}}} \lesssim 1 + \delta \lesssim 1$ by Lemma 9.10. The final estimate then follows by applying these estimates to $\zeta = \phi'$ and $\zeta = \phi$ and applying Lemma 9.12. □

9.3.2. *Preparations for energy estimates.* Before proving the energy estimate, we need to establish two technical lemmas: First, we collect a formula to differentiate integrals, and then some estimates needed to deal with the mean value of ϕ in the base level correction term.

Lemma 9.14 (differentiation of integrals, future stability version). *For any differentiable function ζ, one has*

$$\partial_T \int_M \zeta \operatorname{vol}_{\boldsymbol{g}} = \int_M (\tilde{\partial}_0\zeta + 3\hat{\boldsymbol{n}}\zeta) \operatorname{vol}_{\boldsymbol{g}}. \tag{9-15}$$

Proof. As in the proof of (4-12), we obtain

$$\begin{aligned}
\partial_T \int_M \zeta \operatorname{vol}_{\boldsymbol{g}} &= \int_M \partial_T\zeta + \frac{\partial_T \mu_{\boldsymbol{g}}}{\mu_{\boldsymbol{g}}}\zeta \operatorname{vol}_{\boldsymbol{g}} \\
&= \int_M \partial_T\zeta + 3\hat{\boldsymbol{n}}\zeta - \tfrac{1}{2}(\boldsymbol{g}^{-1})^{ab}\mathcal{L}_{\boldsymbol{X}}\boldsymbol{g}_{ab}\zeta \operatorname{vol}_{\boldsymbol{g}} \\
&= \int_M \partial_T\zeta + 3\hat{\boldsymbol{n}}\zeta - \operatorname{div}_{\boldsymbol{g}} \boldsymbol{X} \cdot \zeta \operatorname{vol}_{\boldsymbol{g}}.
\end{aligned}$$

The statement now follows by applying Stokes' theorem to the final term and rearranging. □

Lemma 9.15 (decay estimate for the integrated time derivative). *For any $T > 0$, we have*

$$\int_{\Sigma_T} \phi' \operatorname{vol}_{\boldsymbol{g}} = \Big(\int_{\Sigma_{T=0}} \phi' \operatorname{vol}_{\boldsymbol{g}} \Big) \cdot e^{-2T}. \tag{9-16}$$

Consequently, the bootstrap assumptions imply

$$\Big| \int_{\Sigma_T} \tilde{\partial}_0 \bar{\phi} \cdot \phi' \operatorname{vol}_{\boldsymbol{g}} \Big| \lesssim \delta^3 e^{-\frac{5}{2}T} \tag{9-17}$$

for $\delta > 0$ small enough.

Proof. Using that the integral of $\operatorname{div}_{\boldsymbol{g}}(\boldsymbol{n}\nabla\phi)$ vanishes, we compute

$$\partial_T \Big(\int_M \phi' \operatorname{vol}_{\boldsymbol{g}} \Big) = \int_M (\tilde{\partial}_0 \phi' + 3\hat{\boldsymbol{n}}\phi') \operatorname{vol}_{\boldsymbol{g}} = \int_M [(1-\boldsymbol{n})\phi' + (\boldsymbol{n}-3)\phi'] \operatorname{vol}_{\boldsymbol{g}} = -2 \Big(\int_M \phi' \operatorname{vol}_{\boldsymbol{g}} \Big).$$

Hence, (9-16) precisely describes the solution to this ODE ($f' = -2f$) with prescribed initial value at $T = 0$, and the initial data assumption (9-8) implies

$$\Big| \int_M \phi' \operatorname{vol}_{\boldsymbol{g}} \Big| \leq \|\phi'\|_{C^0(\Sigma_{T=0})} \operatorname{vol}_{\boldsymbol{g}}(\Sigma_{T=0}) e^{-2T} \lesssim \delta^2 e^{-2T}.$$

Furthermore, one has by (9-15) that

$$\partial_T \operatorname{vol}_{\boldsymbol{g}}(\Sigma_T) = \int_{\Sigma_T} 3\hat{\boldsymbol{n}} \operatorname{vol}_{\boldsymbol{g}}. \tag{9-18}$$

Consequently, one has

$$\tilde{\partial}_0 \bar{\phi} = \Big[-\frac{\partial_T \operatorname{vol}_{\boldsymbol{g}}(\Sigma_T)}{\operatorname{vol}_{\boldsymbol{g}}(\Sigma_T)} \cdot \bar{\phi} + \frac{1}{\operatorname{vol}_{\boldsymbol{g}}(\Sigma_T)} \int_M (\tilde{\partial}_0 \phi + 3\hat{\boldsymbol{n}}\phi) \operatorname{vol}_{\boldsymbol{g}} \Big] = \int_M (\boldsymbol{n}\phi' + 3\hat{\boldsymbol{n}}(\phi - \bar{\phi})) \operatorname{vol}_{\boldsymbol{g}}.$$

By applying $|\boldsymbol{n}| < 3$, the adapted Poincare inequality (9-13) and the bootstrap assumptions (9-9), this implies

$$|\tilde{\partial}_0 \bar{\phi}| \lesssim \|\phi'\|_{L^2} + \|\nabla\phi\|_{L^2} \|\hat{\boldsymbol{n}}\|_{L^2_{\boldsymbol{g}}} \lesssim \delta e^{-T/2}.$$

The bound (9-17) now follows by combining this with (9-16). □

9.3.3. *Energy estimates.* Now, we can collect the following estimates for the corrected scalar field energies:

Lemma 9.16 (base level estimate for the corrected scalar field energy). *Under the bootstrap assumptions, the following estimate holds for some $K > 0$:*

$$\partial_T E^{(0)}_{\mathrm{SF}} \leq -2E^{(0)}_{\mathrm{SF}} + K\delta e^{-T/2} \sqrt{E^{(0)}_{\mathrm{SF}}} \big(\sqrt{E^{(0)}_{\mathrm{SF}}} + \|\boldsymbol{\Sigma}\|_{L^2} + \|\boldsymbol{g} - \gamma\|_{L^2} \big) + K\delta^3 e^{-5T/2}. \tag{9-19}$$

Proof. We compute, using $[\tilde{\partial}_0, \nabla]\phi = 0$, $\tilde{\partial}_0 \phi = \boldsymbol{n}\phi'$ and the rescaled wave equation (9-5g),

$$\begin{aligned} \partial_T \mathbb{E}^{(0)}_{\mathrm{SF}} &= \int_M \big[2\tilde{\partial}_0 \phi' \cdot \phi' + 2\langle \nabla\phi, \nabla\tilde{\partial}_0\phi \rangle_{\boldsymbol{g}} + (\tilde{\partial}_0 \boldsymbol{g}^{-1})^{ab} \nabla_a \phi \nabla_b \phi + 3\hat{\boldsymbol{n}}(|\phi'|^2 + |\nabla\phi|^2_{\boldsymbol{g}}) \big] \operatorname{vol}_{\boldsymbol{g}} \\ &= \int_M \big[2(\langle \nabla\boldsymbol{n}, \nabla\phi \rangle_{\boldsymbol{g}} + \boldsymbol{n}\Delta_{\boldsymbol{g}}\phi + (1-\boldsymbol{n})\phi')\phi' - 2(\boldsymbol{n}\phi') \cdot \Delta_{\boldsymbol{g}}\phi - 2\boldsymbol{n}\langle \boldsymbol{\Sigma}, \nabla\phi\nabla\phi \rangle_{\boldsymbol{g}} + 3\hat{\boldsymbol{n}}|\phi'|^2 + \hat{\boldsymbol{n}}|\nabla\phi|^2_{\boldsymbol{g}} \big] \operatorname{vol}_{\boldsymbol{g}}. \end{aligned}$$

With $2(1-\boldsymbol{n}) = -4-6\hat{\boldsymbol{n}}$, integration by parts and using the bootstrap assumption (9-9) on C-norms, we get for some constant $K>0$ that we update from line to line

$$\begin{aligned}\partial_T \mathbb{E}^{(0)}_{\mathrm{SF}} &\leq \int_M -4|\phi'|^2_{\boldsymbol{g}} \operatorname{vol}_{\boldsymbol{g}} + K[\|\nabla\phi\|_{C^0}\|\hat{\boldsymbol{n}}\|_{H^1}\sqrt{\mathbb{E}^{(0)}_{\mathrm{SF}}} + (\|\boldsymbol{\Sigma}\|_{C^0}+\|\hat{\boldsymbol{n}}\|_{C^0})\mathbb{E}^{(0)}_{\mathrm{SF}}]\\ &\leq \int_M -4|\phi'|^2 \operatorname{vol}_{\boldsymbol{g}} + K\delta e^{-T/2}\big(\sqrt{\mathbb{E}^{(0)}_{\mathrm{SF}}}\|\hat{\boldsymbol{n}}\|_{H^1} + \mathbb{E}^{(0)}_{\mathrm{SF}}\big).\end{aligned}$$

Similarly and using the same evolution equations, we obtain

$$\begin{aligned}\partial_T \mathcal{C}^{(0)}_{\mathrm{SF}} &= \int_M [\tilde{\partial}_0\phi\cdot\phi' - \tilde{\partial}_0\bar{\phi}\cdot\phi' + (\phi-\bar{\phi})\tilde{\partial}_0\phi' + 3\hat{\boldsymbol{n}}(\phi-\bar{\phi})\phi']\operatorname{vol}_{\boldsymbol{g}}\\ &= \int_M [3|\phi'|^2 + 3\hat{\boldsymbol{n}}|\phi'|^2 + (\phi-\bar{\phi})\cdot\operatorname{div}_{\boldsymbol{g}}(\boldsymbol{n}\nabla\phi) - 2(\phi-\bar{\phi})\phi' - \tilde{\partial}_0\bar{\phi}\cdot\phi']\operatorname{vol}_{\boldsymbol{g}}\\ &\leq -2\mathcal{C}^{(0)}_{\mathrm{SF}} + \int_M 3[|\phi'|^2 - |\nabla\phi|^2_{\boldsymbol{g}}]\operatorname{vol}_{\boldsymbol{g}} + 3\|\hat{\boldsymbol{n}}\|_{C^0}\mathbb{E}^{(0)}_{\mathrm{SF}} - \int_M(\tilde{\partial}_0\bar{\phi}\cdot\phi')\operatorname{vol}_{\boldsymbol{g}}.\end{aligned}$$

Applying Lemma 9.15 to the last term, we get

$$\partial_T \mathcal{C}^{(0)}_{\mathrm{SF}} \leq -2\mathcal{C}^{(0)}_{\mathrm{SF}} + K\delta e^{-T/2}\mathbb{E}^{(0)}_{\mathrm{SF}} + K\delta^3 e^{-5T/2}.$$

Combining these two estimates, inserting (9-11a) and (9-12), as well as updating K, yields

$$\begin{aligned}\partial_T E^{(0)}_{\mathrm{SF}} &= \partial_T \mathbb{E}^{(0)}_{\mathrm{SF}} + \tfrac{2}{3}\partial_T \mathcal{C}^{(0)}_{\mathrm{SF}}\\ &= \int_M\big[\big(-4+\tfrac{2}{3}\cdot 3\big)|\phi'|^2 - \tfrac{2}{3}\cdot 3|\nabla\phi|^2_{\boldsymbol{g}}\big]\operatorname{vol}_{\boldsymbol{g}} - 2\cdot\tfrac{2}{3}\mathcal{C}^{(0)}_{\mathrm{SF}} + K\delta e^{-T/2}\big(\sqrt{\mathbb{E}^{(0)}_{\mathrm{SF}}}\sqrt{\|\hat{\boldsymbol{n}}\|_{H^1}} + \mathbb{E}^{(0)}_{\mathrm{SF}}\big) + K\delta^3 e^{-5T/2}\\ &\leq -2E^{(0)}_{\mathrm{SF}} + K\delta e^{-T/2}\sqrt{E^{(0)}_{\mathrm{SF}}}(\|\Sigma\|_{L^2} + \|\boldsymbol{g}-\gamma\|_{L^2} + \sqrt{E^{(0)}_{\mathrm{SF}}}) + K\delta^3 e^{-5T/2}.\end{aligned}$$

□

Lemma 9.17 (higher-order estimates for the corrected scalar field energy). *For any $l\in\{1,\dots,4\}$, the following estimate holds*:

$$\partial_T\big(\mathbb{E}^{(l)}_{\mathrm{SF}} + \tfrac{2}{3}\mathcal{C}^{(l)}_{\mathrm{SF}}\big) \leq -2\big(\mathbb{E}^{(l)}_{\mathrm{SF}} + \tfrac{2}{3}\mathcal{C}^{(l)}_{\mathrm{SF}}\big) + K\delta e^{-T/2}\Big(\sum_{m=0}^{l}\sqrt{\mathbb{E}^{(m)}_{\mathrm{SF}}}\Big)\cdot(\|\phi'\|_{H^l} + \|\nabla\phi\|_{H^l} + \|\Sigma\|_{H^l} + \|\boldsymbol{g}-\gamma\|_{H^l}).$$

Proof. Starting with $l=2k,\ k\in\{1,2\}$, one calculates

$$\begin{aligned}\partial_T \mathbb{E}^{(2k)}_{\mathrm{SF}} = \int_M\Big[&2\Delta^k_{\boldsymbol{g}}\tilde{\partial}_0\phi'\cdot\Delta^k_{\boldsymbol{g}}\phi' + 2\langle\nabla\Delta^k_{\boldsymbol{g}}\phi, \nabla\Delta^k_{\boldsymbol{g}}\tilde{\partial}_0\phi\rangle_{\boldsymbol{g}} && \text{(9-20a)}\\ &+ (\tilde{\partial}_0\boldsymbol{g}^{-1})^{ab}\cdot\nabla_a\Delta^k_{\boldsymbol{g}}\phi\cdot\nabla_b\Delta^k_{\boldsymbol{g}}\phi + 3\hat{\boldsymbol{n}}(|\Delta^k\phi'|^2_{\boldsymbol{g}} + |\nabla\Delta^k\phi|^2_{\boldsymbol{g}}) && \text{(9-20b)}\\ &+ 2[\tilde{\partial}_0, \Delta^k_{\boldsymbol{g}}]\phi'\cdot\Delta^k_{\boldsymbol{g}}\phi' + 2\langle[\tilde{\partial}_0, \nabla\Delta^k_{\boldsymbol{g}}]\phi, \nabla\Delta^k_{\boldsymbol{g}}\phi\rangle_{\boldsymbol{g}}\Big]\operatorname{vol}_{\boldsymbol{g}}. && \text{(9-20c)}\end{aligned}$$

We insert the rescaled wave equation (9-5g) and $\tilde{\partial}_0\phi = \boldsymbol{n}\phi'$ into the right-hand side of (9-20a) and obtain for some constant $K>0$ that we update from line to line

$$\begin{aligned}\text{(9-20a)} &\leq \int_M[-4|\Delta^k_{\boldsymbol{g}}\phi'|^2 - 6\hat{\boldsymbol{n}}|\Delta^k_{\boldsymbol{g}}\phi'|^2 + \boldsymbol{n}\Delta^{k+1}_{\boldsymbol{g}}\phi\cdot\Delta^k_{\boldsymbol{g}}\phi']\operatorname{vol}_{\boldsymbol{g}}\\ &\quad + K\|\Delta^k\phi'\|_{L^2}(\|\hat{\boldsymbol{n}}\|_{H^{2k+1}}\|\nabla\phi\|_{C^0} + \|\nabla\phi\|_{H^{2k}}\|\hat{\boldsymbol{n}}\|_{C^{2k}})\\ &\quad + \int_M[-\boldsymbol{n}\Delta^k_{\boldsymbol{g}}\phi'\cdot\Delta^{k+1}_{\boldsymbol{g}}\nabla\phi - 3\langle\nabla\hat{\boldsymbol{n}}, \nabla\Delta^k_{\boldsymbol{g}}\phi\rangle_{\boldsymbol{g}}\cdot\Delta^k_{\boldsymbol{g}}\phi']\operatorname{vol}_{\boldsymbol{g}}\\ &\quad + K\|\nabla\Delta^k_{\boldsymbol{g}}\phi\|_{L^2}(\|\hat{\boldsymbol{n}}\|_{H^{2k+1}}\|\phi'\|_{C^0} + \|\hat{\boldsymbol{n}}\|_{C^{2k}}\|\phi'\|_{H^{2k}})\\ &\leq \int_M -4|\Delta^k_{\boldsymbol{g}}\phi'|^2\operatorname{vol}_{\boldsymbol{g}} + K\sqrt{\mathbb{E}^{(2k)}_{\mathrm{SF}}}\cdot\big[(\|\nabla\phi\|_{C^0} + \|\phi'\|_{C^0})\cdot\|\hat{\boldsymbol{n}}\|_{H^{2k+1}} + (\|\nabla\phi\|_{H^{2k}} + \|\phi'\|_{H^{2k}})\cdot\|\hat{\boldsymbol{n}}\|_{C^{2k}}\big].\end{aligned}$$

For (9-20b), we use (9-5e) and the bootstrap assumption (9-9) to bound it by $K\delta e^{-T/2}\mathbb{E}_{\mathrm{SF}}^{(2k)}$. Regarding (9-20c), the commutator formulas (B-1a)–(B-1b) imply

$$\|[\tilde{\partial}_0, \Delta_{\boldsymbol{g}}^k]\phi'\|_{L^2} \lesssim \|\boldsymbol{n}\|_{C^{2k-1}}(\|\phi'\|_{C^1}\|\boldsymbol{\Sigma}\|_{\dot{H}^{2k-1}} + \|\boldsymbol{\Sigma}\|_{C^{2k-2}}\|\phi'\|_{H^{2k}}) + \|\hat{\boldsymbol{n}}\|_{C^{2k-1}}\|\phi'\|_{H^{2k}},$$
$$\|[\tilde{\partial}_0, \nabla\Delta_{\boldsymbol{g}}^k]\phi\|_{L^2} \lesssim \|\boldsymbol{n}\|_{C^{2k}}(\|\nabla\phi\|_{C^1}\|\boldsymbol{\Sigma}\|_{H^{2k}} + \|\boldsymbol{\Sigma}\|_{C^{2k-2}}\|\nabla\phi\|_{H^{2k}}) + \|\hat{\boldsymbol{n}}\|_{C^{2k}}\|\nabla\phi\|_{H^{2k}}.$$

Summarizing, inserting the C-norm bounds from the bootstrap assumption (9-9) and updating K, this implies

$$\partial_T\mathbb{E}_{\mathrm{SF}}^{(2k)} \le \int_M -4|\Delta_{\boldsymbol{g}}^k\phi'|^2\,\mathrm{vol}_{\boldsymbol{g}} + K\delta e^{-T/2}\mathbb{E}_{\mathrm{SF}}^{(2k)} + K\delta e^{-T/2}\sqrt{\mathbb{E}_{\mathrm{SF}}^{(2k)}}(\|\phi'\|_{H^{2k}} + \|\nabla\phi\|_{H^{2k}}) + K\delta e^{-T/2}\sqrt{\mathbb{E}_{\mathrm{SF}}^{(2k)}}(\|\hat{\boldsymbol{n}}\|_{H^{2k+1}} + \|\Sigma\|_{H^{2k}}).$$

Moving on to the corrective term, we compute

$$\partial_T\mathcal{C}_{\mathrm{SF}}^{(2k)} = \int_M \big[\Delta_{\boldsymbol{g}}^k\tilde{\partial}_0\phi\cdot\Delta_{\boldsymbol{g}}^k\phi' + \Delta_{\boldsymbol{g}}^k\phi\cdot\Delta_{\boldsymbol{g}}^k\tilde{\partial}_0\phi' \tag{9-21a}$$
$$+ 3\hat{\boldsymbol{n}}\cdot\Delta_{\boldsymbol{g}}^k\phi\cdot\Delta_{\boldsymbol{g}}^k\phi' + [\tilde{\partial}_0, \Delta_{\boldsymbol{g}}^k]\phi\cdot\Delta_{\boldsymbol{g}}\phi' + \Delta_{\boldsymbol{g}}^k\phi\cdot[\tilde{\partial}_0, \Delta_{\boldsymbol{g}}^k]\phi'\big]\,\mathrm{vol}_{\boldsymbol{g}}. \tag{9-21b}$$

Inserting the evolution equations into the right-hand side of (9-21a), we can bound that line by

$$\le \int_M [3|\Delta_{\boldsymbol{g}}^k\phi'|^2 + 3\hat{\boldsymbol{n}}|\Delta_{\boldsymbol{g}}^k\phi'|^2]\mathrm{vol}_{\boldsymbol{g}} + K\|\hat{\boldsymbol{n}}\|_{C^{2k}}\|\phi'\|_{H^{2k-1}}\|\Delta^k\phi'\|_{L^2}$$
$$+ \int_M \big[-2\Delta_{\boldsymbol{g}}^k\phi\cdot\Delta_{\boldsymbol{g}}^k\phi' + 3\hat{\boldsymbol{n}}\Delta_{\boldsymbol{g}}^k\phi\cdot\Delta_{\boldsymbol{g}}^k\phi' + 3\Delta_{\boldsymbol{g}}^k\phi\cdot\Delta_{\boldsymbol{g}}^{k+1}\phi + 3\hat{\boldsymbol{n}}\Delta_{\boldsymbol{g}}^k\phi\cdot\Delta_{\boldsymbol{g}}^{k+1}\phi\big]\,\mathrm{vol}_{\boldsymbol{g}}$$
$$+ K\big[\|\hat{\boldsymbol{n}}\|_{C^{2k}}(\|\nabla\phi\|_{H^{2k}} + \|\phi'\|_{H^{2k-1}}) + (\|\nabla\phi\|_{C^0} + \|\phi'\|_{C^0})\|\hat{\boldsymbol{n}}\|_{H^{2k+1}}\big]\|\Delta_{\boldsymbol{g}}^k\phi\|_{L^2}.$$

Note that, after integrating by parts, the last two terms in the second line can be bounded by

$$\int_M -3|\nabla\Delta_{\boldsymbol{g}}\phi|_{\boldsymbol{g}}^2\,\mathrm{vol}_{\boldsymbol{g}} + \|\hat{\boldsymbol{n}}\|_{C^1}(\|\nabla\Delta^k\phi\|_{L^2} + \|\Delta^k\phi\|_{L^2})\|\nabla\Delta^k\phi\|_{L^2}.$$

For the terms in (9-21b), notice that the first term can be bounded by $\delta e^{-T/2}\|\nabla\phi\|_{H^{2k-1}}\sqrt{\mathbb{E}_{\mathrm{SF}}^{(2k)}}$, while the commutator terms can be estimated as before, with

$$\|[\tilde{\partial}_0, \Delta_{\boldsymbol{g}}^k]\phi\|_{L^2} \lesssim \|\nabla\phi\|_{C^0}\|\boldsymbol{n}\|_{C^0}\|\boldsymbol{\Sigma}\|_{\dot{H}^{2k-1}} + \|\boldsymbol{n}\|_{C^{2k}}\|\boldsymbol{\Sigma}\|_{C^{2k-2}}\|\nabla\phi\|_{H^{2k-1}}.$$

Combining all of the above, we get

$$\partial_T\mathcal{C}_{\mathrm{SF}}^{(2k)} \le -2\mathcal{C}_{\mathrm{SF}}^{(2k)} + \int_M [3|\Delta_{\boldsymbol{g}}^k\phi'| - 3|\nabla\Delta_{\boldsymbol{g}}^k\phi|_{\boldsymbol{g}}]\,\mathrm{vol}_{\boldsymbol{g}} + K\delta e^{-T/2}[\|\phi'\|_{H^{2k}} + \|\nabla\phi\|_{H^{2k}} + \|\hat{\boldsymbol{n}}\|_{H^{2k+1}} + \|\boldsymbol{\Sigma}\|_{H^{2k}}]\cdot\big(\sqrt{\mathbb{E}_{\mathrm{SF}}^{(2k)}} + \sqrt{\mathbb{E}_{\mathrm{SF}}^{(2k-1)}}\big).$$

Finally, combining both differential estimates yields the statement for $l = 2k$. For $l = 2k-1$, $k \in \{1, 2\}$, the argument is completely analogous and hence omitted. □

9.4. ***Geometric variables.*** We can take the following results from prior literature, where we additionally apply the elliptic estimates in Lemma 9.11:

Lemma 9.18 (coercivity of geometric energies [Andersson and Moncrief 2011, Lemma 7.4]). *For sufficiently small $\delta > 0$, the following estimate holds*:

$$\|\boldsymbol{g} - \gamma\|_{H^5}^2 + \|\Sigma\|_{H^4}^2 \lesssim E_{\mathrm{geom}}. \tag{9-22}$$

Lemma 9.19 (geometric energy estimate [Andersson and Fajman 2020, Lemma 20]). *Let $\delta > 0$ be chosen appropriately small, and let*

$$\alpha = \begin{cases} 1, & \lambda_0 > \frac{1}{9}, \\ 1 - 3\sqrt{\delta'}, & \lambda_0 = \frac{1}{9}, \end{cases} \tag{9-23}$$

where $\delta' > 0$ is the same as in (9-7), *in particular, suitably small. Then, there exists some constant $K > 0$ such that the following estimate holds*:

$$\partial_T E_{\mathrm{geom}} \leq -2\alpha E_{\mathrm{geom}} + K E_{\mathrm{geom}}^{3/2} + K\delta e^{-T/2}\sqrt{E_{\mathrm{geom}}}[\|\phi'\|_{H^4} + \|\nabla\phi\|_{H^4}]. \tag{9-24}$$

9.5. ***Closing the bootstrap.*** Now, we can collect our estimates to improve the bootstrap assumptions:

Proposition 9.20 (improved bounds for future stability). *Let the bootstrap assumption* (*see Assumption 9.9*) *be satisfied for $T \in [0, T_{\mathrm{Boot}})$ and assume the initial data assumption holds at $T = 0$* (*see Assumption 9.7*). *For $\delta > 0$ sufficiently small and α as in* (9-23) *with $\delta' > 0$ sufficiently small, the following estimates hold*:

$$\|\phi'\|_{C^2} + \|\nabla\phi\|_{C^2} + \|\phi'\|_{H^4} + \|\nabla\phi\|_{H^4} \lesssim \delta^{3/2} e^{-\alpha T}, \tag{9-25a}$$

$$\|\boldsymbol{g} - \gamma\|_{C^3} + \|\Sigma\|_{C^2} + \|\boldsymbol{g} - \gamma\|_{H^5} + \|\Sigma\|_{H^4} \lesssim \delta^{3/2} e^{-\alpha T}, \tag{9-25b}$$

$$\|\hat{\boldsymbol{n}}\|_{C^4} + \|\boldsymbol{X}\|_{C^4} + \|\hat{\boldsymbol{n}}\|_{H^6} + \|\boldsymbol{X}\|_{H^6} \lesssim \delta^3 e^{-2\alpha T}. \tag{9-25c}$$

Proof. In the following, the positive constant K may be updated from line to line.

Combining the estimate from Lemma 9.16 as well as those from Lemma 9.17 at each level with Lemma 9.19 and applying the (near)-coercivity estimates (9-14) and (9-22) to the right-hand sides, we obtain

$$\begin{aligned}\partial_T(E_{\mathrm{SF}} + E_{\mathrm{geom}}) \leq -2E_{\mathrm{SF}} + K\delta e^{-T/2}\sqrt{E_{\mathrm{SF}}}\big(\sqrt{E_{\mathrm{SF}} + \delta^2 e^{-T}\left\|\mathrm{Ric}[\boldsymbol{g}] + \tfrac{2}{9}\boldsymbol{g}\right\|_{H^2}^2} + \sqrt{E_{\mathrm{geom}}}\big) + K\delta^3 e^{-5T/2} \\ - 2\alpha E_{\mathrm{geom}} + K E_{\mathrm{geom}}^{3/2} + K\delta e^{-T/2}\sqrt{E_{\mathrm{geom}}}\sqrt{E_{\mathrm{SF}} + \delta^2 e^{-T}\left\|\mathrm{Ric}[\boldsymbol{g}] + \tfrac{2}{9}\boldsymbol{g}\right\|_{H^2}^2}.\end{aligned}$$

Applying (9-10a) to the curvature norms, as well as (9-22) to the resulting norms on $\boldsymbol{g} - \gamma$ and (9-9) (which implies $\sqrt{E_{\mathrm{geom}}} \lesssim \delta e^{-T/2}$), this becomes

$$\partial_T(E_{\mathrm{SF}} + E_{\mathrm{geom}}) \leq -2\alpha(E_{\mathrm{SF}} + E_{\mathrm{geom}}) + K\delta e^{-T/2}(E_{\mathrm{SF}} + E_{\mathrm{geom}}) + K\delta^3 e^{-5T/2},$$

and consequently, since $\alpha \leq 1$,

$$\partial_T[e^{2\alpha T}(E_{\mathrm{SF}} + E_{\mathrm{geom}})] \lesssim \delta e^{-T/2} \cdot e^{2\alpha T}(E_{\mathrm{SF}} + E_{\mathrm{geom}}) + \delta^3 e^{-T/2}.$$

The Gronwall lemma, along with the initial data assumption (9-8), now implies

$$E_{\mathrm{SF}} + E_{\mathrm{geom}} \lesssim \delta^3 e^{-2\alpha T}. \tag{9-26}$$

Lemma 9.18 and the standard Sobolev embedding then imply (9-25b). In particular, this means

$$\left\|\mathrm{Ric}[\boldsymbol{g}]+\tfrac{2}{9}\boldsymbol{g}\right\|_{H^2}\lesssim\delta^{3/2}e^{-\alpha T} \tag{9-27}$$

due to Lemma 9.10, and for $\delta'>0$ small enough, inserting (9-26) and (9-27) into (9-14) shows (9-25a). Moreover, (9-25c) follows directly from the proof of Lemma 9.11 and the already obtained improvements. □

Proof of Theorem 9.1. The problem is locally well-posed as outlined in Remark 9.8. There then is some maximal interval $[0, T_{\mathrm{Boot}})$ for the logarithmic time T — or, equivalently, some maximal time interval $[\tau_0, \tau_{\mathrm{Boot}})$ — on which the solution exists and the bootstrap assumptions (see Assumption 9.9) are satisfied. By the analogous argument to the proof of Theorem 8.2, the decay estimates in Proposition 9.20 are strictly stronger than the bootstrap assumptions for small enough $\delta, \delta'>0$. This implies $T_{\mathrm{Boot}}=\infty$ (resp. $\tau_{\mathrm{Boot}}=0$) since we could else extend the solution strictly beyond T_{Boot} while also satisfying the bootstrap assumptions. This proves the convergence statement in Theorem 9.1.

Finally, the decay estimates imply that $|\nabla n|_g$, respectively $|k|_g$, are bounded by $\tau^{\alpha-1}$, respectively $\tau^{\alpha+1}$, up to constant on $[\tau_0,\tau)$. Since α is at worst slightly smaller than 1, both functions are integrable on $[\tau_0,0)$ for suitably small $\delta'>0$. By [Choquet-Bruhat and Cotsakis 2002], this means the spacetime is future complete. □

10. Global stability

To prove Theorem 1.2, what still needs to be shown is that initial data as in Theorem 1.1 develops from Σ_{t_0} to some hypersurface $\Sigma_{t_1}\equiv\Sigma_{\tau(t_1)}$ in its future such that the data in Σ_{t_1} is near-Milne in the sense of Assumption 9.7 and in CMCSH gauge. From there, near-Milne stability yields the behaviour in the future of $\Sigma_{\tau(t_1)}$, and hence future stability of near-FLRW spacetimes as in Theorem 1.2.

Proof of Theorem 1.2. Within this proof, t will denote the "physical" time coordinate used throughout the big bang stability analysis, while τ denotes the mean curvature time used within CMCSH gauge.

Consider initial data $(g, k, \nabla\phi, \partial_0\phi)$ induced on the CMC hypersurface Σ_{t_0} within $\overline{M}$ such that the rescaled variables are close to FLRW reference data in the sense of Theorem 8.2. Moreover, let $(\mathring{g}, \mathring{k}, \mathring{\pi}, \mathring{\psi})$ be the geometric initial data on M that induce it via the embedding $\iota: M\hookrightarrow\overline{M}$. Notice that

$$P: H^{20}_\gamma(M)\to H^{18}_\gamma(M),\quad Y^i\mapsto\Delta_\gamma Y^i+(\gamma^{-1})^{il}\mathrm{Ric}[\gamma]_{lj}Y^j=\Delta_\gamma Y^i-\tfrac{2}{9}Y^i,$$

is an isomorphism since Δ_γ has no positive eigenvalues. Hence, using [Fajman and Kröncke 2020, Theorem 2.5, Remark 2.6], there is a metric $\mathring{g}'$ isometric to $\mathring{g}$ that remains close in $H^{18}_\gamma(M)$ to $a(t_0)^2\gamma$ and satisfies

$$((\mathring{g}')^{-1})^{ij}(\Gamma[\mathring{g}']^k_{ij}-\widehat{\Gamma}[\gamma]^k_{ij})=0.$$

Let $\theta\in\mathrm{Diff}(M)$ be the diffeomorphism such that $\theta^*\mathring{g}=\mathring{g}'$. Then the proof of [Fajman and Kröncke 2020, Theorem 2.5] implies that θ can be chosen close to the identity map within $H^{18}(\mathrm{Diff}(M))$, and consequently that $\theta^*\mathring{k}=\mathring{k}'$, $\theta^*\mathring{\pi}=\mathring{\pi}'$ and $\theta^*\mathring{\psi}=\mathring{\psi}'$ remain close to $-\dot{a}(t_0)a(t_0)\gamma$, 0 and $Ca(t_0)^{-3}$ in

$H^{18}_{\gamma}(M)$. By the same argument as in Remark 8.1, we can now evolve this data locally and obtain a new initial hypersurface Σ' close to Σ_{t_0} that is in CMCSH gauge and that $(g, k, \nabla\phi, \partial_0\phi)$ is close to the reference data in the sense of Assumption 3.10, exchanging the initial time t_0 by some close time t_0'.

Since τ is strictly increasing, $t \equiv t(\tau)$ exists and we can interchangeably view a as a function in t or τ with some abuse of notation. The Friedman equation (2-3) implies $\partial_t a \geq \frac{1}{9}$ and thus $a(t) \geq \frac{1}{9}t$ on $(0, \infty)$, as well as

$$-\tau = 3\frac{\dot{a}}{a} = \frac{1}{a} + \langle \text{lower-order terms} \rangle \quad \text{as } t \to \infty \text{ (resp. } \tau \to 0).$$

We choose $t_1 > \max\{1, t_0'\}$ large enough (resp. $\tau(t_1) \equiv \tau_0$ small enough) that the following estimates hold for some small $\chi \in \left(0, \frac{1}{2}\right)$ that depends only on δ:

$$Ca(t_1)^{-3}\tau(t_1)^{-1} \leq \chi, \tag{10-1}$$

$$-\tau(t_1) \cdot a(t_1) \in [1 - \chi, 1 + \chi]. \tag{10-2}$$

As the solution is Cauchy stable, i.e., it and its maximal time of existence depend continuously upon the initial data,[15] one can choose $\varepsilon > 0$ in the analogue of Assumption 3.10 small enough to ensure the following: The solution exists until $t_1 > t_0'$ and $(a^{-2}g, a\hat{k}, \nabla\phi, a^3\tilde{\partial}_0\phi)$ remain $K\varepsilon$-close to $(\gamma, 0, 0, C)$ in $H^6_\gamma \times H^5_\gamma \times H^5_\gamma \times H^5_\gamma$ for some suitable $K > 0$ along the slab $\bigcup_{s\in[t_0', t_1]} \Sigma_s$. What now remains to be shown is that this implies Assumption 9.7 in the sense that, if ε is small enough, δ can be made as small as necessary for Theorem 9.1 to apply.

Note that the scalings in Definition 9.4 can be rewritten as

$$\boldsymbol{g} - \gamma = (\tau \cdot a)^2 \cdot (a^{-2}g - \gamma) + (\tau^2 \cdot a^2 - 1)\gamma, \quad \boldsymbol{\Sigma} = \frac{\tau}{a}(a\hat{k}),$$
$$\phi' = C(-\tau^{-1} \cdot a^{-3}) + (-\tau^{-1} \cdot a^{-3}) \cdot (a^3 n^{-1}(\partial_\tau - \mathcal{L}_X)\phi - C).$$

Since (10-2) implies $\tau \cdot a$ is close to -1 at t_1, $\|(\tau \cdot a)^2(a^{-2}g - \gamma)\|_{H^6}$ can be bounded by $\frac{1}{2}\delta^3$ for small enough ε. Choosing $\chi < \frac{1}{2}\delta^3$ then implies $\|\boldsymbol{g} - \gamma\|_{H^6(\Sigma_{\tau_0})} < \delta^3$. That $\|\boldsymbol{\Sigma}\|_{H^5}$ can be made smaller than δ^3 for small enough $\varepsilon > 0$ follows since τ/a behaves like $1/a^2$ up to a constant by (10-2).

For the normal derivative of the wave, notice that $|C(-\tau^{-1} \cdot a^{-3})|$ is bounded by χ due to (10-1), and that $-\tau^{-1}a^{-3}$ is equivalent to a^{-2} by (10-2). Hence, we can similarly ensure that ϕ' is bounded in H^5 by δ^3. Since $\nabla\phi$ is not changed in either rescaling, and bounds on lapse and shift (up to constant) follow from the elliptic estimates in Lemma 9.11, it follows each individual norm in Assumption 9.7 can be bounded by δ^3 up to constants that depend only on γ, and hence the initial data assumption itself can be satisfied for suitably small $\delta > 0$.

This proves that we can develop from initial data for the big bang stability proof to near-Milne initial data within a CMCSH foliation, and thus we obtain Theorem 1.2 from Theorems 1.1 and 9.1. □

[15] For the argument for Einstein vacuum in CMCSH gauge, see [Andersson and Moncrief 2004, Theorem 3.1]. As with local existence, the argument in the Einstein scalar-field system is largely identical since the only difference amounts to coupling the hyperbolic parts of the system with a further hyperbolic one.

Appendix A: Big bang stability

A.1. *Basic formulas and estimates.*

A.1.1. *Tools from elementary calculus.*

Lemma A.1 (a Gronwall lemma). *Let $f, \chi, \xi : [a, b] \to \mathbb{R}$ be continuous functions such that $\chi \geq 0$, ξ is decreasing and, for any $s \in [a, b]$,*

$$f(s) \leq \int_s^b \chi(r) f(r)\, dr + \xi(s)$$

is satisfied. Then, for any $t \in [a, b]$, we have

$$f(t) \leq \xi(t) \exp\Big(\int_t^b \chi(r)\, dr\Big).$$

Proof. This follows by standard arguments as in [Dragomir 2003, Corollary 2-3]. □

Lemma A.2 (a weak fundamental theorem of calculus for square roots). *Let $f : (0, t_0] \to \mathbb{R}_0^+$ be a C^1-function. Then, we have for any $t \in (0, t_0]$*

$$\sqrt{f(t)} \leq \sqrt{f(t_0)} + \int_t^{t_0} \frac{|f'(s)|}{2\sqrt{f(s)}}\, ds. \tag{A-1}$$

Proof. This follows from a straightforward application of the monotone convergence theorem to $g_n = \sqrt{f + 1/n}$. □

A.1.2. *Levi-Civita tensor identities.* Herein, we collect some basic identities for the Levi-Civita tensor $\boldsymbol{\varepsilon}[g]$: Firstly, it satisfies the contraction identities, where $\mathbb{I}^a_b$ denotes the Kronecker-symbol:

$$\boldsymbol{\varepsilon}^{ai_1i_2}\boldsymbol{\varepsilon}_{aj_1j_2} = \mathbb{I}^{i_1}_{j_1}\mathbb{I}^{i_2}_{j_2} - \mathbb{I}^{i_1}_{j_2}\mathbb{I}^{i_2}_{j_1}, \tag{A-2a}$$

$$\boldsymbol{\varepsilon}^{abi}\boldsymbol{\varepsilon}_{abj_2} = 2\mathbb{I}^i_j, \tag{A-2b}$$

$$\boldsymbol{\varepsilon}^{abc}\boldsymbol{\varepsilon}_{abc} = 6, \tag{A-2c}$$

$$\nabla\boldsymbol{\varepsilon} = 0. \tag{A-2d}$$

The analogous formulas hold for $\boldsymbol{\varepsilon}[G]$ when raising indices with regard to G instead of g.

For a tracefree and symmetric Σ_t-tangent $(0, 2)$-tensor $\mathfrak{T}$ and a Σ_t-tangent $(0, 2)$-tensor $\mathfrak{A}$, the following simplified identities hold:

$$(\mathfrak{T} \times \mathfrak{A})_{ij} = \boldsymbol{\varepsilon}_i{}^{ab}\boldsymbol{\varepsilon}_j{}^{pq}\mathfrak{T}_{ap}\mathfrak{A}_{bq} + \tfrac{1}{3}(\mathfrak{T} \cdot \mathfrak{A})g_{ij}, \tag{A-3a}$$

$$(\mathfrak{T} \times g)_{ij} = -\mathfrak{T}_{ij}, \tag{A-3b}$$

$$(\mathfrak{T} \times k)_{ij} = -\tfrac{1}{3}\tau\mathfrak{T}_{ij} + (\mathfrak{T} \times \hat{k})_{ij}. \tag{A-3c}$$

Further, note the following formulas (for $\widetilde{\mathfrak{T}}$ as $\mathfrak{T}$, $\widetilde{\mathfrak{A}}$ as $\mathfrak{A}$ and any Σ_t-tangent $(0, 1)$-tensor ξ) (see [Andersson and Moncrief 2004, p. 30]):

$$\operatorname{div}_g(\mathfrak{A} \wedge \widetilde{\mathfrak{A}}) = -\operatorname{curl}\mathfrak{A} \cdot \widetilde{\mathfrak{A}} + \mathfrak{A} \cdot \operatorname{curl}\widetilde{\mathfrak{A}}, \tag{A-3d}$$

$$\mathfrak{A} \cdot (\xi \wedge \widetilde{\mathfrak{A}}) = -2\xi \cdot (\mathfrak{A} \wedge \widetilde{\mathfrak{A}}), \tag{A-3e}$$

$$\mathfrak{T} \cdot (\mathfrak{A} \times \widetilde{\mathfrak{T}}) = (\mathfrak{T} \times \mathfrak{A}) \cdot \widetilde{\mathfrak{T}}. \tag{A-3f}$$

A.1.3. *Estimates on contracted tensors.*

Lemma A.3. *Let $\mathfrak{S}, \mathfrak{T}$ be traceless and symmetric Σ_t-tangent $(0,2)$-tensors, $\mathfrak{M}, \mathfrak{N}$ symmetric Σ_t-tangent $(0,2)$-tensors and ξ a Σ_t-tangent $(0,1)$-tensor. We define G, G^{-1} and $|\cdot|_G$ via* (2-27a)). *Then*

$$|\mathfrak{M} \odot_G \mathfrak{N}|_G \leq |\mathfrak{M}|_G |\mathfrak{N}|_G, \qquad \mathfrak{M} \odot_g \mathfrak{N} = a^{-2} \mathfrak{M} \odot_G \mathfrak{N}, \tag{A-4a}$$

$$|\mathfrak{S} \times_G \mathfrak{T}|_G \lesssim |\mathfrak{S}|_G |\mathfrak{T}|_G, \qquad (\mathfrak{S} \times \mathfrak{T})_{ij} = a^{-3} (\mathfrak{S} \times_G \mathfrak{T})_{ij}, \tag{A-4b}$$

$$|\mathfrak{S} \wedge_G \mathfrak{T}|_G \leq |\mathfrak{S}|_G |\mathfrak{T}|_G, \qquad (\mathfrak{S} \wedge \mathfrak{T})_l = a^{-3} (\mathfrak{S} \wedge_G \mathfrak{T}), \tag{A-4c}$$

$$|\xi \wedge_G \mathfrak{T}|_G \leq |\xi|_G |\mathfrak{T}|_G, \qquad (\xi \wedge \mathfrak{T})_{ij} = a^{-1} (\xi \wedge_G \mathfrak{T})_{ij}, \tag{A-4d}$$

$$|\mathrm{curl}_G \mathfrak{M}|_G \lesssim |\nabla \mathfrak{M}|_G, \qquad \mathrm{curl}\, \mathfrak{M}_{ij} = a^{-1} \mathrm{curl}_G \mathfrak{M}_{ij}. \tag{A-4e}$$

Proof. The estimates with respect to the unrescaled metric are direct consequences of the contraction identities (A-2a)–(A-2c) replacing g with G, and the scalings follow simply by tracking the effects of the rescaling in Definition 2.9. In particular, note

$$\boldsymbol{\varepsilon}[g]_i{}^{cd} = g^{cj} g^{dk} \boldsymbol{\varepsilon}[g]_{ijk} = (a^{-2}(G^{-1})^{cj})(a^{-2}(G^{-1})^{dk}) a^3 \boldsymbol{\varepsilon}[G]_{ijk} = a^{-1} \varepsilon[G]_i{}^{\sharp cd}, \tag{A-5}$$

In particular, (A-5) determines the Levi-Civita symbol. □

A.2. ***Commutators.*** Herein, we collect a variety of commutators of spatial derivative operators with each other as well as with time derivatives. While these mostly follow by standard computations, we use the fact that our spatial hypersurfaces are three-dimensional to significantly simplify the spatial commutator formulas, and need to apply the rescaled equations from Proposition 2.10 for the time derivative formulas.

For higher-order commutators, we denote by $\mathfrak{J}$ terms within the commutator formula that contribute junk terms at any point where this commutator formula is used. Furthermore, in the following, ζ denotes a scalar function on $\overline{M}$ and $\mathfrak{T}$ denotes a Σ_t-tangent, symmetric $(0,2)$-tensor, always with sufficient regularity for the equations to make sense. Moreover, recall the schematic $*$-notation as introduced in Section 2.1.8.

Corollary A.4 (schematic first-order spatial commutators). *For ζ and $\mathfrak{T}$ as above, the following identities hold*:

$$[\Delta, \nabla]\zeta = \mathrm{Ric}[G] * \nabla\zeta, \tag{A-6a}$$

$$[\Delta, \nabla^2]\zeta = \mathrm{Ric}[G] * \nabla^2\zeta + \nabla \mathrm{Ric}[G] * \nabla\zeta, \tag{A-6b}$$

$$[\Delta, \nabla]\mathfrak{T} = \mathrm{Ric}[G] * \nabla\mathfrak{T} + \nabla \mathrm{Ric}[G] * \mathfrak{T}, \tag{A-6c}$$

$$[\Delta, \nabla^2]\mathfrak{T} = \mathrm{Ric}[G] * \nabla^2\mathfrak{T} + \nabla \mathrm{Ric}[G] * \nabla\mathfrak{T} + \nabla^2 \mathrm{Ric}[G] * \mathfrak{T}, \tag{A-6d}$$

$$[\Delta, \mathrm{div}_G]\mathfrak{T} = \mathrm{Ric}[G] * \nabla\mathfrak{T} + \nabla \mathrm{Ric}[G] * \mathfrak{T}, \tag{A-6e}$$

$$[\Delta, \mathrm{curl}_G] = \boldsymbol{\varepsilon}[G] * (\mathrm{Ric}[G] * \nabla\mathfrak{T} + \nabla \mathrm{Ric}[G] * \mathfrak{T}). \tag{A-6f}$$

Proof. Since we are working in three spatial dimensions, the following identity holds:

$$\mathrm{Riem}[G]_{ijkl} = G_{ik}\,\mathrm{Ric}[G]_{jl} - G_{il}\,\mathrm{Ric}[G]_{jk} + G_{jl}\,\mathrm{Ric}[G]_{ik} - G_{jk}\,\mathrm{Ric}[G]_{il} - \tfrac{1}{2}(G^{-1})^{mn}\,\mathrm{Ric}[G]_{mn}(G_{ik}G_{jl} - G_{il}G_{jk}).$$

Hence, for any $I \in \mathbb{N}_0$, any $\nabla^I \operatorname{Riem}[G]$-term reduces to a sum of products and contractions of $\nabla^I \operatorname{Ric}[G]$ with various metric tensors that are all suppressed in schematic notation. With this in mind, the above statements are simply direct consequences of standard commutation cormulas and (A-5). □

Lemma A.5 (higher-order spatial commutators). *For $l \in \mathbb{N}$, $l \geq 2$, the following formulas hold* (*and extend to $l = 1$ when dropping any term involving Δ^{l-2}*):

$$[\Delta^l, \nabla]\zeta = \Delta^{l-1} \operatorname{Ric}[G] * \nabla\zeta + \nabla\Delta^{l-2} \operatorname{Ric}[G] * \nabla^2\zeta + \mathfrak{J}([\Delta^l, \nabla]\zeta), \tag{A-7a}$$

$$[\Delta^l, \nabla^2]\zeta = \nabla\Delta^{l-1} \operatorname{Ric}[G] * \nabla\zeta + \nabla^2\Delta^{l-2} \operatorname{Ric}[G] * \nabla^2\zeta + \mathfrak{J}([\Delta^l, \nabla^2]\zeta), \tag{A-7b}$$

$$[\Delta^l, \nabla]\mathfrak{T} = \nabla\Delta^{l-1} \operatorname{Ric}[G] * \mathfrak{T} + \nabla^2\Delta^{l-2} \operatorname{Ric}[G] * \nabla\mathfrak{T} + \mathfrak{J}([\Delta^l, \nabla]\mathfrak{T}), \tag{A-7c}$$

$$[\Delta^l, \nabla^2]\mathfrak{T} = \nabla^2\Delta^{l-1} \operatorname{Ric}[G] * \mathfrak{T} + \nabla^3\Delta^{l-2} \operatorname{Ric}[G] * \nabla\mathfrak{T} + \mathfrak{J}([\Delta^l, \nabla^2]\mathfrak{T}), \tag{A-7d}$$

$$[\Delta^l, \operatorname{div}_G]\mathfrak{T} = \nabla\Delta^{l-1} \operatorname{Ric}[G] * \mathfrak{T} + \nabla^2\Delta^{l-2} \operatorname{Ric}[G] * \nabla\mathfrak{T} + \mathfrak{J}([\Delta^l, \operatorname{div}_G]\mathfrak{T}), \tag{A-7e}$$

$$[\Delta^l, \operatorname{curl}_G]\mathfrak{T} = \boldsymbol{\varepsilon}[G] * (\nabla\Delta^{l-1} \operatorname{Ric}[G] * \mathfrak{T} + \nabla^2\Delta^{l-2} \operatorname{Ric}[G] * \nabla\mathfrak{T}) + \mathfrak{J}([\Delta^l, \operatorname{curl}_G]\mathfrak{T}), \tag{A-7f}$$

with junk terms, where $\mathcal{I} = I_1 + \cdots + I_{l-m}$,

$$\mathfrak{J}([\Delta^l,\nabla]\zeta) = \sum_{\substack{I_1+I_\zeta=2(l-1)\\ I_\zeta\geq 2}} \nabla^{I_1} \operatorname{Ric}[G]*\nabla^{I_\zeta+1}\zeta + \sum_{m=0}^{l-2} \sum_{\mathcal{I}+I_\zeta=2m} \nabla^{I_1} \operatorname{Ric}[G]*\cdots*\nabla^{I_{l-m}} \operatorname{Ric}[G]*\nabla^{I_\zeta+1}\zeta,$$

$$\mathfrak{J}([\Delta^l,\nabla^2]\zeta) = \sum_{\substack{I_1+I_\zeta=2(l-1)+1\\ I_1,I_\zeta\geq 2}} \nabla^{I_1} \operatorname{Ric}[G]*\nabla^{I_\zeta+1}\zeta + \sum_{m=0}^{l-2} \sum_{\mathcal{I}+I_\zeta=2m+1} \nabla^{I_1} \operatorname{Ric}[G]*\cdots*\nabla^{I_{l-m}} \operatorname{Ric}[G]*\nabla^{I_\zeta+1}\zeta,$$

$$\mathfrak{J}([\Delta^l,\nabla]\mathfrak{T}) = \sum_{\substack{I_1+I_\mathfrak{T}=2(l-1)+1\\ I_\mathfrak{T}\geq 2}} \nabla^{I_1} \operatorname{Ric}[G]*\nabla^{I_\mathfrak{T}}\mathfrak{T} + \sum_{m=0}^{l-2} \sum_{\mathcal{I}+I_\mathfrak{T}=2m+1} \nabla^{I_1} \operatorname{Ric}[G]*\cdots*\nabla^{I_{l-m}} \operatorname{Ric}[G]*\nabla^{I_\mathfrak{T}}\mathfrak{T},$$

$$\mathfrak{J}([\Delta^l,\nabla^2]\mathfrak{T}) = \sum_{\substack{I_1+I_\mathfrak{T}=2l\\ I_\mathfrak{T}\geq 2}} \nabla^{I_1} \operatorname{Ric}[G]*\nabla^{I_\mathfrak{T}}\mathfrak{T} + \sum_{m=0}^{l-2} \sum_{\mathcal{I}+I_\mathfrak{T}=2m+2} \nabla^{I_1} \operatorname{Ric}[G]*\cdots*\nabla^{I_{l-m}} \operatorname{Ric}[G]*\nabla^{I_\mathfrak{T}} T,$$

$$\mathfrak{J}([\Delta^l,\operatorname{div}_G]\mathfrak{T}) = \sum_{\substack{I_1+I_\mathfrak{T}=2(l-1)+1\\ I_\mathfrak{T}\geq 2}} \nabla^{I_1} \operatorname{Ric}[G]*\nabla^{I_\mathfrak{T}}\mathfrak{T} + \sum_{m=0}^{l-2} \sum_{\mathcal{I}+I_\mathfrak{T}=2m+1} \nabla^{I_1} \operatorname{Ric}[G]*\cdots*\nabla^{I_{l-m}} \operatorname{Ric}[G]*\nabla^{I_\mathfrak{T}}\mathfrak{T},$$

$$\begin{aligned}\mathfrak{J}([\Delta^l,\operatorname{curl}_G]\mathfrak{T}) = \boldsymbol{\varepsilon}[G]*\Bigg[&\sum_{\substack{I_1+I_\mathfrak{T}=2(l-1)+1\\ I_\mathfrak{T}\geq 2}} \nabla^{I_1} \operatorname{Ric}[G]*\nabla^{I_\mathfrak{T}}\mathfrak{T}\\ &+ \sum_{m=0}^{l-2} \sum_{\mathcal{I}+I_\mathfrak{T}=2m+1} \nabla^{I_1} \operatorname{Ric}[G]*\cdots*\nabla^{I_{l-m}} \operatorname{Ric}[G]*\nabla^{I_\mathfrak{T}}\mathfrak{T}\Bigg].\end{aligned}$$

Proof. The formulas follow by applying the formulas from Corollary A.4 inductively. □

Lemma A.6 (time derivative commutators). *With respect to a solution to the Einstein scalar-field system as in Proposition 2.6, the following commutator formulas hold*:

$$[\partial_t, \nabla_i]\zeta = 0, \tag{A-8a}$$

$$[\partial_t, \nabla^{\sharp i}]\zeta = 2(N+1)a^{-3}\Sigma^{\sharp ij}\nabla_j\zeta - 2N\frac{\dot{a}}{a}\nabla^{\sharp i}\zeta, \tag{A-8b}$$

$$[\partial_t,\Delta]\zeta = 2(N+1)a^{-3}\langle\Sigma,\nabla^2\zeta\rangle_G - 2N\frac{\dot a}{a}\Delta\zeta \\ -2(N+1)a^{-3}\langle \mathrm{div}_G\,\Sigma,\nabla\zeta\rangle_G - 2a^{-3}\langle\Sigma,\nabla N\nabla\zeta\rangle_G + \frac{\dot a}{a}\langle\nabla N,\nabla\zeta\rangle_G, \quad \text{(A-8c)}$$

$$[\partial_t,\nabla]\mathfrak{T} = a^{-3}((N+1)\nabla\Sigma+\Sigma*\nabla N)*\mathfrak{T}+\frac{\dot a}{a}\nabla N*\mathfrak{T}, \quad \text{(A-8d)}$$

$$[\partial_t,\Delta]\mathfrak{T} = a^{-3}(N+1)\Sigma*\nabla^2\mathfrak{T}+\frac{\dot a}{a}N\Delta\mathfrak{T}+a^{-3}\nabla((N+1)\Sigma)*\nabla\mathfrak{T}+\frac{\dot a}{a}\nabla N*\nabla\mathfrak{T} \\ +a^{-3}\nabla^2((N+1)\Sigma)*\mathfrak{T}-\frac{\dot a}{a}\nabla^2 N*\mathfrak{T}. \quad \text{(A-8e)}$$

Proof. Equation (A-8a) is simply that coordinate derivatives commute, and (A-8b) follows by applying (2-28b) and the product rule.

For the commutators (A-8c), (A-8d) and (A-8e), we write out the covariant derivatives in local coordinates, apply the product rule, and then the evolution equations (2-28b) and (2-34) for the inverse metric and Christoffel symbols. □

Lemma A.7 (high-order time derivative commutators). *For $l\in\mathbb{N}$, $l\geq 2$, the time derivative commutators take the form*

$$[\partial_t,\Delta^l]\zeta = 2a^{-3}(N+1)\langle\Sigma,\nabla^2\Delta^{l-1}\zeta\rangle_G + a^{-3}\nabla\Sigma*\nabla^3\Delta^{l-2}\zeta - 2(N+1)a^{-3}\langle\mathrm{div}_G\,\Delta^{l-1}\Sigma,\nabla\zeta\rangle_G \\ +(N+1)a^{-3}\nabla^{2l-3}\mathrm{Ric}*\Sigma*\nabla\zeta+\mathfrak{J}([\partial_t,\Delta^l]\zeta), \quad \text{(A-9a)}$$

$$[\partial_t,\nabla\Delta^l]\zeta = 2a^{-3}(N+1)\langle\Sigma,\nabla^3\Delta^{l-1}\zeta\rangle_G + a^{-3}(N+1)\nabla\Sigma*\nabla^{2l}\zeta \\ -2(N+1)a^{-3}\langle\nabla\,\mathrm{div}_G\,\Delta^{l-1}\Sigma,\nabla\zeta\rangle_G \\ +\frac{\dot a}{a}\langle\nabla^2\Delta^{l-1}N,\nabla\zeta\rangle_G + (N+1)a^{-3}\nabla^{2l-2}\mathrm{Ric}[G]*\Sigma*\nabla\zeta+\mathfrak{J}([\partial_t,\nabla\Delta^l]\zeta), \quad \text{(A-9b)}$$

$$[\partial_t,\Delta^l]\mathfrak{T} = a^{-3}\big(\Sigma*\nabla^2\Delta^{l-1}\mathfrak{T}+\nabla\Sigma*\nabla^3\Delta^{l-2}\mathfrak{T}+\nabla\mathfrak{T}*\nabla\Delta^{l-1}\Sigma+\mathfrak{T}*\Delta^l\Sigma\big) \\ +a^{-3}\big((N+1)\Sigma*\mathfrak{T}*\nabla^2\Delta^{l-2}\mathrm{Ric}[G]+\nabla((N+1)\Sigma*\mathfrak{T})*\nabla^{2l-3}\mathrm{Ric}[G]\big) \\ +\frac{\dot a}{a}\Delta^l N\cdot\mathfrak{T}+\frac{\dot a}{a}\nabla\Delta^{l-1}N*\nabla\mathfrak{T}+\mathfrak{J}([\partial_t,\Delta^l])\mathfrak{T}, \quad \text{(A-9c)}$$

$$[\partial_t,\nabla\Delta^l]\mathfrak{T} = a^{-3}\nabla\Sigma*\Delta^l\mathfrak{T}+a^{-3}(N+1)\Sigma*\nabla^3\Delta^{l-1}\mathfrak{T}+a^{-3}(N+1)\mathfrak{T}*\nabla\Delta^l\Sigma \\ +\frac{\dot a}{a}\nabla\Delta^l N*\mathfrak{T}+\frac{\dot a}{a}\nabla^2\Delta^{l-1}N*\nabla\mathfrak{T}+a^{-3}(N+1)\Sigma*\nabla^3\Delta^{l-2}\mathrm{Ric}[G]*\mathfrak{T}+\mathfrak{J}([\partial_t,\nabla\Delta^l]\mathfrak{T}), \quad \text{(A-9d)}$$

where the junk terms are, where $\mathcal{I}=\sum_{i=1}^{l-m-1}I_i$,

$$\mathfrak{J}([\partial_t,\Delta^l]\zeta) = a^{-3}\sum_{\substack{I_N+I_\Sigma+I_\zeta=2(l-1)\\ I_\zeta\leq 2(l-2)}}\nabla^{I_N}(N+1)*\nabla^{I_\Sigma}\Sigma*\nabla^{I_\zeta+2}\zeta+\frac{\dot a}{a}\sum_{\substack{I_N+I_\zeta=2l\\ I_\zeta\geq 2}}\nabla^{I_N}N*\nabla^{I_\zeta}\zeta \\ +a^{-3}\sum_{m=0}^{l-2}\sum_{I_N+I_\Sigma+I_\zeta+\mathcal{I}=2m}\nabla^{I_N}(N+1)*\nabla^{I_\Sigma}\Sigma*\nabla^{I_1}\mathrm{Ric}[G]*\cdots*\nabla^{I_{l-m-1}}\mathrm{Ric}[G]*\nabla^{I_\zeta+2}\zeta \\ +a^{-3}\sum_{m=0}^{l-2}\sum_{I_N+I_\Sigma+I_\zeta+\mathcal{I}=2m}\nabla^{I_N}(N+1)*\nabla^{I_\Sigma}\Sigma*\nabla^{I_1+1}\mathrm{Ric}[G]*\cdots*\nabla^{I_{l-m-1}}\mathrm{Ric}[G]*\nabla^{I_\zeta+1}\zeta \\ +\frac{\dot a}{a}\sum_{m=0}^{l-1}\sum_{\substack{I_N+\mathcal{I}+I_\zeta=2m-1\\ I_\zeta\neq 2(l-1)}}\nabla^{I_N}N*\nabla^{I_1}\mathrm{Ric}[G]*\cdots*\nabla^{I_{l-m-1}}\mathrm{Ric}[G]*\nabla^{I_\zeta+1}\zeta, \quad \text{(A-9e)}$$

$$\begin{aligned}\mathfrak{J}([\partial_t,\nabla\Delta^l]\zeta)=&\frac{\dot a}{a}\sum_{\substack{I_N+I_\zeta=2l\\ I_\zeta\neq 0}}\nabla^{I_N}N*\nabla^{I_\zeta+1}\zeta+a^{-3}\sum_{\substack{I_N+I_\Sigma+I_\zeta=2(l-1)+1\\ (I_\Sigma,I_\zeta)\neq(0,2(l-1)+1),(1,2(l-1))}}\nabla^{I_N}(N+1)*\nabla^{I_\Sigma}\Sigma*\nabla^{I_\zeta+1}\zeta\\&+a^{-3}\sum_{m=0}^{l-2}\sum_{I_N+I_\Sigma+I_\zeta+\mathcal{I}=2m+1}\nabla^{I_N}(N+1)*\nabla^{I_\Sigma}\Sigma*\nabla^{I_1}\mathrm{Ric}[G]*\cdots*\nabla^{I_{l-m-1}}\mathrm{Ric}[G]*\nabla^{I_\zeta+2}\zeta\\&+a^{-3}\sum_{m=0}^{l-2}\sum_{\substack{I_N+I_\Sigma+I_\zeta+\mathcal{I}=2m+1\\ I_1\neq 2l-3}}\nabla^{I_N}(N+1)*\nabla^{I_\Sigma}\Sigma*\nabla^{I_1+1}\mathrm{Ric}[G]*\cdots*\nabla^{I_{l-m-1}}\mathrm{Ric}[G]*\nabla^{I_\zeta+1}\zeta\\&+\frac{\dot a}{a}\sum_{m=0}^{l-1}\sum_{\substack{I_N+\mathcal{I}+I_\zeta=2m\\ I_\zeta\neq 2(l-1)}}\nabla^{I_N}N*\nabla^{I_1}\mathrm{Ric}[G]*\cdots*\nabla^{I_{l-m-1}}\mathrm{Ric}[G]*\nabla^{I_\zeta+1}\zeta,\end{aligned}\tag{A-9f}$$

$$\begin{aligned}\mathfrak{J}([\partial_t,\Delta^l])\mathfrak{T}=&a^{-3}\sum_{I_N+I_\Sigma+I_{\mathfrak{T}}=2l}\nabla^{I_N}N*\nabla^{I_\Sigma}\Sigma*\nabla^{I_{\mathfrak{T}}}\mathfrak{T}+a^{-3}\sum_{\substack{I_\Sigma+I_{\mathfrak{T}}=2l\\ I_\Sigma,I_{\mathfrak{T}}\geq 2}}\nabla^{I_\Sigma}\Sigma*\nabla^{I_{\mathfrak{T}}}\mathfrak{T}\\&+a^{-3}\sum_{m=0}^{l-1}\sum_{\substack{I_N+I_\Sigma+I_{\mathfrak{T}}+\mathcal{I}=2m\\ I_1<2l-3}}\nabla^{I_N}(N+1)*\nabla^{I_\Sigma}\Sigma*\nabla^{I_1}\mathrm{Ric}[G]*\cdots*\nabla^{I_{l-m-1}}\mathrm{Ric}[G]*\nabla^{I_{\mathfrak{T}}}\mathfrak{T}\\&+\frac{\dot a}{a}\sum_{\substack{I_N+I_{\mathfrak{T}}=2l\\ I_{\mathfrak{T}}\geq 2}}\nabla^{I_N}N*\nabla^{I_{\mathfrak{T}}}\mathfrak{T},\end{aligned}\tag{A-9g}$$

$$\mathfrak{J}([\partial_t,\nabla\Delta^l]\mathfrak{T})=a^{-3}\Sigma*\nabla N*\Delta^l\mathrm{Ric}[G]+a^{-3}N*\nabla\Sigma*\Delta^l\mathfrak{T}+\frac{\dot a}{a}\nabla N*\Delta^l\mathfrak{T}+\nabla\mathfrak{J}([\partial_t,\Delta^l]\mathfrak{T}).\tag{A-9h}$$

We can extend the formulas to $l=1$ by dropping any term which would contain negative powers of Δ or a multiindex of negative order.

Proof. This follows by iteratively applying the commutators in Lemma A.6. □

While all of the above commutators will be essential for the mainline argument, the a priori estimates require the following commutators:

Lemma A.8 (auxiliary commutators). *Let $J\in\mathbb{N}$. Then, we have*

$$[\partial_t,\nabla^J]\zeta=a^{-3}\sum_{\substack{I_N+I_\Sigma+I_\zeta=l-1\\ I_\zeta<J-1}}\nabla^{I_N}(N+1)*\nabla^{I_\Sigma}\Sigma*\nabla^{I_\zeta+1}\zeta+\frac{\dot a}{a}\sum_{\substack{I_N+I_\zeta=J-1\\ I_N>0}}\nabla^{I_N}N*\nabla^{I_\zeta+1}\zeta,\tag{A-10a}$$

$$[\partial_t,\nabla^J]\mathfrak{T}=a^{-3}\sum_{\substack{I_N+I_\Sigma+I_{\mathfrak{T}}=J\\ I_{\mathfrak{T}}<J}}\nabla^{I_N}(N+1)*\nabla^{I_\Sigma}\Sigma*\nabla^{I_{\mathfrak{T}}}\mathfrak{T}+\frac{\dot a}{a}\sum_{\substack{I_N+I_{\mathfrak{T}}=J\\ I_N>0}}\nabla^{I_N}N*\nabla^{I_{\mathfrak{T}}}\mathfrak{T}.\tag{A-10b}$$

Proof. For $J=1$, this has already been shown in (A-8a) and (A-8d). For higher orders, the formulas follow from a straightforward induction argument using that, in local coordinates, we schematically have

$$[\partial_t,\nabla^J]\zeta=[\partial_t,\nabla]\nabla^{J-1}\zeta+\nabla[\partial_t,\nabla^{J-1}]\zeta=(\partial_t\Gamma[G])*\nabla^{J-1}\zeta+\nabla[\partial_t,\nabla^{J-1}]\zeta$$

and analogously replacing ζ with $\mathfrak{T}$. □

A.3. *Borderline and junk terms.*

Definition A.9 (error terms). Let $L\in 2\mathbb{N}$, $L\geq 2$. Then, the error terms in the Laplace-commuted equations stated in Lemma 2.11 take the following form:

For the constraint equations, we have

$$\mathfrak{M}_{L,\mathrm{Junk}} = -8\pi(\Psi + C)\nabla\Delta^{L/2-2}\,\mathrm{Ric}[G] * \nabla^2\phi + \nabla^{L-2}\,\mathrm{Ric}[G] * \nabla\Sigma + \underbrace{\nabla^{L-3}\,\mathrm{Ric}[G] * \nabla^2\Sigma}_{\text{if } L\neq 2} + \sum_{\substack{I_\Psi + I_\phi = L \\ I_\Psi \neq 0}} \nabla^{I_\Psi}\Psi * \nabla^{I_\phi+1}\phi + 8\pi(\Psi + C)\mathfrak{J}([\Delta^{L/2}, \nabla]\phi) - \mathfrak{J}([\Delta^{L/2}, \mathrm{div}_G]\Sigma), \tag{A-11a}$$

$$\widetilde{\mathfrak{M}}_{L,\mathrm{Junk}} = \underbrace{-\boldsymbol{\varepsilon}[G] * \nabla^{L-3}\,\mathrm{Ric}[G] * \nabla\Sigma}_{\text{if } L\neq 2} - \mathfrak{J}([\Delta^{L/2}, \mathrm{curl}_G]\Sigma), \tag{A-11b}$$

$$\mathfrak{H}_{L,\mathrm{Border}} = a^{-4}[\Sigma * \Delta^{L/2}\Sigma + \nabla\Sigma * \nabla^{L-1}\Sigma], \tag{A-11c}$$

$$\mathfrak{H}_{L,\mathrm{Junk}} = \sum_{I_1+I_2=L} \nabla^{I_1+1}\phi * \nabla^{I_2+1}\phi + a^{-4} \sum_{\substack{I_1+I_2=L \\ I_i\geq 2}} \nabla^{I_1}\Sigma * \nabla^{I_2}\Sigma + \Delta^{L/2}\big[\tfrac{4\pi}{3}|\nabla\phi|_G^2 + \tfrac{8\pi}{3}a^{-4}\Psi^2 + \tfrac{16\pi}{3}Ca^{-4}\Psi\big]\cdot G. \tag{A-11d}$$

The lapse equation error terms are

$$\mathfrak{N}_{L,\mathrm{Border}} = a^{-4}(N+1)\big(\Sigma * \Delta^{L/2}\Sigma + \nabla\Sigma * \nabla^{L-1}\Sigma + \Psi * \Delta^{L/2}\Psi + \nabla\Psi * \nabla^{L-1}\Psi\big) + a^{-4}[|\Sigma|_G^2 + \Psi^2 + \Psi] * \Delta^{L/2}N + a^{-4}\nabla[|\Sigma|_G^2 + \Psi^2 + \Psi] * \nabla^{L-1}N, \tag{A-12a}$$

$$\mathfrak{N}_{L,\mathrm{Junk}} = a^{-4} \sum_{\substack{I_N+I_1+I_2=L; \\ I_N\leq L-2;\, I_N>0 \text{ or } I_1\leq I_2\leq L-2}} \nabla^{I_N}(N+1) * (\nabla^{I_1}\Sigma * \nabla^{I_2}\Sigma + \nabla^{I_1}\Psi * \nabla^{I_2}\Psi) + a^{-4}N * \Delta^{L/2}\Psi + a^{-4} \sum_{\substack{I_N+I_\Psi=L \\ I_\Psi\geq 2,\, I_N\geq 1}} \nabla^{I_N}N * \nabla^{I_\Psi}\Psi, \tag{A-12b}$$

as well as

$$\mathfrak{N}_{L+1,\mathrm{Border}} = a^{-4}(N+1)\big(\Sigma * \nabla\Delta^{L/2}\Sigma + \nabla\Sigma * \nabla^{L}\Sigma + \nabla^2\Sigma * \nabla^{L-1}\Sigma + \Psi * \nabla\Delta^{L/2}\Psi + \nabla\Psi * \nabla^{L}\Psi + \nabla^2\Psi * \nabla^{L-1}\Psi\big) + a^{-4}[|\Sigma|_G^2 + \Psi^2 + \Psi] * \nabla\Delta^{L/2}N + \nabla\Psi * \nabla^{L}N + \nabla^2\Psi * \nabla^{L-1}\Psi, \tag{A-12c}$$

$$\mathfrak{N}_{L+1,\mathrm{Junk}} = a^{-4} \sum_{\substack{I_N+I_1+I_2=L; \\ I_N<L+1;\, I_N>0 \text{ or } I_1\geq I_2>2}} \nabla^{I_N}(N+1) * (\nabla^{I_1}\Sigma * \nabla^{I_2}\Sigma + \nabla^{I_1}\Psi * \nabla^{I_2}\Psi) + a^{-4}N * \nabla\Delta^{L/2}\Psi + a^{-4} \sum_{\substack{I_N+I_\Psi=L+1 \\ I_N, I_\Psi>2}} \nabla^{I_N}N * \nabla^{I_\Psi}\Psi, \tag{A-12d}$$

whereas the scalar field error terms read

$$\mathfrak{P}_{L,\mathrm{Border}} = -3\Psi\frac{\dot a}{a}\Delta^{L/2}N + \frac{\dot a}{a}\nabla\Psi * \nabla^{L-1}N + 2a^{-3}(N+1)\langle\Sigma, \nabla^2\Delta^{L/2-1}\Psi\rangle_G + 2a^{-3}(N+1)\nabla^{L-3}\,\mathrm{Ric} * \Sigma * \nabla\Psi - 2a^{-3}(N+1)\langle\mathrm{div}_G\,\Delta^{L/2-1}\Sigma, \nabla\Psi\rangle_G + a^{-3}(N+1)\nabla\Sigma * \nabla^3\Delta^{L/2-2}\Psi, \tag{A-13a}$$

$$\mathfrak{P}_{L,\mathrm{Junk}} = \frac{\dot a}{a} \sum_{\substack{I_N+I_\Psi=L \\ I_\Psi\geq 2}} \nabla^{I_N}N * \nabla^{I_\Psi}\Psi + a \sum_{\substack{I_N+I_\phi=L+1 \\ I_N, I_\phi\neq 0}} \nabla^{I_N}N * \nabla^{I_\phi+1}\phi + \mathfrak{J}([\partial_t, \Delta^{L/2}]\Psi), \tag{A-13b}$$

$$\mathfrak{Q}_{L,\mathrm{Border}} = a^{-3}\Psi\nabla\Delta^{L/2}N + a^{-3}(N+1)\Sigma * \nabla^3\Delta^{L/2-1}\phi + a^{-3}(N+1)\nabla^{L}\phi * \nabla\Sigma, \tag{A-13c}$$

$$\mathfrak{Q}_{L,\mathrm{Junk}} = a^{-3} \sum_{\substack{I_N+I_\Psi=L+1 \\ I_N, I_\Psi \neq 0}} \nabla^{I_N} N * \nabla^{I_\Psi} \Psi + a^{-3} \nabla \Delta^{L/2-1} N * \nabla^2 \phi * \Sigma + a^{-3}(N+1) \nabla^2 \Delta^{L/2-1} \Sigma * \nabla \phi$$
$$+ (N+1) a^{-3} \nabla \Delta^{L/2-1} \operatorname{Ric}[G] * \Sigma * \nabla \phi + a^{-3} \nabla^{L-2} \operatorname{Ric}[G] * ((N+1) * \Sigma * \nabla \phi)$$
$$+ \frac{\dot{a}}{a} \langle \nabla^2 \Delta^{L/2-1} N, \nabla \phi \rangle_G + \mathfrak{J}([\partial_t, \nabla \Delta^{L/2}] \phi) \tag{A-13d}$$

and

$$\mathfrak{P}_{L+1,\mathrm{Border}} = -3\Psi \frac{\dot{a}}{a} \nabla \Delta^{L/2} N + \frac{\dot{a}}{a} \nabla \Psi * \nabla^2 \Delta^{L/2-1} N + 2a^{-3} \langle \Sigma, \nabla^3 \Delta^{L/2-1} \Psi \rangle_G$$
$$+ a^{-3}(N+1) \nabla \Sigma * \nabla^L \Psi + 2a^{-3} \nabla^{L-2} \mathrm{Ric} * \Sigma * \nabla \Psi + a^{-3}(N+1) \nabla^2 \Delta^{L/2-1} \Sigma * \nabla \Psi, \tag{A-13e}$$

$$\mathfrak{P}_{L+1,\mathrm{Junk}} = \frac{\dot{a}}{a} \sum_{\substack{I_N+I_\Psi=L+1 \\ I_\Psi \geq 2}} \nabla^{I_N} N * \nabla^{I_\Psi} \Psi + a \sum_{\substack{I_N+I_\phi=L+2 \\ I_N, I_\phi \neq 0}} \nabla^{I_N} N * \nabla^{I_\phi+1} \phi + \mathfrak{J}([\partial_t, \nabla \Delta^{L/2}] \Psi), \tag{A-13f}$$

$$\mathfrak{Q}_{L+1,\mathrm{Border}} = a^{-3} \Psi \Delta^{L/2+1} N + a^{-3} \nabla \Psi * \nabla \Delta^{L/2} N + a^{-3}(N+1) \Sigma * \nabla^2 \Delta^{L/2} \phi$$
$$+ \frac{\dot{a}}{a} \nabla \Delta^{L/2} N * \nabla \phi + a^{-3}(N+1) \nabla \Sigma * \nabla^2 \Delta^{L/2-1} \phi, \tag{A-13g}$$

$$\mathfrak{Q}_{L+1,\mathrm{Junk}} = a^{-3} \sum_{\substack{I_N+I_\Psi=L+2 \\ 2 \leq I_\Psi \leq L+1}} \nabla^{I_N} N * \nabla^{I_\Psi} \Psi + a^{-3}(N+1) \nabla^{L-2} \mathrm{Ric}[G] * \Sigma * \nabla \phi$$
$$+ a^{-3}(N+1) \nabla^2 \Delta^{L/2-1} \Sigma * \nabla \phi + \mathfrak{J}([\partial_t, \Delta^{L/2+1}] \phi), \tag{A-13h}$$

as well as

$$\mathfrak{Q}_{1,\mathrm{Border}} = a^{-3} \Psi \Delta N + a^{-3}(N+1) \Sigma * \nabla^2 \phi, \tag{A-13i}$$
$$\mathfrak{Q}_{1,\mathrm{Junk}} = a^{-3} \nabla \Psi * \nabla N + a^{-3}(N+1) \nabla \Sigma * \nabla \phi + \mathfrak{J}([\partial_t, \Delta] \phi). \tag{A-13j}$$

The commuted rescaled evolution equation for Σ has the error terms

$$\mathfrak{S}_{L,\mathrm{Border}} = a^{-3}(N+1)(\Sigma * \nabla^2 \Delta^{L/2-1} \Sigma + \nabla \Sigma * \nabla^3 \Delta^{L/2-2} \Sigma)$$
$$+ a^{-3}(\Delta^{L/2} N \cdot (\Sigma * \Sigma) + \nabla \Delta^{L/2-1} N * \nabla \Sigma * \Sigma)$$
$$+ a^{-3}(N+1) \Sigma * \Sigma * \nabla^2 \Delta^{L/2-2} \operatorname{Ric}[G] + \frac{\dot{a}}{a} \Delta^{L/2} N * \Sigma + \frac{\dot{a}}{a} \nabla \Delta^{L/2-1} N * \nabla \Sigma$$
$$+ \underbrace{a^{-3}[(N+1) \nabla \Sigma * \Sigma + \nabla N * \Sigma * \Sigma] * \nabla^{L-3} \operatorname{Ric}[G]}_{\text{not present for } L=2}, \tag{A-14a}$$

$$\mathfrak{S}_{L,\mathrm{Junk}} = -a[\Delta^{L/2}, \nabla^2] N + a \sum_{\substack{I_N+I_{\mathrm{Ric}}=L \\ I_N \neq 0}} \nabla^{I_N} N * \nabla^{I_{\mathrm{Ric}}} \operatorname{Ric}[G]$$
$$+ \frac{\dot{a}}{a} \sum_{I_N+I_\Sigma=L} \nabla^{I_N} N * \nabla^{I_\Sigma} \Sigma + a^{-3} \sum_{\substack{I_1+I_2=L \\ I_i>0}} \nabla^{I_1} \Sigma * \nabla^{I_2} \Sigma$$
$$+ a^{-3} \sum_{\substack{I_N+I_1+I_2=L \\ I_N<L}} \nabla^{I_N} N * \nabla^{I_1} \Sigma * \nabla^{I_2} \Sigma + a \sum_{I_N+I_1+I_2=L} \nabla^{I_N} (N+1) * \nabla^{I_1+1} \phi * \nabla^{I_2+1} \phi$$
$$+ \left(4\pi C^2 a^{-3} + \tfrac{1}{3} a\right) \Delta^{L/2} N \cdot G + \mathfrak{J}([\partial_t, \Delta^{L/2}] \Sigma), \tag{A-14b}$$

while the commuted Ricci tensor evolution equations have error terms, where $\mathcal{I} = I_N + I_\Sigma + \sum_{i=1}^{L/2-m+1} I_i$,

$$\mathfrak{R}_{L,\mathrm{Border}} = a^{-3}[\nabla^{L+2}N\cdot\Sigma + \nabla^{L+1}N * \nabla\Sigma + \Sigma * \nabla^2\Delta^{L/2-1}\,\mathrm{Ric}[G] + \nabla\Sigma * \nabla^{L-1}\,\mathrm{Ric}[G]], \tag{A-15a}$$

$$\mathfrak{R}_{L+1,\mathrm{Border}} = a^{-3}[\nabla^{L+3}N\cdot\Sigma + \nabla^{L+2}N * \nabla\Sigma + \Sigma * \nabla^3\Delta^{L/2-1}\,\mathrm{Ric}[G] + \nabla\Sigma * \nabla^{L}\,\mathrm{Ric}[G]], \tag{A-15b}$$

$$\begin{aligned}\mathfrak{R}_{L,\mathrm{Junk}} = {} & a^{-3}\sum_{\substack{I_N+I_\Sigma=L+2\\ I_\Sigma\geq 2}} \nabla^{I_N}N * \nabla^{I_\Sigma}\Sigma \\ & + a^{-3}\sum_{\substack{I_N+I_\Sigma+I_{\mathrm{Ric}}=L\\ (I_\Sigma,I_{\mathrm{Ric}})\neq(0,L),(1,L-1)}} \nabla^{I_N}(N+1) * \nabla^{I_\Sigma}\Sigma * \nabla^{I_{\mathrm{Ric}}}\,\mathrm{Ric}[G] \\ & + a^{-3}\sum_{m=0}^{L/2-1}\sum_{\mathcal{I}=2m} \nabla^{I_N}(N+1) * \nabla^{I_\Sigma}\Sigma * \nabla^{I_1}\,\mathrm{Ric}[G] * \cdots * \nabla^{I_{L/2-m+1}}\,\mathrm{Ric}[G] \\ & + \frac{\dot a}{a}\big([\Delta^{L/2},\nabla^2]N + \Delta^{L/2}N * \mathrm{Ric}[G] + \nabla^{L-1}N * \nabla\,\mathrm{Ric}[G]\big) \\ & + \mathfrak{J}([\partial_t,\Delta^{L/2}]\,\mathrm{Ric}[G]),\end{aligned} \tag{A-15c}$$

$$\begin{aligned}\mathfrak{R}_{L+1,\mathrm{Junk}} = {} & a^{-3}\sum_{\substack{I_N+I_\Sigma=L+3\\ I_\Sigma\geq 2}} \nabla^{I_N}N * \nabla^{I_\Sigma}\Sigma \\ & + a^{-3}\sum_{\substack{I_N+I_\Sigma+I_{\mathrm{Ric}}=L\\ (I_\Sigma,I_{\mathrm{Ric}})\neq(0,L+1),(1,L)}} \nabla^{I_N}(N+1) * \nabla^{I_\Sigma}\Sigma * \nabla^{I_{\mathrm{Ric}}}\,\mathrm{Ric}[G] \\ & + a^{-3}\sum_{m=0}^{L/2-1}\sum_{\mathcal{I}=2m+1} \nabla^{I_N}(N+1) * \nabla^{I_\Sigma}\Sigma * \nabla^{I_1}\,\mathrm{Ric}[G] * \cdots \nabla^{I_{L/2-m+1}}\,\mathrm{Ric}[G] \\ & + \frac{\dot a}{a}\big(\nabla[\Delta^{L/2},\nabla^2]N + \nabla\Delta^{L/2}N * \mathrm{Ric}[G] + \nabla^2\Delta^{L/2-1}N * \nabla\,\mathrm{Ric}[G]\big) \\ & + \mathfrak{J}([\partial_t,\nabla\Delta^{L/2}]\,\mathrm{Ric}[G]).\end{aligned} \tag{A-15d}$$

Finally, the Bel–Robinson evolution error terms are

$$\begin{aligned}\mathfrak{E}_{L,\mathrm{Border}} = {} & \frac{\tau}{3}(\Delta^{L/2}N\cdot\boldsymbol{E} + \nabla^{L-1}N * \nabla\boldsymbol{E}) - a^{-1}(\Delta^{L/2}\boldsymbol{E}\times\Sigma + \boldsymbol{E}\times\Delta^{L/2}\Sigma) \\ & + a^{-3}\boldsymbol{\varepsilon}[G] * \boldsymbol{\varepsilon}[G] * (\nabla^{L-1}\boldsymbol{E} * \nabla\Sigma + \nabla\boldsymbol{E} * \nabla^{L-1}\Sigma) \\ & + a^{-3}\Delta^{L/2}N\cdot(\boldsymbol{E} * \Sigma) + a^{-3}\nabla\Delta^{L/2-1}N * [\nabla\boldsymbol{E} * \Sigma + \boldsymbol{E} * \nabla\Sigma] \\ & + a^{-3}(\Sigma * \nabla^2\Delta^{L/2-1}\boldsymbol{E} + \nabla\Sigma * \nabla^3\Delta^{L/2-2}\boldsymbol{E} + \nabla\boldsymbol{E} * \nabla\Delta^{L/2-1}\Sigma + \boldsymbol{E} * \Delta^{L/2}\Sigma) \\ & + a^{-3}\big[(N+1)\Sigma * \boldsymbol{E} * \nabla^2\Delta^{L/2-2}\,\mathrm{Ric}[G] + \underbrace{\nabla((N+1) * \Sigma * \boldsymbol{E}) * \nabla^{L-3}\,\mathrm{Ric}[G]}_{\text{if } L\neq 2}\big] \\ & + 4\pi a^{-3}(\Psi+C)^2\Delta^{L/2}N\cdot\Sigma + 4\pi a^{-3}\nabla^{L-1}N * [(\Psi+C)^2\nabla\Sigma + 2(\Psi+C) * \nabla\Psi * \Sigma] \\ & + 4\pi a^{-3}(N+1)[(\Psi^2+2C\Psi)\Delta^{L/2}\Sigma + 2(\Psi+C)\Delta^{L/2}\Psi\cdot\Sigma] \\ & + 4\pi a^{-3}\nabla^{L-1}\Sigma * [(\Psi+C)^2\nabla N + 2(N+1)(\Psi+C)\nabla\Psi] \\ & + 4\pi a^{-3}(\Psi+C)\nabla^{L-1}\Psi * [(N+1)\nabla\Sigma + \nabla N * \Sigma],\end{aligned} \tag{A-16a}$$

$$\mathfrak{E}_{L,\mathrm{top}} = a^{-1}(N+1)\boldsymbol{\varepsilon}[G] * \boldsymbol{B} * \nabla\Delta^{L/2-1}\,\mathrm{Ric}[G] + a(N+1)(\Psi+C)\nabla\Delta^{L/2-1}\,\mathrm{Ric}[G] * \nabla\phi, \tag{A-16b}$$

$$\mathfrak{E}_{L,\mathrm{Junk}} = \frac{\dot a}{a}\sum_{\substack{I_N+I_{\boldsymbol E}=L\\ I_N\le L-2}} \nabla^{I_N}N*\nabla^{I_{\boldsymbol E}}\boldsymbol E + a^{-1}\boldsymbol\varepsilon[G]*\Big[\sum_{\substack{I_N+I_{\boldsymbol B}=L+1\\ I_N,I_{\boldsymbol B}\le L}}\nabla^{I_N}N*\nabla^{I_{\boldsymbol B}}\boldsymbol B + (N+1)\nabla^2\Delta^{L/2-2}\mathrm{Ric}[G]*\nabla\boldsymbol B\Big]$$
$$+a^{-3}\boldsymbol\varepsilon[G]*\boldsymbol\varepsilon[G]*\sum_{\substack{I_N+I_{\boldsymbol E}+I_\Sigma=L\\ I_N\le L-2;\ I_N>0 \text{ or } I_{\boldsymbol E},I_\Sigma\ge 2}}\nabla^{I_N}N*\nabla^{I_{\boldsymbol E}}\boldsymbol E*\nabla^{I_\Sigma}\Sigma$$
$$+a\sum_{\substack{I_N+I_\Psi+I_\phi=L+1\\ I_N,I_\Psi,I_\phi\ne L+1}}\nabla^{I_N}(N+1)*\nabla^{I_\Psi}(\Psi+C)*\nabla^{I_\phi+1}\phi$$
$$+a^{-3}\sum_{\substack{I_N+I_\Sigma+I_1+I_2=L\\ I_N,I_\Sigma,I_i\le L-2}}\nabla^{I_N}(N+1)*\nabla^{I_\Sigma}\Sigma*\nabla^{I_1}(\Psi+C)*\nabla^{I_2}(\Psi+C)$$
$$+\dot a a^3\sum_{I_N+I_1+I_2=L}\nabla^{I_N}(N+1)*\nabla^{I_1+1}\phi*\nabla^{I_2+1}\phi$$
$$+a\sum_{I_N+I_\Sigma+I_1+I_2=L}\nabla^{I_N}(N+1)*\nabla^{I_\Sigma}\Sigma*\nabla^{I_1+1}\phi*\nabla^{I_2+1}\phi$$
$$+a^{-1}\varepsilon[G]*\boldsymbol B*[\Delta^{L/2},\nabla]N+4\pi a(N+1)(\Psi+C)\big[\nabla^{L-2}\mathrm{Ric}[G]*\nabla^2\phi+\mathfrak J([\Delta^{L/2},\nabla^2]\phi)\big]$$
$$+a\{(\Psi+C)[\Delta^{L/2},\nabla]N+(N+1)[\Delta^{L/2},\nabla]\Psi\}*\nabla\phi$$
$$+(N+1)a^{-1}\mathfrak J([\Delta^{L/2},\mathrm{curl}_G]\boldsymbol B)+\mathfrak J([\partial_t,\Delta^{L/2}]\boldsymbol E)$$
$$+\Delta^{L/2}\Big[a^{-3}(N+1)\boldsymbol E*\Sigma$$
$$+\frac{2\pi}{3}a^6(N+1)\Big(\partial_0\cdot(a^{-6}(\Psi+C)^2+a^{-2}|\nabla\phi|_G^2)+4\pi\frac{\dot a}{a}(\Psi+C)^2\Big)\Big]\cdot G, \quad \text{(A-16c)}$$

$$\mathfrak B_{L,\mathrm{Border}} = \frac{\tau}{3}(\Delta^{L/2}N\cdot\boldsymbol B+\nabla^{L-1}N*\nabla\boldsymbol B)-a^{-1}(\Delta^{L/2}\boldsymbol B\times\Sigma+\boldsymbol B*\Delta^{L/2}\Sigma)$$
$$+a^{-3}\boldsymbol\varepsilon[G]*\boldsymbol\varepsilon[G]*(\nabla^{L-1}\boldsymbol B*\nabla\Sigma+\nabla\boldsymbol B*\nabla^{L-1}\Sigma)$$
$$+a^{-3}\Delta^{L/2}N\cdot(\boldsymbol B*\Sigma)+a^{-3}\nabla\Delta^{L/2-1}N\cdot[\nabla\boldsymbol B*\Sigma+\boldsymbol B*\nabla\Sigma]$$
$$+a^{-3}\big(\Sigma*\nabla^2\Delta^{L/2-1}\boldsymbol B+\nabla\Sigma*\nabla^3\Delta^{L/2-2}\boldsymbol B+\nabla\boldsymbol B*\nabla\Delta^{L/2-1}\Sigma+\boldsymbol B*\Delta^{L/2}\Sigma\big)$$
$$+a^{-3}[(N+1)\Sigma*\boldsymbol B*\nabla^{L-2}\mathrm{Ric}[G]+\underbrace{\nabla((N+1)*\Sigma*\boldsymbol B)*\nabla^{L-3}\mathrm{Ric}[G]]}_{\text{if } L\ne 2}$$
$$+a^{-1}(N+1)(\Psi+C)\cdot\boldsymbol\varepsilon[G]*\nabla\Delta^{L/2}\phi*\Sigma$$
$$+a^{-1}\boldsymbol\varepsilon[G]*\nabla^2\nabla^L\phi*\nabla((N+1)(\Psi+C)\Sigma), \quad \text{(A-16d)}$$

$$\mathfrak B_{L,\mathrm{top}} = a^3(N+1)\boldsymbol\varepsilon[G]*\nabla\Delta^{L/2-1}\mathrm{Ric}[G]*\nabla\phi*\nabla\phi+a^{-1}\boldsymbol\varepsilon[G]*\boldsymbol E*\nabla\Delta^{L/2-1}\mathrm{Ric}[G], \quad \text{(A-16e)}$$

$$\mathfrak B_{L,\mathrm{Junk}} = \frac{\dot a}{a}\sum_{\substack{I_N+I_{\boldsymbol B}=L\\ I_N\le L-2}}\nabla^{I_N}N*\nabla^{I_{\boldsymbol B}}\boldsymbol B+a^{-1}\boldsymbol\varepsilon[G]*\Big[\sum_{\substack{I_N+I_{\boldsymbol E}=L+1\\ I_N,I_{\boldsymbol E}\le L}}\nabla^{I_N}N*\nabla^{I_{\boldsymbol E}}\boldsymbol E+\nabla^2\Delta^{L/2-2}\mathrm{Ric}[G]*\nabla\boldsymbol E\Big]$$
$$+a^{-3}\boldsymbol\varepsilon[G]*\boldsymbol\varepsilon[G]*\sum_{\substack{I_N+I_{\boldsymbol E}+I_\Sigma=L,\\ I_N\le L-2;\ I_N>0 \text{ or } I_{\boldsymbol B},I_\Sigma\ge 2}}\nabla^{I_N}(N+1)*\nabla^{I_{\boldsymbol B}}\boldsymbol B*\nabla^{I_\Sigma}\Sigma$$
$$+a^3\boldsymbol\varepsilon[G]*\sum_{\substack{I_N+I_1+I_2=L\\ I_N>0 \text{ or } I_2<L}}\nabla^{I_N}(N+1)*\nabla^{I_1+1}\phi*\nabla^{I_2+2}\phi$$
$$+a^{-1}\boldsymbol\varepsilon[G]*\sum_{\substack{I_N+I_\Psi+I_\phi+I_\Sigma=L\\ I_\phi\le L-2}}\nabla^{I_N}(N+1)*\nabla^{I_\Psi}(\Psi+C)*\nabla^{I_\phi+1}\phi*\nabla^{I_\Sigma}\Sigma$$

$$
\begin{aligned}
&+a^{-1}\varepsilon[G]*\boldsymbol{E}*[\Delta^{L/2},\nabla]N+a^{-1}(N+1)(\Psi+C)\cdot\boldsymbol{\varepsilon}[G]*\Sigma*[\Delta^{L/2},\nabla]\phi\\
&+a^{3}(N+1)\boldsymbol{\varepsilon}[G]*\nabla\phi*\big(\nabla^2\Delta^{L/2-2}\mathrm{Ric}[G]*\nabla^2\phi+\mathfrak{J}([\Delta^{L/2},\nabla]\phi)\big)\\
&-a^{-1}(N+1)\mathfrak{J}([\Delta^{L/2},\mathrm{curl}_G]\boldsymbol{E})+\mathfrak{J}([\partial_t,\Delta^{L/2}]\boldsymbol{B})+\Delta^{L/2}[a^{-3}(N+1)\boldsymbol{B}*\Sigma]\cdot G\\
&+\Delta^{L/2}\Big[4\pi a^2\nabla^{\sharp m}\phi(\Psi+C)+\frac{2\pi}{3}a^5\Delta^{L/2}\nabla^{\sharp m}(a^{-6}(\Psi+C)^2+a^{-2}|\nabla\phi|_G^2)\Big]\boldsymbol{\varepsilon}[G]_{(\cdot)m(\cdot)}.
\end{aligned}
\tag{A-16f}
$$

A.4. L^2_G *error term estimates.* In this subsection, we collect how the error terms can be controlled in terms of energies as well as homogeneous Sobolev norms of ϕ. We don't claim that these estimates are optimal — in particular, we note that at low order (like $L=2$), many of the curvature errors that appear in the estimates below could be avoided entirely: These arise as a result of applying the general estimates in Lemma 4.5 where the Ricci tensor doesn't naturally occur in the respective equations, and can be avoided at low orders by applying (4-4f) on all curvature terms that occur.

Instead of optimality, we try to keep both notation and form of the error term estimates as simple as possible and the energy estimates between base and top level as unified as possible. In particular, we track the "worst" curvature energy occurring at high orders for all estimates below, even if these terms are added in artificially for low orders.

Lemma A.10 (estimates for borderline error terms). *Let $L\in 2\mathbb{Z}_+$, $L\leq 20$. Then, the following estimates hold*:

$$
\begin{aligned}
\|\mathfrak{H}_{L,\mathrm{Border}}\|_{L^2_G}\lesssim{}&\varepsilon a^{-4}\sqrt{\mathcal{E}^{(L)}(\Sigma,\cdot)}\\
&+\varepsilon a^{-4-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\Sigma,\cdot)}+\varepsilon^2 a^{-4-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\mathrm{Ric},\cdot)},
\end{aligned}
\tag{A-17a}
$$

$$
\begin{aligned}
\|\mathfrak{N}_{L,\mathrm{Border}}\|_{L^2_G}\lesssim{}&\varepsilon a^{-4}\Big[\sqrt{\mathcal{E}^{(L)}(\phi,\cdot)}+\sqrt{\mathcal{E}^{(L)}(\Sigma,\cdot)}\Big]+\varepsilon a^{-4}\sqrt{\mathcal{E}^{(L)}(N,\cdot)}\\
&+\varepsilon a^{-4-c\sqrt{\varepsilon}}\Big[\sqrt{\mathcal{E}^{(\leq L-2)}(\phi,\cdot)}+\sqrt{\mathcal{E}^{(\leq L-2)}(\Sigma,\cdot)}+\sqrt{\mathcal{E}^{(\leq L-2)}(N,\cdot)}\Big]\\
&+\underbrace{\varepsilon^2 a^{-4}\sqrt{\mathcal{E}^{(\leq L-2)}(\mathrm{Ric},\cdot)}+\varepsilon a^{-4-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-4)}(\mathrm{Ric},\cdot)}}_{\textit{not present for } L=2},
\end{aligned}
\tag{A-17b}
$$

$$
\begin{aligned}
\|\mathfrak{N}_{L+1,\mathrm{Border}}\|_{L^2_G}\lesssim{}&\varepsilon a^{-6}\Big[\sqrt{a^4\mathcal{E}^{(L+1)}(\phi,\cdot)}+\sqrt{a^4\mathcal{E}^{(L+1)}(\Sigma,\cdot)}\Big]+\varepsilon a^{-6}\sqrt{a^4\mathcal{E}^{(L+1)}(N,\cdot)}\\
&+\varepsilon a^{-4}\Big[\sqrt{\mathcal{E}^{(L)}(\phi,\cdot)}+\sqrt{\mathcal{E}^{(L)}(\Sigma,\cdot)}+\sqrt{\mathcal{E}^{(L)}(N,\cdot)}\Big]\\
&+\varepsilon a^{-4-c\sqrt{\varepsilon}}\Big[\sqrt{\mathcal{E}^{(\leq L-2)}(\phi,\cdot)}+\sqrt{\mathcal{E}^{(\leq L-2)}(\Sigma,\cdot)}+\sqrt{\mathcal{E}^{(\leq L-2)}(N,\cdot)}\Big]\\
&+\varepsilon^2 a^{-4-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\mathrm{Ric},\cdot)},
\end{aligned}
\tag{A-17c}
$$

$$
\begin{aligned}
\|\mathfrak{P}_{L,\mathrm{Border}}\|_{L^2_G}\lesssim{}&\varepsilon a^{-3}\sqrt{\mathcal{E}^{(L)}(\phi,\cdot)}+\varepsilon a^{-3}\sqrt{\mathcal{E}^{(L)}(N,\cdot)}+\varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(N,\cdot)}\\
&+\varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\phi,\cdot)}+\varepsilon a^{-3}\sqrt{\mathcal{E}^{(L)}(\Sigma,\cdot)}+\varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(L-2)}(\Sigma,\cdot)}\\
&+\underbrace{\varepsilon^2 a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-3)}(\mathrm{Ric},\cdot)}}_{\textit{not present for } L=2},
\end{aligned}
\tag{A-17d}
$$

$$
\begin{aligned}
\|\mathfrak{Q}_{L,\mathrm{Border}}\|_{L^2_G}\lesssim{}&\varepsilon a^{-3}\sqrt{\mathcal{E}^{(L+1)}(N,\cdot)}\\
&+\varepsilon a^{-3}\sqrt{a^{-4}\mathcal{E}^{(L)}(\phi,\cdot)}+\varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{a^{-4}\mathcal{E}^{(\leq L-2)}(\phi,\cdot)},
\end{aligned}
\tag{A-17e}
$$

$$\begin{aligned}\|\mathfrak{P}_{L+1,\mathrm{Border}}\|_{L^2_G} \lesssim{}& \varepsilon a^{-3}\sqrt{\mathcal{E}^{(L+1)}(\phi,\cdot)} + \varepsilon a^{-3}\sqrt{\mathcal{E}^{(L+1)}(N,\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-1)}(N,\cdot)}\\ &+ \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-1)}(\phi,\cdot)} + \varepsilon a^{-3}\sqrt{\mathcal{E}^{(L+1)}(\Sigma,\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(L-1)}(\Sigma,\cdot)}\\ &+ \varepsilon^2 a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\mathrm{Ric},\cdot)},\end{aligned} \tag{A-17f}$$

$$\begin{aligned}\|\mathfrak{Q}_{L+1,\mathrm{Border}}\|_{L^2_G} \lesssim{}& \varepsilon a^{-3}\sqrt{\mathcal{E}^{(L+2)}(N,\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(L+1)}(N,\cdot)}\\ &+ \varepsilon a^{-3}\sqrt{a^{-4}\mathcal{E}^{(L+1)}(\phi,\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{a^{-4}\mathcal{E}^{(\leq L-1)}(\phi,\cdot)},\end{aligned} \tag{A-17g}$$

$$\begin{aligned}\|\mathfrak{S}_{L,\mathrm{Border}}\|_{L^2_G} \lesssim{}& \varepsilon a^{-3}\sqrt{\mathcal{E}^{(L)}(\Sigma,\cdot)} + \varepsilon a^{-3}\sqrt{\mathcal{E}^{(L)}(N,\cdot)} + \varepsilon^2 a^{-3}\sqrt{\mathcal{E}^{(L-2)}(\mathrm{Ric},\cdot)}\\ &+ \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\Sigma,\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(N,\cdot)}\\ &+ \underbrace{\varepsilon^2 a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-4)}(\mathrm{Ric},\cdot)}}_{\textit{not present for } L=2},\end{aligned} \tag{A-17h}$$

$$\begin{aligned}\|\mathfrak{R}_{L,\mathrm{Border}}\|_{L^2_G} \lesssim{}& \varepsilon a^{-3}\sqrt{\mathcal{E}^{(L+2)}(N,\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L)}(N,\cdot)}\\ &+ \varepsilon a^{-3}\sqrt{\mathcal{E}^{(L)}(\mathrm{Ric},\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\mathrm{Ric},\cdot)},\end{aligned} \tag{A-17i}$$

$$\begin{aligned}\|\mathfrak{R}_{L+1,\mathrm{Border}}\|_{L^2_G} \lesssim{}& \varepsilon a^{-3}\sqrt{\mathcal{E}^{(L+3)}(N,\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L+1)}(N,\cdot)}\\ &+ \varepsilon a^{-3}\sqrt{\mathcal{E}^{(L+1)}(\mathrm{Ric},\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-1)}(\mathrm{Ric},\cdot)},\end{aligned} \tag{A-17j}$$

$$\begin{aligned}\|\mathfrak{E}_{L,\mathrm{Border}}\|_{L^2_G} + \|\mathfrak{B}_{L,\mathrm{Border}}\|_{L^2_G} \lesssim{}& \varepsilon a^{-3}\sqrt{\mathcal{E}^{(L)}(\phi,\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\phi,\cdot)}\\ &+ \varepsilon a^{-3}\Big(\sqrt{\mathcal{E}^{(L)}(N,\cdot)} + \sqrt{\mathcal{E}^{(L)}(\Sigma,\cdot)}\Big) + \varepsilon a^{-3}\sqrt{\mathcal{E}^{(L)}(W,\cdot)}\\ &+ \varepsilon a^{-3-c\sqrt{\varepsilon}}\Big(\sqrt{\mathcal{E}^{(\leq L-2)}(N,\cdot)} + \sqrt{\mathcal{E}^{(\leq L-2)}(\Sigma,\cdot)} + \sqrt{\mathcal{E}^{(\leq L-2)}(W,\cdot)}\Big)\\ &+ \varepsilon^2 a^{-3}\sqrt{\mathcal{E}^{(L-2)}(\mathrm{Ric},\cdot)} + \underbrace{\varepsilon^2 a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-4)}(\mathrm{Ric},\cdot)}}_{\textit{not present for } L=2}.\end{aligned} \tag{A-17k}$$

Proof. All of these estimates follow from applying L^2_G-L^∞_G-type Hölder estimates to the individual nonlinear terms. The lower-order terms are either controlled by the zero order estimates in Section 4.1 or the a priori estimates in Lemma 4.3. Furthermore, we apply Lemma 4.5, along with again Lemma 4.3, to translate L^2_G-norms into energies up to additional curvature energy terms. For the sake of simplicity, we always estimate $\dot{a}/a$ by a^{-3} up to constant (see (2-3)), and liberally apply (3-8) to deal with odd order energies and to distribute $a^{-c\sqrt{\varepsilon}}$ factors to lower orders while updating $c > 0$ wherever this is convenient. □

Lemma A.11 (estimates for top-order error terms).

$$\|\mathfrak{E}_{L,\mathrm{top}}\|_{L^2_G} \lesssim \sqrt{\varepsilon} a^{1-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(L-1)}(\mathrm{Ric},\cdot)} = \sqrt{\varepsilon} a^{-1-c\sqrt{\varepsilon}}\sqrt{a^4\mathcal{E}^{(L-1)}(\mathrm{Ric},\cdot)}, \tag{A-18a}$$

$$\|\mathfrak{B}_{L,\mathrm{top}}\|_{L^2_G} \lesssim \varepsilon a^{-1}\sqrt{\mathcal{E}^{(L-1)}(\mathrm{Ric},\cdot)} = \varepsilon a^{-3}\sqrt{a^4\mathcal{E}^{(L-1)}(\mathrm{Ric},\cdot)}. \tag{A-18b}$$

Proof. This follows directly using (4-2c) and (4-4g) for the Bel–Robinson terms as well as (4-4e). □

Lemma A.12 (junk terms). *Recalling the* $\|$*-notation from Remark 2.12, the following hold*:

$$\begin{aligned}\|\mathfrak{M}_{L,\mathrm{Junk}}\|_{L^2_G} \lesssim{}& \varepsilon a^{-2-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-1)}(\phi,\cdot)} + a^{-2-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\phi,\cdot)}\\ &+ a^{-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-1)}(\Sigma,\cdot)} + \sqrt{\varepsilon} a^{-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\mathrm{Ric},\cdot)},\end{aligned} \tag{A-19a}$$

$$\|\widetilde{\mathfrak{M}}_{L,\mathrm{Junk}}\|_{L^2_G} \lesssim \varepsilon a^{-1-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\mathrm{Ric},\cdot)} + a^{-1-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-1)}(\Sigma,\cdot)}, \tag{A-19b}$$

$$\begin{aligned}\|\mathfrak{H}^{\|}_{L,\mathrm{Junk}}\|_{L^2_G} \lesssim{}& \varepsilon a^{-4-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\Sigma,\cdot)} + \sqrt{\varepsilon}a^{-2-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L)}(\phi,\cdot)}\\ &+ \varepsilon a^{-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\mathrm{Ric},\cdot)},\end{aligned} \tag{A-19c}$$

$$\begin{aligned}\|\mathfrak{N}_{L,\mathrm{Junk}}\|_{L^2_G} \lesssim{}& \varepsilon a^{-4-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(N,\cdot)} + \varepsilon a^{-c\sigma}\sqrt{\mathcal{E}^{(L)}(\phi,\cdot)}\\ &+ \varepsilon a^{-4-c\sqrt{\varepsilon}}\big[\sqrt{\mathcal{E}^{(\leq L-2)}(\phi,\cdot)} + \sqrt{\mathcal{E}^{(\leq L-2)}(\Sigma,\cdot)}\big]\\ &+ \underbrace{\varepsilon a^{-4}\sqrt{\mathcal{E}^{(\leq L-2)}(\mathrm{Ric},\cdot)} + \varepsilon a^{-4-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-4)}(\mathrm{Ric},\cdot)}}_{\textit{not present for } L=2},\end{aligned} \tag{A-19d}$$

$$\begin{aligned}\|\mathfrak{N}_{L+1,\mathrm{Junk}}\|_{L^2_G} \lesssim{}& \varepsilon a^{-4-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-1)}(N,\cdot)} + \varepsilon a^{-c\sigma}\big(\sqrt{\mathcal{E}^{(L+1)}(\phi,\cdot)} + \sqrt{\mathcal{E}^{(L+1)}(\Sigma,\cdot)}\big)\\ &+ \varepsilon a^{-4-c\sqrt{\varepsilon}}\big[\sqrt{\mathcal{E}^{(\leq L-1)}(\phi,\cdot)} + \sqrt{\mathcal{E}^{(\leq L-1)}(\Sigma,\cdot)}\big]\\ &+ \varepsilon^2 a^{-4}\sqrt{\mathcal{E}^{(\leq L-1)}(\mathrm{Ric},\cdot)} + \underbrace{\varepsilon^2 a^{-4-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-3)}(\mathrm{Ric},\cdot)}}_{\textit{not present for } L=2},\end{aligned} \tag{A-19e}$$

$$\begin{aligned}\|\mathfrak{P}_{L,\mathrm{Junk}}\|_{L^2_G} \lesssim{}& \varepsilon a^{1-c\sigma}\sqrt{\mathcal{E}^{(L)}(\phi,\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\phi,\cdot)}\\ &+ \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\Sigma,\cdot)} + \sqrt{\varepsilon}a^{1-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(L)}(N,\cdot)}\\ &+ \big[\varepsilon a^{-3-c\sqrt{\varepsilon}} + \sqrt{\varepsilon}a^{-1-c\sqrt{\varepsilon}}\big]\sqrt{\mathcal{E}^{(\leq L-2)}(N,\cdot)}\\ &+ \underbrace{\varepsilon^2 a^{1-c\sigma}\sqrt{\mathcal{E}^{(\leq L-2)}(\mathrm{Ric},\cdot)} + \varepsilon^2 a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-3)}(\mathrm{Ric},\cdot)}}_{\textit{not present for } L=2},\end{aligned} \tag{A-19f}$$

$$\begin{aligned}\|\mathfrak{P}_{L+1,\mathrm{Junk}}\|_{L^2_G} \lesssim{}& \varepsilon a^{1-c\sigma}\sqrt{\mathcal{E}^{(L+1)}(\phi,\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-1)}(\phi,\cdot)}\\ &+ \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-1)}(\Sigma,\cdot)} + \sqrt{\varepsilon}a^{1-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(L+1)}(N,\cdot)}\\ &+ \big[\varepsilon a^{-3-c\sqrt{\varepsilon}} + \sqrt{\varepsilon}a^{-1-c\sqrt{\varepsilon}}\big]\sqrt{\mathcal{E}^{(\leq L-1)}(N,\cdot)}\\ &+ \varepsilon^2 a^{-1-c\sigma}\sqrt{a^4\mathcal{E}^{(L-1)}(\mathrm{Ric},\cdot)}\\ &+ (\varepsilon^2 a^{-1-c\sigma} + \varepsilon^2 a^{-3-c\sqrt{\varepsilon}})\sqrt{\mathcal{E}^{(\leq L-2)}(\mathrm{Ric},\cdot)},\end{aligned} \tag{A-19g}$$

$$\begin{aligned}\|\mathfrak{Q}_{L,\mathrm{Junk}}\|_{L^2_G} \lesssim{}& \varepsilon a^{-1-c\sqrt{\varepsilon}}\sqrt{a^{-4}\mathcal{E}^{(L)}(\phi,\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\mathcal{E}^{(\leq L-2)}(\phi,\cdot)\\ &+ \sqrt{\varepsilon}a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L)}(\Sigma,\cdot)} + \sqrt{\varepsilon}a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L)}(N,\cdot)}\\ &+ \underbrace{\varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\mathrm{Ric},\cdot)}}_{\textit{not present for } L=2},\end{aligned} \tag{A-19h}$$

$$\|\mathfrak{Q}_{1,\mathrm{Junk}}\| \lesssim \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(1)}(N,\cdot)} + \varepsilon^{3/2}a^{-3-c\sqrt{\varepsilon}}\|\nabla\phi\|_{L^2_G} + \sqrt{\varepsilon}a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq 1)}(\Sigma,\cdot)}, \tag{A-19i}$$

$$\begin{aligned}\|\mathfrak{Q}_{L+1,\mathrm{Junk}}\|_{L^2_G} \lesssim{}& \varepsilon a^{-1-c\sqrt{\varepsilon}}\sqrt{a^{-4}\mathcal{E}^{(L+1)}(\phi,\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-1)}(\phi,\cdot)}\\ &+ \sqrt{\varepsilon}a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L+1)}(\Sigma,\cdot)} + \sqrt{\varepsilon}a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L+1)}(N,\cdot)}\\ &+ \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-1)}(\mathrm{Ric},\cdot)},\end{aligned} \tag{A-19j}$$

$$\begin{aligned}\|\mathfrak{S}^{\parallel}_{L,\mathrm{Junk}}\|_{L^2_G} &\lesssim \varepsilon a^{1-c\sigma}\sqrt{\mathcal{E}^{(L)}(\Sigma,\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\Sigma,\cdot)} + \sqrt{\varepsilon}a^{-1-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(L)}(\phi,\cdot)}\\ &\quad + (\varepsilon a^{-3} + a^{1-c\sqrt{\varepsilon}})\sqrt{\mathcal{E}^{(\leq L)}(N,\cdot)} + \varepsilon a^{5-c\sigma}\sqrt{\mathcal{E}^{(\leq L-1)}(\mathrm{Ric},\cdot)}\\ &\quad + \underbrace{\varepsilon a^{-3}\sqrt{\mathcal{E}^{(L-2)}(\mathrm{Ric},\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-4)}(\mathrm{Ric},\cdot)}}_{\textit{not present for } L=2},\end{aligned} \tag{A-19k}$$

$$\begin{aligned}\|\mathfrak{R}_{L,\mathrm{Junk}}\|_{L^2_G} &\lesssim \varepsilon^2 a^{1-c\sigma}\sqrt{\mathcal{E}^{(\leq L-1)}(\mathrm{Ric},\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\mathrm{Ric},\cdot)}\\ &\quad + \varepsilon a^{1-c\sigma}\sqrt{\mathcal{E}^{(\leq L+2)}(\Sigma,\cdot)} + a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L)}(\Sigma,\cdot)}\\ &\quad + a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L)}(N,\cdot)},\end{aligned} \tag{A-19l}$$

$$\begin{aligned}\|\mathfrak{R}_{L+1,\mathrm{Junk}}\|_{L^2_G} &\lesssim \varepsilon^2 a^{1-c\sigma}\sqrt{\mathcal{E}^{(\leq L)}(\mathrm{Ric},\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-1)}(\mathrm{Ric},\cdot)}\\ &\quad + \varepsilon a^{1-c\sigma}\sqrt{\mathcal{E}^{(\leq L+3)}(\Sigma,\cdot)} + a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L+1)}(\Sigma,\cdot)}\\ &\quad + a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L+1)}(N,\cdot)},\end{aligned} \tag{A-19m}$$

$$\begin{aligned}\|\mathfrak{E}^{\parallel}_{L,\mathrm{Junk}}\|_{L^2_G} + \|\mathfrak{B}^{\parallel}_{L,\mathrm{Junk}}\|_{L^2_G} &\lesssim \varepsilon a^{-1-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L)}(W,\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(W,\cdot)}\\ &\quad + \varepsilon a^{-1-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(L)}(\phi,\cdot)} + (\varepsilon a^{-3-c\sqrt{\varepsilon}} + a^{-1-c\sqrt{\varepsilon}})\sqrt{\mathcal{E}^{(\leq L-2)}(\phi,\cdot)}\\ &\quad + \sqrt{\varepsilon}a^{1-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L)}(N,\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(N,\cdot)}\\ &\quad + \varepsilon a^{-1-c\sigma}\sqrt{\mathcal{E}^{(L)}(\Sigma,\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\sqrt{\mathcal{E}^{(\leq L-2)}(\Sigma,\cdot)}\\ &\quad + \varepsilon a^{-3}\sqrt{\mathcal{E}^{(L-2)}(\mathrm{Ric},\cdot)} + \varepsilon a^{-3-c\sqrt{\varepsilon}}\underbrace{\sqrt{\mathcal{E}^{(\leq L-4)}(\mathrm{Ric},\cdot)}}_{\textit{not present for } L=2}.\end{aligned} \tag{A-19n}$$

Proof. Once again, this follows by applying the a priori estimates from Section 4.1 and Lemma 4.3, as well as the bootstrap assumption (3-17h) for the lapse, to deal with the lower-order terms in the nonlinearities, and then applying Lemma 4.5, as well as (3-8), wherever this is necessary. Further, especially in (A-19n), it is often more convenient to use the bootstrap assumption for $\|\nabla\phi\|_{C_G}$ instead of the a priori estimate (4-4e) to gain higher powers of ε in prefactors.

Recognizing that every low-order curvature term can be estimated up to constant by $a^{-c\sqrt{\varepsilon}}$ at worst (see (4-4f)), we also note that any of the highly nonlinear curvature terms in $\mathfrak{J}$-expressions turn out to be negligible after updating c compared to Ricci energies arising from applying Lemma 4.5 or compared to junk terms in which $\mathrm{Ric}[G]$ is tracked explicitly. □

Appendix B: Future stability

Here, we collect the commutators in CMCSH gauge necessary to study the commuted scalar-field equations:

Lemma B.1 (commutator formulas for future stabilty). *Let ζ be a scalar function on Σ_T. Then, the following formulas hold*:

$$\begin{aligned}[\tilde{\partial}_0, \nabla]\zeta &= 0,\\ [\tilde{\partial}_0, \Delta_{\boldsymbol{g}}]\zeta &= (\tilde{\partial}_0(\boldsymbol{g}^{-1})^{ab})\nabla_a\nabla_b\zeta - 2(\boldsymbol{g}^{-1})^{ab}(\mathrm{div}_{\boldsymbol{g}}(\boldsymbol{n}\boldsymbol{\Sigma})_a - 2\nabla_a\boldsymbol{n})\nabla_b\zeta.\end{aligned}$$

Schematically, for $k \in \mathbb{N}$, this implies

$$[\tilde{\partial}_0, \Delta_{\boldsymbol{g}}^k]\zeta = \sum_{I_{\boldsymbol{n}}+I_{\boldsymbol{\Sigma}}+I_\zeta=2k-1} \nabla^{I_{\boldsymbol{n}}}\boldsymbol{n} *_{\boldsymbol{g}} \nabla^{I_{\boldsymbol{\Sigma}}}\boldsymbol{\Sigma} *_{\boldsymbol{g}} \nabla^{I_\zeta+1}\zeta + \sum_{I_{\hat{\boldsymbol{n}}}+I_\zeta=2k-1} \nabla^{I_{\hat{\boldsymbol{n}}}}\hat{\boldsymbol{n}} *_{\boldsymbol{g}} \nabla^{I_\zeta+1}\zeta, \tag{B-1a}$$

$$[\tilde{\partial}_0, \nabla\Delta_{\boldsymbol{g}}^k]\zeta = \sum_{I_{\boldsymbol{n}}+I_{\boldsymbol{\Sigma}}+I_\zeta=2k} \nabla^{I_{\boldsymbol{n}}}\boldsymbol{n} *_{\boldsymbol{g}} \nabla^{I_{\boldsymbol{\Sigma}}}\boldsymbol{\Sigma} *_{\boldsymbol{g}} \nabla^{I_\zeta+1}\zeta + \sum_{I_{\hat{\boldsymbol{n}}}+I_\zeta=2k} \nabla^{I_{\hat{\boldsymbol{n}}}}\hat{\boldsymbol{n}} *_{\boldsymbol{g}} \nabla^{I_\zeta+1}\zeta. \tag{B-1b}$$

Proof. This follows from straightfoward computations, similar to Lemma A.6 for the low-order commutators and to Lemma A.7 for higher orders. □

Acknowledgements

This research was funded in whole or in part by the Austrian Science Fund (FWF) 10.55776/Y963 and 10.55776/P34313. For open access purposes, the author has applied a CC BY public copyright license to any author-accepted manuscript version arising from this submission. Liam Urban is a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Faculty of Mathematics at the University of Vienna. Urban also thanks the German Academic Scholarship Foundation (Studienstiftung des deutschen Volkes) for their scholarship. The authors thank Ian Agol, Klaus Kröncke, Michael Lipnowski, Dalimil Mazac and Roman Prosanov for their help in seeking out numerical and analytic evidence for the spectral condition used in Section 9, and Michael Eichmair and the anonymous referees for their detailed, constructive and warm feedback on previous versions of this manuscript. The authors would like to thank the Erwin Schrödinger International Institute for Mathematics and Physics in Vienna for hosting the authors during the Thematic Programs "Mathematical Perspectives of Gravitation beyond the Vacuum Regime", "Spectral Theory and Mathematical Relativity" and "Nonlinear Waves and General Relativity" during which research for this work was done and parts of this paper were written.

References

[Alho et al. 2019] A. Alho, G. Fournodavlos, and A. T. Franzen, "The wave equation near flat Friedmann–Lemaître–Robertson–Walker and Kasner big bang singularities", *J. Hyperbolic Differ. Equ.* **16**:2 (2019), 379–400. MR Zbl

[Allen and Rendall 2010] P. T. Allen and A. D. Rendall, "Asymptotics of linearized cosmological perturbations", *J. Hyperbolic Differ. Equ.* **7**:2 (2010), 255–277. MR Zbl

[Andersson and Fajman 2020] L. Andersson and D. Fajman, "Nonlinear stability of the Milne model with matter", *Comm. Math. Phys.* **378**:1 (2020), 261–298. MR Zbl

[Andersson and Moncrief 2003] L. Andersson and V. Moncrief, "Elliptic-hyperbolic systems and the Einstein equations", *Ann. Henri Poincaré* **4**:1 (2003), 1–34. MR Zbl

[Andersson and Moncrief 2004] L. Andersson and V. Moncrief, "Future complete vacuum spacetimes", pp. 299–330 in *The Einstein equations and the large scale behavior of gravitational fields*, edited by P. T. Chruściel and H. Friedrich, Birkhäuser, Basel, 2004. MR Zbl

[Andersson and Moncrief 2011] L. Andersson and V. Moncrief, "Einstein spaces as attractors for the Einstein flow", *J. Differential Geom.* **89**:1 (2011), 1–47. MR Zbl

[Andersson and Rendall 2001] L. Andersson and A. D. Rendall, "Quiescent cosmological singularities", *Comm. Math. Phys.* **218**:3 (2001), 479–511. MR Zbl

[Andersson et al. 1997] L. Andersson, V. Moncrief, and A. J. Tromba, "On the global evolution problem in $2+1$ gravity", *J. Geom. Phys.* **23**:3-4 (1997), 191–205. MR Zbl

[Bachelot 2019] A. Bachelot, "Wave asymptotics at a cosmological time-singularity: classical and quantum scalar fields", *Comm. Math. Phys.* **369**:3 (2019), 973–1020. MR Zbl

[Barrow 1978] J. D. Barrow, "Quiescent cosmology", *Nature* **272** (1978), 211–215.

[Barzegar and Fajman 2022] H. Barzegar and D. Fajman, "Stable cosmologies with collisionless charged matter", *J. Hyperbolic Differ. Equ.* **19**:4 (2022), 587–634. MR Zbl

[Belinskiĭ and Khalatnikov 1973] V. A. Belinskiĭ and I. M. Khalatnikov, "Effect of scalar and vector fields on the nature of the cosmological singularity", *Soviet Phys. JETP* **36**:4 (1973), 591–597. MR

[Bergeron 2003] N. Bergeron, "Lefschetz properties for arithmetic real and complex hyperbolic manifolds", *Int. Math. Res. Not.* **2003**:20 (2003), 1089–1122. MR Zbl

[Beyer and Oliynyk 2024a] F. Beyer and T. A. Oliynyk, "Localized big bang stability for the Einstein-scalar field equations", *Arch. Ration. Mech. Anal.* **248**:1 (2024), art. id. 3. MR Zbl

[Beyer and Oliynyk 2024b] F. Beyer and T. A. Oliynyk, "Relativistic perfect fluids near Kasner singularities", *Comm. Anal. Geom.* **32**:6 (2024), 1701–1794. MR Zbl

[Bonifacio et al. 2025] J. Bonifacio, D. Mazáč, and S. Pal, "Spectral bounds on hyperbolic 3-manifolds: associativity and the trace formula", *Comm. Math. Phys.* **406**:3 (2025), art. id. 51. MR Zbl

[Booker and Strömbergsson 2007] A. R. Booker and A. Strömbergsson, "Numerical computations with the trace formula and the Selberg eigenvalue conjecture", *J. Reine Angew. Math.* **607** (2007), 113–161. MR Zbl

[Branding et al. 2019] V. Branding, D. Fajman, and K. Kröncke, "Stable cosmological Kaluza–Klein spacetimes", *Comm. Math. Phys.* **368**:3 (2019), 1087–1120. MR Zbl

[Callahan 1994] P. J. Callahan, *Spectral geometry of hyperbolic* 3*-manifolds*, Ph.D. thesis, University of Illinois at Urbana-Champaign, 1994, available at https://www.proquest.com/docview/304119964.

[Cheng and Zhou 1995] X. Cheng and D. T. Zhou, "First eigenvalue estimate on Riemannian manifolds", *Hokkaido Math. J.* **24**:3 (1995), 453–472. MR Zbl

[Choquet-Bruhat and Cotsakis 2002] Y. Choquet-Bruhat and S. Cotsakis, "Global hyperbolicity and completeness", *J. Geom. Phys.* **43**:4 (2002), 345–350. MR Zbl

[Choquet-Bruhat and Geroch 1969] Y. Choquet-Bruhat and R. Geroch, "Global aspects of the Cauchy problem in general relativity", *Comm. Math. Phys.* **14** (1969), 329–335. MR Zbl

[Choquet-Bruhat and Moncrief 2001] Y. Choquet-Bruhat and V. Moncrief, "Future global in time Einsteinian spacetimes with U(1) isometry group", *Ann. Henri Poincaré* **2**:6 (2001), 1007–1064. MR Zbl

[Choquet-Bruhat et al. 2004] Y. Choquet-Bruhat, J. Isenberg, and V. Moncrief, "Topologically general U(1) symmetric vacuum space-times with AVTD behavior", *Nuovo Cimento Soc. Ital. Fis. B* **119**:7-9 (2004), 625–638. MR

[Chow et al. 2006] B. Chow, P. Lu, and L. Ni, *Hamilton's Ricci flow*, Grad. Stud. in Math. **77**, Amer. Math. Soc., Providence, RI, 2006. MR Zbl

[Christodoulou and Klainerman 1990] D. Christodoulou and S. Klainerman, "Asymptotic properties of linear field equations in Minkowski space", *Comm. Pure Appl. Math.* **43**:2 (1990), 137–199. MR Zbl

[Christodoulou and Klainerman 1993] D. Christodoulou and S. Klainerman, *The global nonlinear stability of the Minkowski space*, Princeton Math. Ser. **41**, Princeton Univ. Press, 1993. MR Zbl

[Chruściel and Isenberg 1993] P. T. Chruściel and J. Isenberg, "Nonisometric vacuum extensions of vacuum maximal globally hyperbolic spacetimes", *Phys. Rev. D* (3) **48**:4 (1993), 1616–1628. MR

[Chruściel and Rendall 1995] P. T. Chruściel and A. D. Rendall, "Strong cosmic censorship in vacuum space-times with compact, locally homogeneous Cauchy surfaces", *Ann. Physics* **242**:2 (1995), 349–385. MR Zbl

[Cicoli et al. 2024] M. Cicoli, J. P. Conlon, A. Maharana, S. Parameswaran, F. Quevedo, and I. Zavala, "String cosmology: from the early universe to today", *Phys. Rep.* **1059** (2024), 1–155. MR Zbl

[Cornish and Spergel 1999] N. J. Cornish and D. N. Spergel, "On the eigenmodes of compact hyperbolic 3-manifolds", preprint, 1999. arXiv math/9906017

[Damour et al. 2002] T. Damour, M. Henneaux, A. D. Rendall, and M. Weaver, "Kasner-like behaviour for subcritical Einstein-matter systems", *Ann. Henri Poincaré* **3**:6 (2002), 1049–1111. MR Zbl

[Dragomir 2003] S. S. Dragomir, *Some Gronwall type inequalities and applications*, Nova Sci., Hauppauge, NY, 2003. MR Zbl

[Fajman 2017] D. Fajman, "The nonvacuum Einstein flow on surfaces of negative curvature and nonlinear stability", *Comm. Math. Phys.* **353**:2 (2017), 905–961. MR Zbl

[Fajman 2020] D. Fajman, "Future attractors in $2+1$ dimensional Λ gravity", *Phys. Rev. Lett.* **125**:12 (2020), art. id. 121102. MR

[Fajman and Kröncke 2020] D. Fajman and K. Kröncke, "Stable fixed points of the Einstein flow with positive cosmological constant", *Comm. Anal. Geom.* **28**:7 (2020), 1533–1576. MR Zbl

[Fajman and Urban 2022] D. Fajman and L. Urban, "Blow-up of waves on singular spacetimes with generic spatial metrics", *Lett. Math. Phys.* **112**:2 (2022), art. id. 42. MR Zbl

[Fajman and Wyatt 2021] D. Fajman and Z. Wyatt, "Attractors of the Einstein–Klein–Gordon system", *Comm. Partial Differential Equations* **46**:1 (2021), 1–30. MR Zbl

[Fajman et al. 2024] D. Fajman, M. Ofner, and Z. Wyatt, "Slowly expanding stable dust spacetimes", *Arch. Ration. Mech. Anal.* **248**:5 (2024), art. id. 83. MR Zbl

[Fourès-Bruhat 1952] Y. Fourès-Bruhat, "Théorème d'existence pour certains systèmes d'équations aux dérivées partielles non linéaires", *Acta Math.* **88** (1952), 141–225. MR Zbl

[Fournodavlos and Luk 2023] G. Fournodavlos and J. Luk, "Asymptotically Kasner-like singularities", *Amer. J. Math.* **145**:4 (2023), 1183–1272. MR Zbl

[Fournodavlos et al. 2023] G. Fournodavlos, I. Rodnianski, and J. Speck, "Stable big bang formation for Einstein's equations: the complete sub-critical regime", *J. Amer. Math. Soc.* **36**:3 (2023), 827–916. MR Zbl

[Girão et al. 2019] P. M. Girão, J. Natário, and J. D. Silva, "Solutions of the wave equation bounded at the big bang", *Classical Quantum Gravity* **36**:7 (2019), art. id. 075016. MR Zbl

[Hawking 1967] S. W. Hawking, "The occurrence of singularities in cosmology, III: Causality and singularities", *Proc. Roy. Soc. London Ser. A* **300**:1461 (1967), 187–201. Zbl

[Inoue 2001] K. T. Inoue, "Numerical study of length spectra and low-lying eigenvalue spectra of compact hyperbolic 3-manifolds", *Classical Quantum Gravity* **18**:4 (2001), 629–652. MR Zbl

[Isenberg and Moncrief 2002] J. Isenberg and V. Moncrief, "Asymptotic behaviour in polarized and half-polarized U(1) symmetric vacuum spacetimes", *Classical Quantum Gravity* **19**:21 (2002), 5361–5386. MR Zbl

[Kröncke 2015] K. Kröncke, "On the stability of Einstein manifolds", *Ann. Global Anal. Geom.* **47**:1 (2015), 81–98. MR Zbl

[Lin and Lipnowski 2022] F. Lin and M. Lipnowski, "The Seiberg–Witten equations and the length spectrum of hyperbolic three-manifolds", *J. Amer. Math. Soc.* **35**:1 (2022), 233–293. MR Zbl

[Lin and Lipnowski 2024] F. Lin and M. Lipnowski, "Closed geodesics and Frøyshov invariants of hyperbolic three-manifolds", *J. Eur. Math. Soc.* (online publication April 2024).

[Moncrief 2008] V. Moncrief, "Relativistic Teichmüller theory: a Hamilton–Jacobi approach to $2+1$-dimensional Einstein gravity", pp. 203–249 in *Geometric flows*, Surv. Differ. Geom. **12**, Int. Press, Somerville, MA, 2008. MR Zbl

[Mondal 2023] P. Mondal, "Big-bang limit of $2+1$ gravity and Thurston boundary of Teichmüller space", *J. Math. Phys.* **64**:11 (2023), art. id. 112501. MR Zbl

[O'Neill 1983] B. O'Neill, *Semi-Riemannian geometry*: *with applications to relativity*, Pure Appl. Math. **103**, Academic Press, New York, 1983. MR Zbl

[Oude Groeniger et al. 2023] H. Oude Groeniger, O. Petersen, and H. Ringström, "Formation of quiescent big bang singularities", preprint, 2023. arXiv 2309.11370

[Penrose 1965] R. Penrose, "Gravitational collapse and space-time singularities", *Phys. Rev. Lett.* **14** (1965), 57–59. MR Zbl

[Rendall 2008] A. D. Rendall, *Partial differential equations in general relativity*, Oxford Grad. Texts in Math. **16**, Oxford Univ. Press, 2008. MR Zbl

[Ringström 2008] H. Ringström, "Future stability of the Einstein-non-linear scalar field system", *Invent. Math.* **173**:1 (2008), 123–208. MR Zbl

[Ringström 2009] H. Ringström, *The Cauchy problem in general relativity*, Eur. Math. Soc., Zürich, 2009. MR Zbl

[Ringström 2019] H. Ringström, "A unified approach to the Klein–Gordon equation on Bianchi backgrounds", *Comm. Math. Phys.* **372**:2 (2019), 599–656. MR Zbl

[Ringström 2020] H. Ringström, *Linear systems of wave equations on cosmological backgrounds with convergent asymptotics*, Astérisque **420**, Soc. Math. France, Paris, 2020. MR Zbl

[Ringström 2021] H. Ringström, "Wave equations on silent big bang backgrounds", preprint, 2021. To appear in *Mem. Amer. Math. Soc.* arXiv 2101.04939

[Rodnianski and Speck 2018a] I. Rodnianski and J. Speck, "A regime of linear stability for the Einstein-scalar field system with applications to nonlinear big bang formation", *Ann. of Math.* (2) **187**:1 (2018), 65–156. MR Zbl

[Rodnianski and Speck 2018b] I. Rodnianski and J. Speck, "Stable big bang formation in near-FLRW solutions to the Einstein-scalar field and Einstein-stiff fluid systems", *Selecta Math.* (*N.S.*) **24**:5 (2018), 4293–4459. MR Zbl

[Rodnianski and Speck 2022] I. Rodnianski and J. Speck, "On the nature of Hawking's incompleteness for the Einstein-vacuum equations: the regime of moderately spatially anisotropic initial data", *J. Eur. Math. Soc.* **24**:1 (2022), 167–263. MR Zbl

[Speck 2018] J. Speck, "The maximal development of near-FLRW data for the Einstein-scalar field system with spatial topology $\mathbb{S}^3$", *Comm. Math. Phys.* **364**:3 (2018), 879–979. MR Zbl

[Wang 2019] J. Wang, "Future stability of the $1+3$ Milne model for the Einstein–Klein–Gordon system", *Classical Quantum Gravity* **36**:22 (2019), art. id. 225010. MR Zbl

[Wang 2021] J. Wang, "Nonlinear wave equation in a cosmological Kaluza Klein spacetime", *J. Math. Phys.* **62**:6 (2021), art. id. 062504. MR Zbl

Received 14 Dec 2022. Revised 19 Jul 2023. Accepted 3 Jun 2024.

DAVID FAJMAN: david.fajman@univie.ac.at
Faculty of Physics, University of Vienna, Vienna, Austria

LIAM URBAN: liam.urban@univie.ac.at
Faculty of Mathematics, University of Vienna, Vienna, Austria

ANALYSIS AND PDE
Vol. 18 (2025), No. 7, pp. 1715–1768
DOI: 10.2140/apde.2025.18.1715

msp

SPECTRAL ESTIMATES FOR FREE BOUNDARY MINIMAL SURFACES VIA MONTIEL–ROS PARTITIONING METHODS

ALESSANDRO CARLOTTO, MARIO B. SCHULZ AND DAVID WIYGUL

We adapt and extend the Montiel–Ros methodology to compact manifolds with boundary, allowing for mixed (including oblique) boundary conditions and also accounting for the action of a finite group G together with an additional twisting homomorphism $\sigma : G \to \mathrm{O}(1)$. We then apply this machinery in order to obtain quantitative lower and upper bounds on the growth rate of the Morse index of free boundary minimal surfaces with respect to the topological data (i.e., the genus and the number of boundary components) of the surfaces in question. In particular, we compute the exact values of the equivariant Morse index and nullity for two infinite families of examples, with respect to their maximal symmetry groups, and thereby derive explicit two-sided linear bounds when the equivariance constraint is lifted.

1. Introduction

Despite a profusion of constructions of free boundary minimal surfaces in the Euclidean unit ball $\mathbb{B}^3$ over the course of the past decade — [Fraser and Schoen 2011; 2016; Girouard and Lagacé 2021; Karpukhin et al. 2014] via optimization of the first Steklov eigenvalue, [Carlotto et al. 2022a; Ketover 2016a; 2016b] via min-max methods for the area functional, and [Carlotto et al. 2022b; Folha et al. 2017; Kapouleas and Li 2021; Kapouleas and McGrath 2023; Kapouleas and Wiygul 2023; Kapouleas and Zou 2021] via gluing methods — many basic questions about the space of such surfaces remain open. The reader is referred to [Franz 2022; Fraser 2020; Li 2020] for recent overviews of the field. In particular, so far it is only for the rotationally symmetric examples, planar discs through the origin and critical catenoids, that the exact value of the Morse index is actually known; see [Devyver 2019; Smith and Zhou 2019; Tran 2020]. The present manuscript is the first in a series of works aimed at shedding new light on this fundamental invariant, which (also due to its variational content, and thus to its natural connection with min-max theory, see [Marques and Neves 2016; 2018; 2020]) has acquired great importance within geometric analysis.

MSC2020: primary 53A10; secondary 49Q05, 58C40.
Keywords: minimal surfaces, Morse index, equivariant spectrum.

Partly motivated by the corresponding conjectures concerning closed minimal hypersurfaces in manifolds of positive Ricci curvature (see [Ambrozio et al. 2018a; Neves 2014]), five years ago the first author proved with Ambrozio and Sharp a universal lower bound for the index of any free boundary minimal surface in any mean-convex subdomain Ω of $\mathbb{R}^3$ in terms of the topological data of the surface under consideration. Specifically, it was shown in [Ambrozio et al. 2018b] that the following estimate holds:

$$\operatorname{index}(\Sigma) \geq \tfrac{1}{3}(2g + b - 1), \tag{1-1}$$

where Σ is any free boundary minimal surface in Ω and g and b denote its genus and the number of its boundary components, respectively. This result was then partly complemented by [Lima 2022, Theorem 4], that is, an affine upper bound with a very large, yet in principle computable, numerical constant. In this article we shall develop a general methodology, building upon the fundamental work [Montiel and Ros 1991], which allows us, among other things, to significantly refine such universal estimates bringing the geometry and symmetry group of the surfaces under consideration into play. This approach, while motivated by our goal to better understand the behavior of certain infinite families of free boundary minimal surfaces in $\mathbb{B}^3$ (aiming for two-sided bounds in terms of explicit, affine functions of the topological data), turns out to be of independent interest and much wider applicability.

In more abstract terms, we shall be concerned here with proving effective estimates for (part of) the spectrum of Schrödinger-type operators on bounded Lipschitz domains of Riemannian manifolds, combined with mixed boundary conditions, which will be — on disjoint portions of the boundary in question — of Dirichlet or Robin (oblique) type. Summarizing and oversimplifying things to the extreme, the number of eigenvalues of any such operator *below a given threshold* can be estimated by suitably partitioning the domain into finitely many subdomains, provided one adjoins Dirichlet boundary conditions in the interior boundaries when aiming for lower bounds, and Neumann boundary conditions in the interior boundaries for upper bounds instead. We refer the reader to Section 2 for the setup of our problem together with our standing assumptions, and to the first part of Section 3 (specifically to Proposition 3.1, and Corollary 3.2) for precise statements.

Often times (yet not always) the partitions mentioned above naturally relate to the underlying symmetries of the problem in question, which is in particular the case for some of the classes of free boundary minimal surfaces in $\mathbb{B}^3$ that have so far been constructed. With this remark in mind, a peculiar (and, a posteriori, fundamental) feature of our work is the development of the Montiel–Ros methodology in the presence of the action of a group G together with an additional twisting homomorphism $\sigma : G \to \mathrm{O}(1)$, in the terms explained in Section 2.4. This allows us, for instance, to explicitly and transparently study how the Morse index of a given free boundary minimal surface depends on the symmetries one imposes, namely to look at the "functor" $(G, \sigma) \to \operatorname{ind}_G^\sigma(T)$, where T denotes the index (Jacobi) form of the surface in question. As apparent even from the simplest examples we shall discuss, this perspective turns out to be very natural and effective in tackling the geometric problems we are interested in.

With this approach, *lower* bounds are sometimes relatively cheap to obtain. One way they can be derived is from ambient Killing vector fields, once it is shown that the associated (scalar-valued) Jacobi field on the surface under consideration vanishes along the (interior) boundary of any domain of the

chosen partition, which in practice amounts to suitably *designing the partition and picking the Killing field* given the geometry of the problem. We present one such simple yet paradigmatic result in Proposition 4.2, which concerns free boundary minimal surfaces with pyramidal or prismatic symmetry in $\mathbb{B}^3$. Instead, *upper* bounds are often a lot harder to obtain and shall typically rely on finer information than the sole symmetries of the scene one deals with. Said otherwise, one needs to know *how* (i.e., by which method) the surface under study has been obtained.

We will develop here a detailed analysis of the Morse index of the two families of free boundary minimal surfaces we constructed in our recent, previous work [Carlotto et al. 2022b]. Very briefly, using gluing methods of essentially PDE-theoretic character, we obtained there a sequence $\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}$ of surfaces having genus m, three boundary components and antiprismatic symmetry group $\mathbb{A}_{m+1}$, and a sequence $\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}$ of surfaces having genus zero, $n+2$ boundary components and prismatic symmetry group $\mathbb{P}_n$. As we described at length in Section 7 therein, with data (see Tables 2 and 3 in [Carlotto et al. 2022b]) and heuristics, numerical simulations for the Morse index of the surfaces in the former sequence display a seemingly "erratic" behavior, as such values do not align on the graph of any affine function, nor seem to exhibit any obvious periodic pattern. This is a rather unexpected behavior (by comparison, e.g., with other families of examples, say in the round three-dimensional sphere, see [Kapouleas and Wiygul 2020]), which obviously calls for a careful study that we carry through in Section 5 of the present article. In particular, we establish the following statement.

Theorem 1.1 (index estimates for $\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}$ and $\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}$). *There exist $m_0, n_0 > 0$ such that, for all integers $m > m_0$ and $n > n_0$, the Morse index and nullity of the free boundary minimal surfaces $\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}$, $\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0} \subset \mathbb{B}^3$ satisfy the bounds*

$$\begin{aligned} 2m+1 \le \operatorname{ind}(\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}), \quad \operatorname{ind}(\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}) + \operatorname{nul}(\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}) \le 12m+12, \\ 2n+2 \le \operatorname{ind}(\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}), \qquad \operatorname{ind}(\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}) + \operatorname{nul}(\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}) \le 8n. \end{aligned}$$

In fact, the upper bound in this "absolute estimate" follows quite easily by combining the "relative estimate" associated to the equivariant Morse index of these surfaces (with respect to their respective *maximal* symmetry groups) with the aforementioned Proposition 3.1. The next statement thus pertains to such equivariant bounds for which we do obtain equality, thus settling part of Conjectures 7.7 (iv) and 7.9 (iv) of [Carlotto et al. 2022b]. We stress that neither family is constructed variationally, and thus there is actually no cheap index bound one can extract from the design methodology itself; on the contrary, this statement indicates a posteriori that the families of surfaces in question may in principle be constructed (even in a nonasymptotic regime) by means of min-max schemes generated by 2-parameter sweepouts, modulo the well-known problem of fully controlling the topology in the process (see [Carlotto et al. 2022a]).

Theorem 1.2 (equivariant index and nullity of $\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}$ and $\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}$). *There exist $m_0, n_0 > 0$ such that, for all integers $m > m_0$ and $n > n_0$, the equivariant Morse index and nullity of the free boundary minimal surfaces $\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}$, $\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0} \subset \mathbb{B}^3$ satisfy*

$$\begin{aligned} \operatorname{ind}_{\mathbb{A}_{m+1}}(\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}) = 2, \quad \operatorname{nul}_{\mathbb{A}_{m+1}}(\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}) = 0, \\ \operatorname{ind}_{\mathbb{P}_n}(\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}) = 2, \qquad \operatorname{nul}_{\mathbb{P}_n}(\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}) = 0. \end{aligned}$$

The main idea behind the proof of these results, or — more precisely — for the upper bounds, can only be explained by recalling, in a few words, how the surfaces in question have been constructed. Following the general methodology of [Kapouleas 1997], one first considers a singular configuration, that is a formal union of minimal surfaces in $\mathbb{B}^3$ (not necessarily free boundary), then its regularization — which needs the use of (wrapped) periodic minimal surfaces in $\mathbb{R}^3$, to desingularize near the divisors, and controlled interpolation processes between the building blocks in play — and, thirdly and finally, the perturbation of such configurations to exact minimality (at least for *some* values of the parameters), while also ensuring proper embeddedness and accommodating the free boundary condition. Here we first get a complete understanding of the index and nullities of the building blocks for the concrete cases under consideration in Section 5. In somewhat more detail, the analysis of the Karcher–Scherk towers (the periodic building blocks employed in either construction) exploits, in a substantial fashion, the use of the Gauss map, which allows one to rephrase the initial geometric question into one for the spectrum of simple elliptic operators of the form $\Delta_{g^{\mathbb{S}^2}} + 2$ on suitable (typically singular, i.e., spherical triangles, wedges or lunes) subdomains of round $\mathbb{S}^2$, with mixed boundary conditions, and possibly subject to additional symmetry requirements. The analysis of the other building blocks — disks and asymmetric catenoidal annuli — is more direct, although, in the latter case, trickier than it may first look (see, e.g., Lemma 5.8).

Once that preliminary analysis is done, we then prove that, corresponding to the (local) geometric convergence results (that are implied by the very gluing methodology), there are robust spectral convergence results that serve our scopes. However, a general challenge in the process is that gluing constructions typically have *transition regions* where different scales interact with each another: in our constructions of the sequences

$$\Sigma_m^{-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0} \quad \text{and} \quad \Xi_n^{-\mathbb{K}_0 \cup \mathbb{K}_0},$$

such regions occur between the catenoidal annuli $\mathbb{K}_0$ (as well as the disk $\mathbb{B}^2$ in the former case) and the wrapped Karcher–Scherk towers, roughly at distances between m^{-1} and $m^{-1/2}$ (respectively n^{-1} and $n^{-1/2}$) from the equatorial $\mathbb{S}^1$. As a result, we need to deal with delicate scale-picking arguments, an ad hoc study of the geometry of such regions (see Lemma 5.21) and — most importantly — prove the corresponding uniform bounds for eigenvalues and eigenfunctions (collected in Lemma 5.25), which allow us to rule out pathologic concentration phenomena, thereby leading to the desired conclusions.

2. Notation and standing assumptions

2.1. ***Boundary value problems for Schrödinger operators on Lipschitz domains.*** Let Ω be a Lipschitz domain of a smooth, compact d-dimensional manifold M with (possibly empty) boundary ∂M, by which we mean here a nonempty, open subset of M whose boundary is everywhere locally representable as the graph of a Lipschitz function. We do not require — at least in general — Ω to be connected, and we admit the case $\overline{\Omega} = M$ (where $\overline{\Omega}$ denotes the closure of Ω in M), when of course $\partial\Omega = \partial M$, the boundary of the ambient manifold in question. Throughout this article we will in fact assume $d \geq 2$.

We are going to study the spectrum of a given Schrödinger operator on Ω subject to boundary conditions and, sometimes, symmetry constraints. Such symmetry constraints will be encoded in terms of equivariance with respect to a certain group action, which we shall specify in due time.

The Schrödinger operator

$$\Delta_g + q$$

is determined by the data of a given smooth Riemannian metric g on $\overline{\Omega}$ and a given smooth (i.e., C^∞) function $q : \overline{\Omega} \to \mathbb{R}$. To avoid ambiguities, we remark here that a function (or tensor field) on $\overline{\Omega}$ is smooth if it is the restriction of a smooth tensor field on M or — equivalently — on a relatively open set containing $\overline{\Omega}$.

The boundary conditions are specified by another smooth function $r : \overline{\Omega} \to \mathbb{R}$ and a decomposition

$$\partial\Omega = \overline{\partial_{\mathrm{D}}\Omega} \cup \overline{\partial_{\mathrm{N}}\Omega} \cup \overline{\partial_{\mathrm{R}}\Omega}, \tag{2-1}$$

where the sets on the right-hand side are the closures of pairwise disjoint open subsets $\partial_{\mathrm{D}}\Omega$, $\partial_{\mathrm{N}}\Omega$, and $\partial_{\mathrm{R}}\Omega$ of $\partial\Omega$.

Somewhat more specifically, we will consider the spectrum of the operator $\Delta_g + q$ subject to the Dirichlet, Neumann, and Robin conditions

$$\begin{cases} u = 0 & \text{on } \partial_{\mathrm{D}}\Omega, \\ du(\eta_g^\Omega) = 0 & \text{on } \partial_{\mathrm{N}}\Omega, \\ du(\eta_g^\Omega) = ru & \text{on } \partial_{\mathrm{R}}\Omega, \end{cases} \tag{2-2}$$

where η_g^Ω is the almost-everywhere defined outward unit normal induced by g on $\partial\Omega$.

It is obviously the case that the Neumann boundary conditions can be regarded as a special case of their inhomogeneous counterpart, however it is convenient — somewhat artificially — to distinguish them in view of the later applications we have in mind — the study of the Morse index of free boundary minimal surfaces.

2.2. ***Sobolev spaces and traces.*** To pose the problem precisely, we introduce the Sobolev space $H^1(\Omega, g)$ consisting of all real-valued functions in $L^2(\Omega, g)$ which have a weak g-gradient whose pointwise g norm is also in $L^2(\Omega, g)$; then $H^1(\Omega, g)$ is a Hilbert space equipped with the inner product

$$\langle u, v\rangle_{H^1(\Omega,g)} := \int_\Omega (uv + g(\nabla_g u, \nabla_g v))\, d\mathscr{H}^d(g),$$

integrating with respect to the d-dimensional Hausdorff measure induced by g. (We say a function $u \in L^1_{\mathrm{loc}}(\Omega, g)$ has a weak g-gradient $\nabla_g u$ if $\nabla_g u$ is a measurable vector field on Ω with pointwise g norm in $L^1_{\mathrm{loc}}(\Omega, g)$ and $\int_\Omega g(X, \nabla_g u)\, d\mathscr{H}^d(g) = -\int_\Omega u \operatorname{div}_g X\, d\mathscr{H}^d(g)$ for every smooth vector field X on Ω of relatively compact support, where $\operatorname{div}_g X$ is the g divergence of X; $\nabla_g u$ is uniquely defined whenever it exists, modulo vector fields vanishing almost everywhere.)

Under our assumptions on $\partial\Omega$, we have a bounded trace map $H^1(\Omega, g) \to L^2(\partial\Omega, g)$ extending the restriction map $C^1(\overline{\Omega}) \to C^0(\partial\Omega)$. (The Hilbert space $L^2(\partial\Omega, g)$ is defined using either the $(d-1)$-dimensional Hausdorff measure $\mathscr{H}^{d-1}(g)$ induced by g or, equivalently, the almost-everywhere defined volume density induced by g on $\partial\Omega$.) In fact, we have not only boundedness of this map but also the stronger inequality

$$\|u|_{\partial\Omega}\|_{L^2(\partial\Omega,g)} \le C(\Omega, g)(\epsilon\|u\|_{H^1(\Omega,g)} + C(\epsilon)\|u\|_{L^2(\Omega,g)}) \tag{2-3}$$

for all $u \in H^1(\Omega, g)$, all $\epsilon > 0$, some $C(\Omega, g)$ independent of u and ϵ, and some $C(\epsilon)$ independent of u and (Ω, g). (This can be deduced, for example, by inspecting the proof of Theorem 4.6 in [Evans and Gariepy 2015]: specifically, we can apply the Cauchy–Schwarz inequality (weighting with ϵ, as is standard) to the inequality immediately above the line labeled ($\star\star\star$) on page 158 of the preceding reference, whose treatment of Lipschitz domains in Euclidean space is readily adapted to our setting.)

For each $\mathrm{C} \in \{\mathrm{D}, \mathrm{N}, \mathrm{R}\}$, indicating one of the boundary conditions we wish to impose, by composing the preceding trace map with the restriction $L^2(\partial\Omega, g) \to L^2(\partial_{\mathrm{C}}\Omega, g)$, since $\partial_{\mathrm{C}}\Omega$ is open in $\partial\Omega$, we also get a trace map $\cdot|_{\partial_{\mathrm{C}}} : H^1(\Omega, g) \to L^2(\partial_{\mathrm{C}}\Omega, g)$. In practice we will consider traces on just $\partial_{\mathrm{D}}\Omega$ and $\partial_{\mathrm{R}}\Omega$. Considering the condition on $\partial_{\mathrm{D}}\Omega$, we will then define

$$H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g) := \{u \in H^1(\Omega, g) : u|_{\partial_{\mathrm{D}}\Omega} = 0\},$$

which is obviously to be understood in the sense of traces, in the terms we just described, and we remark that (2-3) also clearly holds with $\partial\Omega$ on the left-hand side replaced by $\partial_{\mathrm{R}}\Omega$ (or by $\partial_{\mathrm{D}}\Omega$ or $\partial_{\mathrm{N}}\Omega$, but we have no need of the inequality in these cases).

2.3. ***Bilinear forms and their eigenvalues and eigenspaces.*** Corresponding to the above data we define the bilinear form $T = T[\Omega, g, q, r, \partial_{\mathrm{D}}\Omega, \partial_{\mathrm{N}}\Omega, \partial_{\mathrm{R}}\Omega]$ by

$$T : H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g) \times H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g) \to \mathbb{R},$$
$$(u, v) \mapsto \int_\Omega (g(\nabla_g u, \nabla_g v) - quv)\, d\mathscr{H}^d(g) - \int_{\partial_{\mathrm{R}}\Omega} ruv\, d\mathscr{H}^{d-1}(g). \tag{2-4}$$

Then T is symmetric, bounded, and coercive as encoded in the following three equations, respectively:

$$\text{for all } u, v \in H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g), \quad T(u, v) = T(v, u),$$
$$\text{for all } u \in H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g), \quad T(u, u) \le (1 + C(\Omega, g, q, r))\|u\|^2_{H^1(\Omega,g)}, \tag{2-5}$$
$$\text{for all } u \in H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g), \quad T(u, u) \ge \tfrac{1}{2}\|u\|^2_{H^1(\Omega,g)} - C(\Omega, g, q, r)\|u\|^2_{L^2(\Omega,g)}, \tag{2-6}$$

where, for (2-5) and (2-6), one can take $C(\Omega, g, q, r) = \|q\|_{C^0(\bar\Omega)} + C(\Omega, g)\|r\|_{C^0(\overline{\partial_{\mathrm{R}}\Omega})}$ thanks to the trace inequality (2-3). From these three properties and the Riesz representation theorem for Hilbert spaces, it follows that, for some constant $\Lambda = \Lambda(\Omega, g, q, r) > 0$, there exists a linear map $R : L^2(\Omega, g) \to H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g)$ such that

$$T(Rf, v) + \Lambda\langle \iota Rf, \iota v\rangle_{L^2(\Omega,g)} = \langle f, \iota v\rangle_{L^2(\Omega,g)}$$

for all functions $f \in L^2(\Omega, g)$ and $v \in H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g)$, where we have introduced the inclusion map $\iota : H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g) \to L^2(\Omega, g)$.

(Of course, if f is smooth then standard elliptic *interior* regularity results ensure that u is as well smooth on Ω and there satisfies the equation $-(\Delta_g + q - \Lambda)u = f$ in a classical pointwise sense.) Since the inclusion $H^1(\Omega, g) \hookrightarrow L^2(\Omega, g)$ is compact (see for example Section 7 of Chapter 4 of [Taylor 1996]) and of course the inclusion of the closed subspace $H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g) \hookrightarrow H^1(\Omega, g)$ is bounded, the aforementioned maps $\iota : H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g) \to L^2(\Omega, g)$ and the composite $\iota R : L^2(\Omega, g) \to L^2(\Omega, g)$ are also

both compact operators. Furthermore, to confirm that ιR is symmetric, we simply note that (by appealing to the equation defining the operator R, with Rf_1 and Rf_2 in place of v)

$$\begin{aligned}\langle f_2, \iota R f_1\rangle_{L^2(\Omega,g)} &= T(Rf_2, Rf_1) + \Lambda\langle \iota R f_2, \iota R f_1\rangle_{L^2(\Omega,g)}\\ &= T(Rf_1, Rf_2) + \Lambda\langle \iota R f_1, \iota R f_2\rangle_{L^2(\Omega,g)} = \langle f_1, \iota R f_2\rangle_{L^2(\Omega,g)}\end{aligned}$$

for all $f_1, f_2 \in L^2(\Omega, g)$. That being clarified, to improve readability we will from now on refrain from explicitly indicating the inclusion map ι in our equations.

With slight abuse of language, in the setting above we call $\lambda \in \mathbb{R}$ an *eigenvalue* of T if there exists a nonzero $u \in H^1_{\partial_D\Omega}(\Omega, g)$ such that,

$$\text{for all } v \in H^1_{\partial_D\Omega}(\Omega, g), \quad T(u, v) = \lambda\langle u, v\rangle_{L^2(\Omega,g)}, \tag{2-7}$$

and we call any such u an *eigenfunction* of T with eigenvalue λ. (We caution that the notions of eigenfunctions and eigenvalues depend not only on T but also on the underlying metric g; for the sake of convenience we choose to suppress the latter dependence from our notation.)

Hence, as a consequence of the key facts we presented before this definition, one can prove by well-known arguments the existence of a discrete spectrum for the "shifted" elliptic operator $(\Delta_g + q) - \Lambda$ subject to the very same boundary conditions (2-2). As a straightforward corollary, by accounting for the shift, we obtain the following conclusions for T:

- The set of eigenvalues of T is discrete in $\mathbb{R}$ and bounded below.
- For each eigenvalue of T, the corresponding eigenspace has finite dimension.
- There exists a Hilbertian basis $\{e_j\}_{j=1}^{\infty}$ for $L^2(\Omega, g)$ consisting of eigenfunctions of T.
- $\{e_j\}_{j=1}^{\infty}$ has dense span in $H^1_{\partial_D\Omega}(\Omega, g)$.

(To avoid ambiguities, we remark that the phrase *Hilbertian basis* refers to a countable, complete orthonormal system for the Hilbert space in question.) For each integer $i \geq 1$, we write $\lambda_i(T)$ for the i-th eigenvalue of T (listed with repetitions in nondecreasing order, in the usual fashion). We have the usual min-max characterization

$$\lambda_i(T) = \min\left\{\max\left\{\frac{T(w, w)}{\|w\|^2_{L^2(\Omega,g)}} : 0 \neq w \in W\right\} : W \underset{\text{subspace}}{\subset} H^1_{\partial_D\Omega}(\Omega, g),\ \dim W = i\right\}. \tag{2-8}$$

Next, for any $t \in \mathbb{R}$, we let $E^{=t}(T)$ denote the (possibly trivial) linear span, in $H^1_{\partial_D\Omega}(\Omega, g)$, of the eigenfunctions of T with eigenvalue t, and, more generally, for any $t \in \mathbb{R}$ and any binary relation $\sim$ on $\mathbb{R}$ (in practice $<$, $\leq$, $>$, $\geq$, or $=$), we set

$$E^{\sim t}(T) := \text{Closure}_{L^2(\Omega,g)}\left(\text{Span}\left(\bigcup_{s\sim t} E^{=s}(T)\right)\right),$$

and we denote the corresponding orthogonal projection by

$$\pi_T^{\sim t} : L^2(\Omega, g) \to E^{\sim t}(T).$$

That is, the space $E^{\sim t}(T)$ has been defined to be the closure in $L^2(\Omega, g)$ of the span of all eigenfunctions of T having eigenvalue λ such that $\lambda \sim t$. Of course $E^{\sim t}(T)$ is a subspace of $H^1_{\partial_D\Omega}(\Omega, g)$ — in particular — whenever the former has finite dimension. Taking $\sim$ to be equality clearly reproduces the originally defined space $E^{=t}(T)$.

For future use, observe that the above spectral theorem for T implies

$$\begin{gathered}(E^{\sim t}(T))^{\perp_{L^2(\Omega,g)}} = E^{\not\sim t}(T), \quad E^{<t}(T) \underset{\text{subspace}}{\subset} E^{\leq t}(T) \underset{\text{subspace}}{\subset} H^1_{\partial_D\Omega}(\Omega, g), \\ \text{for all } u \in E^{\sim t}(T) \cap H^1_{\partial_D\Omega}(\Omega, g), \quad T(u,u) \sim t\|u\|^2_{L^2(\Omega,g)} \text{ for } \sim \text{ any one of } <, \leq, >, \geq,\end{gathered} \tag{2-9}$$

and,

$$\text{for all } u \in H^1_{\partial_D\Omega}(\Omega, g) \cap (E^{\leq t}(T) \cup E^{\geq t}(T)), \quad T(u,u) = t\|u\|^2_{L^2(\Omega,g)} \implies u \in E^{=t}(T),$$

throughout which t is any real number (not necessarily an eigenvalue of T) and where in the first equality of (2-9) $\sim$ is any relation on $\mathbb{R}$ and $\not\sim$ its negation (so that $\{s \not\sim t\} = \mathbb{R} \setminus \{s \sim t\}$ for any $t \in \mathbb{R}$).

Index and nullity. In the setting above and under the corresponding standing assumption, we shall define the nonnegative integers

$$\operatorname{ind}(T) := \dim E^{<0}(T) \quad \text{and} \quad \operatorname{nul}(T) := \dim E^{=0}(T)$$

called, respectively, the *index* and *nullity* of T. Such invariants will be of primary interest in our applications.

2.4. ***Group actions.*** Let G be a finite group of smooth diffeomorphisms of M, each restricting to an isometry of $(\overline{\Omega}, g)$. Then, as for any group of diffeomorphism of Ω, we have the standard (left) action of G on functions on Ω via pullback:

$$(\phi, u) \mapsto u \circ \phi^{-1} = \phi^{-1*}u \quad \text{for all } \phi \in G, \;\; u : \Omega \to \mathbb{R}.$$

We say that a function u is G-invariant if it is invariant under this action: equivalently $u \circ \phi = u$ for all $\phi \in G$.

We can also twist this action by orthogonal transformations on the fiber $\mathbb{R}$: given in addition to G a group homomorphism $\sigma : G \to \mathrm{O}(1) = \{-1, 1\}$, we define the action

$$(\phi, u) \mapsto \sigma(\phi)(u \circ \phi^{-1}) = \sigma(\phi)\phi^{-1*}u \quad \text{for all } \phi \in G, \;\; u : \Omega \to \mathbb{R},$$

and we call a function (G, σ)-invariant if it is invariant under this action. Obviously the above standard action $(\phi, u) \mapsto u \circ \phi^{-1}$ is recovered by taking the trivial homomorphism $\sigma \equiv 1$. We also comment that one could of course replace $\mathbb{R}$ by $\mathbb{C}$ and correspondingly $\mathrm{O}(1)$ by $\mathrm{U}(1)$ (and in the preceding sections instead work with Sobolev spaces over $\mathbb{C}$), though we restrict our attention to real-valued functions in this article.

Since, by virtue of our initial requirement, G is a group of isometries of $(\overline{\Omega}, g)$, the above twisted action yields a unitary representation of G in $L^2(\Omega, g)$, i.e., a group homomorphism

$$\hat{\sigma} : G \to \mathrm{O}(L^2(\Omega, g)), \quad \phi \mapsto \sigma(\phi)\phi^{-1*}, \tag{2-10}$$

whose targets are the global isometries of $L^2(\Omega, g)$; we note that the same conclusions hold with $H^1(\Omega, g)$ in place of $L^2(\Omega, g)$. The corresponding subspaces of (G, σ)-invariant functions, in $L^2(\Omega, g)$ or $H^1(\Omega, g)$, are readily checked to be closed and thus Hilbert spaces themselves. That said, we define the orthogonal projection

$$\pi_{G,\sigma} : L^2(\Omega, g) \to L^2(\Omega, g), \quad u \mapsto \frac{1}{|G|} \sum_{\phi \in G} \hat{\sigma}(\phi) u. \tag{2-11}$$

Here $|G|$ is the order of G, which — we recall — is assumed throughout to be finite. The image of $L^2(\Omega, g)$ under $\pi_{G,\sigma}$ thus consists of (G, σ)-invariant functions.

Remark 2.1. One could lift the finiteness assumption, say by allowing G to be a compact Lie group, requiring σ to be continuous, and replacing the finite average in (2-11) with the average over G with respect to its Haar measure (which reduces to the former for finite G). However, with a view towards our later applications, in this article we content ourselves with the finiteness assumption, which allows for a lighter exposition.

Henceforth we make the additional assumptions that G *globally* (i.e., as sets) preserves each of $\partial_{\mathrm{D}}\Omega$, $\partial_{\mathrm{N}}\Omega$, and $\partial_{\mathrm{R}}\Omega$, and that q and r are both G-invariant. Each element of $\hat{\sigma}(G)$ then preserves also $H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g)$ and the bilinear form T, and the projection $\pi_{G,\sigma}$ commutes with the projection $\pi_T^{\sim t}$ for any $t \in \mathbb{R}$ and binary relation $\sim$ on $\mathbb{R}$ (as above). In particular $\pi_{G,\sigma}$ preserves each eigenspace $E^{=t}(T)$ of T, and more generally the space

$$E^{\sim t}_{G,\sigma}(T) := \pi_{G,\sigma}(E^{\sim t}(T)) \tag{2-12}$$

is a subspace of $E^{\sim t}(T)$.

For each integer $i \geq 1$, we can then define $\lambda_i^{G,\sigma}(T)$, the i-th (G, σ)-eigenvalue of T, to be the i-th eigenvalue of T having a (G, σ)-invariant eigenfunction (by definition nonzero), counting with multiplicity as before; equivalently one can work with spaces of (G, σ)-invariant functions and derive the analogous conclusions as in Section 2.3 directly in that setting.

Remark 2.2. We explicitly note, for the sake of completeness, that under no additional assumptions on the group G and the homomorphism σ it is possible that the space of (G, σ)-invariant functions be finite dimensional (possibly even of dimension zero). This type of phenomenon happens, for instance, when every point of the manifold M is a fixed point of an isometry on which σ takes the value -1. In this case, all conclusions listed above still hold but need to be understood with a bit of care: the corresponding sequence of eigenvalues $\lambda_1^{G,\sigma}(T) \leq \lambda_2^{G,\sigma}(T) \leq \cdots$ will in fact just be a finite sequence, consisting say of $I(G, \sigma)$ elements, counted with multiplicity as usual; we shall use the convention that $\lambda_i^{G,\sigma}(T) = +\infty$ for $i > I(G, \sigma)$. That being said, we also remark that this phenomenon patently does not occur for the Jacobi form of the two sequences of free boundary minimal surfaces we examine in Sections 4 and 5.

In this equivariant framework we still have the corresponding min-max characterization

$$\lambda_i^{G,\sigma}(T) = \min\left\{ \max\left\{ \frac{T(w, w)}{\|w\|^2_{L^2(\Omega,g)}} : 0 \neq w \in W \right\} : W \underset{\text{subspace}}{\subset} \pi_{G,\sigma}(H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g)), \dim W = i \right\}. \tag{2-13}$$

We also define the (G,σ)-index and (G,σ)-nullity

$$\operatorname{ind}_G^\sigma(T) := \dim E^{<0}_{G,\sigma}(T) \quad \text{and} \quad \operatorname{nul}_G^\sigma(T) := \dim E^{=0}_{G,\sigma}(T)$$

of T. Obviously we can recover $E^{\sim t}(T)$, $\lambda_i(T)$, and the standard index and nullity by taking G to be the trivial group. As mentioned in the introduction, we reiterate that it is one of the goals of the present article to study, *for fixed g and T*, how these numbers (index $\operatorname{ind}_G^\sigma(T)$ and nullity $\operatorname{nul}_G^\sigma(T)$) depend on G and σ.

Terminology. For the sake of brevity, we shall employ the phrase *admissible data* to denote any tuple $(\Omega, g, q, r, \partial_{\mathrm{D}}\Omega, \partial_{\mathrm{N}}\Omega, \partial_{\mathrm{R}}\Omega, G, \sigma)$ satisfying all the standing assumptions presented up to now. We digress briefly to highlight two important special cases, which warrant additional notation.

Example 2.3 (actions of order-2 groups). When $|G| = 2$, there are precisely two homomorphisms $G \to \mathrm{O}(1)$. Considering such homomorphisms and the corresponding (G,σ)-invariant functions, we may define G-even or G-odd functions. Hence, we may call ind_G^+ and ind_G^- the G-even and G-odd index, and likewise for the nullity. Clearly, we always have

$$\begin{cases} \operatorname{ind}(T) = \operatorname{ind}_G^+(T) + \operatorname{ind}_G^-(T), \\ \operatorname{nul}(T) = \operatorname{nul}_G^+(T) + \operatorname{nul}_G^-(T). \end{cases} \tag{2-14}$$

Example 2.4 (actions of self-congruences of two-sided hypersurfaces). Suppose, momentarily, that (M, g) is isometrically embedded (as a codimension-one submanifold) in a Riemannian manifold (N, h), that the set Ω is connected, and assume further that the normal bundle of M over Ω is trivial. Then we can pick a unit normal ν on Ω and thereby identify — as usual — sections of the normal bundle of $M|_\Omega$ with functions on Ω. With this interpretation of functions on Ω in mind and G now a finite group of diffeomorphisms of N that map Ω onto itself (as a set) and everywhere on Ω preserve the ambient metric h meaning that $\phi^* h = h$ for any $\phi \in G$, we have a natural action given by

$$(\phi, u) \mapsto \operatorname{sgn}_\nu(\phi)(u \circ \phi^{-1}) \quad \text{for all } \phi \in G, \quad u : \Omega \to \mathbb{R},$$

where $\operatorname{sgn}_\nu(\phi) := h(\phi_* \nu, \nu)$ is a constant in $\mathrm{O}(1) = \{1, -1\}$. We shall further assume that the action of G on Ω is faithful, meaning that only the identity element fixes Ω pointly; this assumption is always satisfied in our applications.

In this context we continue to say that a function $u : \Omega \to \mathbb{R}$ is G-invariant if $u = u \circ \phi$ for all $\phi \in G$, and we say rather that u is G-equivariant if $u = \operatorname{sgn}_\nu(\phi) u \circ \phi$ for all $\phi \in G$ (that is, noting the identity $\operatorname{sgn}_\nu(\phi) = \operatorname{sgn}_\nu(\phi^{-1})$, provided u is invariant under the sgn_ν-twisted G action).

Similarly, in this context, we set

$$\operatorname{ind}_G(T) := \operatorname{ind}_G^{\operatorname{sgn}_\nu}(T) \quad \text{and} \quad \operatorname{nul}_G T := \operatorname{nul}_G^{\operatorname{sgn}_\nu}(T), \tag{2-15}$$

which we may refer to as simply the G-equivariant index and G-equivariant nullity of T. We point out that we are abusing notation in the above definitions in that, on the right-hand side of each, in place of G we mean really the group, isomorphic to G by virtue of the faithfulness assumption, obtained by restricting each element of G to Ω, and in place of sgn_ν we mean really the corresponding homomorphism, well-defined by the faithfulness assumption, on this last group of isometries of Ω.

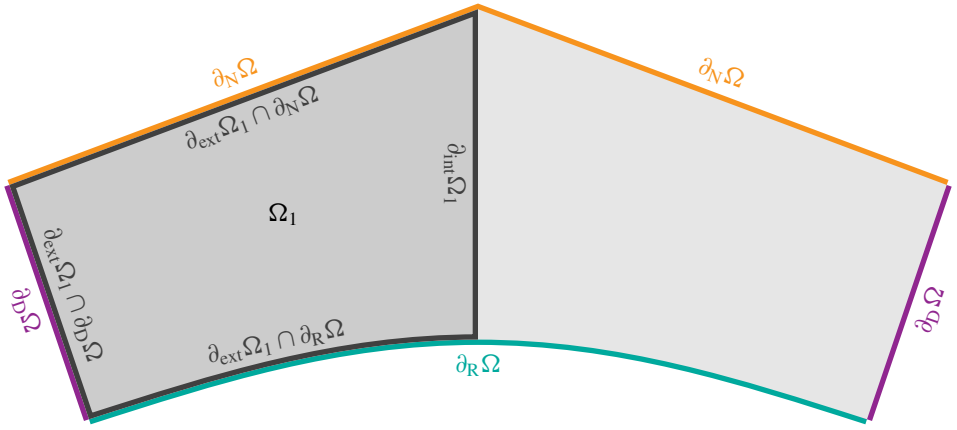

Figure 1. Example of a Lipschitz domain Ω with subdomain Ω_1.

We now return to the more general assumptions on G preceding this paragraph.

2.5. ***Subdomains.*** Suppose that $\Omega_1 \subset \Omega$ is another Lipschitz domain of M (see Figure 1). We shall define

$$\begin{aligned}
\partial_{\mathrm{int}}\Omega_1 &:= \partial\Omega_1 \cap \Omega, & \partial_{\mathrm{ext}}\Omega_1 &:= \partial\Omega_1 \setminus \overline{\partial_{\mathrm{int}}\Omega_1},\\
\partial_{\mathrm{D}}^{\mathrm{D}_{\mathrm{int}}}\Omega_1 &:= (\partial_{\mathrm{ext}}\Omega_1 \cap \partial_{\mathrm{D}}\Omega) \cup \partial_{\mathrm{int}}\Omega_1, & \partial_{\mathrm{D}}^{\mathrm{N}_{\mathrm{int}}}\Omega_1 &:= \partial_{\mathrm{ext}}\Omega_1 \cap \partial_{\mathrm{D}}\Omega,\\
\partial_{\mathrm{N}}^{\mathrm{D}_{\mathrm{int}}}\Omega_1 &:= \partial_{\mathrm{ext}}\Omega_1 \cap \partial_{\mathrm{N}}\Omega, & \partial_{\mathrm{N}}^{\mathrm{N}_{\mathrm{int}}}\Omega_1 &:= (\partial_{\mathrm{ext}}\Omega_1 \cap \partial_{\mathrm{N}}\Omega) \cup \partial_{\mathrm{int}}\Omega_1,\\
\partial_{\mathrm{R}}^{\mathrm{D}_{\mathrm{int}}}\Omega_1 &:= \partial_{\mathrm{ext}}\Omega_1 \cap \partial_{\mathrm{R}}\Omega, & \partial_{\mathrm{R}}^{\mathrm{N}_{\mathrm{int}}}\Omega_1 &:= \partial_{\mathrm{ext}}\Omega_1 \cap \partial_{\mathrm{R}}\Omega.
\end{aligned} \tag{2-16}$$

In this way we prepare to pose two different sets of boundary conditions on Ω_1, whereby, roughly speaking, in both cases $\partial\Omega_1$ inherits whatever boundary condition is in effect on $\partial\Omega$ wherever the two meet (corresponding to $\partial_{\mathrm{ext}}\Omega_1$) and the two sets of conditions are distinguished by placing either the Dirichlet or the Neumann condition on the remainder of the boundary (corresponding to $\partial_{\mathrm{int}}\Omega_1$). Naturally associated to these two sets of conditions are the bilinear forms

$$\begin{aligned}
T_{\Omega_1}^{\mathrm{D}_{\mathrm{int}}} &:= T\big[\Omega_1, g, q, r, \partial_{\mathrm{D}}^{\mathrm{D}_{\mathrm{int}}}\Omega_1, \partial_{\mathrm{N}}^{\mathrm{D}_{\mathrm{int}}}\Omega_1, \partial_{\mathrm{R}}^{\mathrm{D}_{\mathrm{int}}}\Omega_1\big],\\
T_{\Omega_1}^{\mathrm{N}_{\mathrm{int}}} &:= T\big[\Omega_1, g, q, r, \partial_{\mathrm{D}}^{\mathrm{N}_{\mathrm{int}}}\Omega_1, \partial_{\mathrm{N}}^{\mathrm{N}_{\mathrm{int}}}\Omega_1, \partial_{\mathrm{R}}^{\mathrm{N}_{\mathrm{int}}}\Omega_1\big],
\end{aligned} \tag{2-17}$$

defined, respectively, on the Sobolev spaces

$$H^1_{\partial_{\mathrm{D}}^{\mathrm{D}_{\mathrm{int}}}\Omega_1}(\Omega_1, g) \quad \text{and} \quad H^1_{\partial_{\mathrm{D}}^{\mathrm{N}_{\mathrm{int}}}\Omega_1}(\Omega_1, g).$$

Recalling (G, σ) from above, with the tacit understanding that $(\Omega, g, q, r, \partial_{\mathrm{D}}\Omega, \partial_{\mathrm{N}}\Omega, \partial_{\mathrm{R}}\Omega, G, \sigma)$ is admissible, we further assume that each element of G maps Ω_1 onto itself; since G preserves Ω and respects the decomposition (2-1), it follows that it also respects the decompositions (2-16). Somewhat abusively, we shall write $\hat{\sigma}$ and $\pi_{G,\sigma}$ not only for the maps (2-10) and (2-11) but also for their counterparts with Ω replaced by Ω_1, which are well-defined under our assumptions. The spaces $E^{\sim t}_{G,\sigma}(T_{\Omega_1}^{\mathrm{D}_{\mathrm{int}}})$ and $E^{\sim t}_{G,\sigma}(T_{\Omega_1}^{\mathrm{N}_{\mathrm{int}}})$ as in (2-12) are then also well-defined.

3. Fundamental tools

3.1. ***Index and nullity bounds in the style of Montiel and Ros.*** Recalling the notation and assumptions of Section 2, suppose now that we have not only $\Omega_1 \subset \Omega$ as above, but also (open) Lipschitz subdomains $\Omega_1, \dots, \Omega_n \subset \Omega$ which are pairwise disjoint, each of which satisfies the same assumptions as Ω_1 in Section 2.5 and whose closures cover $\overline{\Omega}$. In particular, we assume that each element of the group G maps each subdomain Ω_i onto itself. We assume further that G acts transitively on the connected components of Ω and note that this last condition is always satisfied in the important special case that Ω is connected.

Proposition 3.1 (Montiel–Ros bounds on the number of eigenvalues below a threshold). *With assumptions as in the preceding paragraph and notation as in Section 2, the following inequalities hold for any $t \in \mathbb{R}$:*

(i) $\dim E^{<t}_{G,\sigma}(T) \geq \dim E^{<t}_{G,\sigma}(T^{\mathrm{D_{int}}}_{\Omega_1}) + \sum_{i=2}^{n} \dim E^{\leq t}_{G,\sigma}(T^{\mathrm{D_{int}}}_{\Omega_i})$,

(ii) $\dim E^{\leq t}_{G,\sigma}(T) \leq \dim E^{\leq t}_{G,\sigma}(T^{\mathrm{N_{int}}}_{\Omega_1}) + \sum_{i=2}^{n} \dim E^{<t}_{G,\sigma}(T^{\mathrm{N_{int}}}_{\Omega_i})$.

The statement and proof of Proposition 3.1 are adapted from Lemmas 12 and 13 of [Montiel and Ros 1991], which concern the spectrum of the Laplacian on branched coverings of the round sphere and rely on standard, fundamental facts about eigenvalues and eigenfunctions of Schrödinger operators, much as in the proof of the classical Courant nodal domain theorem. These arguments are readily applied to more general Schrödinger operators on more general domains, as observed for instance in [Kapouleas and Wiygul 2020], where such bounds in the style of Montiel and Ros played a major role in the computation of the index and nullity of the $\xi_{g,1}$ Lawson surfaces. Here, instead, we present an extended version allowing for the imposition of mixed (Robin and Dirichlet) boundary conditions and invariance under a group action; as mentioned in the introduction, this level of generality is motivated by the goal of bounding (from above and below) the G-equivariant Morse index of free boundary minimal surfaces. (Our treatment of course includes the fundamental case when G is the trivial group.)

Proof. Throughout the proof we will make free use of the consequences (2-9) of the spectral theorem for the various bilinear forms appearing in the statement. Fix $t \in \mathbb{R}$. For (i) we will verify injectivity of the map

$$\iota^{\mathrm{D_{int}}} : E^{<t}_{G,\sigma}(T^{\mathrm{D_{int}}}_{\Omega_1}) \oplus \bigoplus_{i=2}^{n} E^{\leq t}_{G,\sigma}(T^{\mathrm{D_{int}}}_{\Omega_i}) \to E^{<t}_{G,\sigma}(T), \quad (u_1, u_2, \dots, u_n) \mapsto \pi^{<t}_{T}\left(\sum_{i=1}^{n} U_i\right),$$

where each U_i is the extension to Ω of u_i such that U_i vanishes on $\Omega \setminus \Omega_i$. Clearly, each such extension lies in the image of $\pi_{G,\sigma}$, which, as observed above, commutes with $\pi^{<t}_{T}$, so that the map is indeed well-defined with its asserted target. Now suppose that $(u_1, \dots, u_n)$ belongs to the domain of $\iota^{\mathrm{D_{int}}}$, and set $v := \sum_{i=1}^{n} U_i$. Then $v \in H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g)$ and

$$T(v, v) = \sum_{i=1}^{n} T^{\mathrm{D_{int}}}_{\Omega_i}(u_i, u_i) \leq t \|v\|^2_{L^2(\Omega,g)},$$

with equality possible only when $u_1 = 0$. To check injectivity suppose next that $\iota^{\mathrm{D_{int}}}(u_1, \dots, u_n) = 0$. By definition of $\iota^{\mathrm{D_{int}}}$ this assumption means that v is $L^2(\Omega, g)$-orthogonal to $E^{<t}_{G,\sigma}(T)$, and so in view of the preceding inequality and (2-9) we have $v \in E^{=t}_{G,\sigma}(T)$. Thus, v satisfies the elliptic equation

$(\Delta_g + q + t)u = 0$; moreover, we must also have $v|_{\Omega_1} = u_1 = 0$, but now the unique continuation principle [Aronszajn 1957] implies that $v = 0$, whence $(u_1, \dots, u_n) = 0$, completing the proof of (i).

For (ii) we verify injectivity of

$$\iota^{\mathrm{N_{int}}} : E^{\leq t}_{G,\sigma}(T) \to E^{\leq t}_{G,\sigma}(T^{\mathrm{N_{int}}}_{\Omega_1}) \oplus \bigoplus_{i=2}^{n} E^{<t}_{G,\sigma}(T^{\mathrm{N_{int}}}_{\Omega_i}),$$
$$u \mapsto (\pi^{\leq t}_{T^{\mathrm{N_{int}}}_{\Omega_1}} u|_{\Omega_1}, \pi^{<t}_{T^{\mathrm{N_{int}}}_{\Omega_2}} u|_{\Omega_2}, \dots, \pi^{<t}_{T^{\mathrm{N_{int}}}_{\Omega_n}} u|_{\Omega_n})$$

instead. Note that

$$u|_{\Omega_i} \in \pi_{G,\sigma}(L^2(\Omega_i, g)) \cap H^1_{\partial^{\mathrm{N_{int}}}_{\mathrm{D}}\Omega_i}(\Omega_i, g) \tag{3-1}$$

for each i; in particular, the left inclusion and the commutativity of $\pi_{G,\sigma}$ with each of the spectral projections appearing in the definition of $\iota^{\mathrm{N_{int}}}$ ensure that the latter really is well-defined. Suppose then that u belongs to the domain of $\iota^{\mathrm{N_{int}}}$ and $\iota^{\mathrm{N_{int}}} u = (0, \dots, 0)$. The second assumption (making use of the right inclusion in (3-1) in addition to (2-9)) implies

$$T(u, u) = \sum_{i=1}^{n} T^{\mathrm{N_{int}}}_{\Omega_i}(u|_{\Omega_i}, u|_{\Omega_i}) \geq t \|u\|^2_{L^2(\Omega, g)},$$

with equality possible only when $u|_{\Omega_1} = 0$. Recalling that, by assumption, $u \in E^{\leq t}_{G,\sigma}(T)$, we therefore conclude, appealing to (2-9), that $u \in E^{=t}_{G,\sigma}(T)$ and indeed this equality case holds. In particular, u satisfies the elliptic equation $(\Delta_g + q + t)u = 0$, but then the condition $u|_{\Omega_1} = 0$ and the unique continuation principle imply $u = 0$, ending the proof. □

In particular, in our applications we will repeatedly (yet not always) appeal to the special case when $t = 0$ and Ω (most often equal to the whole ambient manifold itself M) is partitioned into a finite collection of pairwise isometric domains.

Corollary 3.2 (Montiel–Ros index and nullity bounds from isometric pieces). *In the setting of the previous proposition, let us suppose the domains $\Omega_1, \dots, \Omega_n$ to be pairwise isometric via isometries of Ω. Then*

(i) $\mathrm{ind}^\sigma_G(T) \geq n \,\mathrm{ind}^\sigma_G(T^{\mathrm{D_{int}}}_{\Omega_1}) + (n-1)\,\mathrm{nul}^\sigma_G(T^{\mathrm{D_{int}}}_{\Omega_1})$,

(ii) $\mathrm{ind}^\sigma_G(T) + \mathrm{nul}^\sigma_G(T) \leq n \,\mathrm{ind}^\sigma_G(T^{\mathrm{N_{int}}}_{\Omega_1}) + \mathrm{nul}^\sigma_G(T^{\mathrm{N_{int}}}_{\Omega_1})$.

Remark 3.3. We further explicitly note how the two inequalities given in the previous corollary jointly imply the "compatibility condition" that

$$(n-1)\,\mathrm{nul}^\sigma_G(T^{\mathrm{D_{int}}}_{\Omega_1}) - \mathrm{nul}^\sigma_G(T^{\mathrm{N_{int}}}_{\Omega_1}) \leq n(\mathrm{ind}^\sigma_G(T^{\mathrm{N_{int}}}_{\Omega_1}) - \mathrm{ind}^\sigma_G(T^{\mathrm{D_{int}}}_{\Omega_1})), \tag{3-2}$$

which in general has nontrivial content.

Remark 3.4. The requirement that the domains in question be G-invariant implies, in certain examples, that some of them may in fact have to be taken disconnected. We will however discuss, in the next subsection, how this nuisance may actually be avoided in the totality of our later applications.

3.2. ***Reduction and extension of domain under symmetries.*** With our standing assumptions on (Ω, g), T and (G, σ) in place, encoded in the requirement that they determine admissible data, we again assume that $\Omega_1, \ldots, \Omega_n \subset \Omega$ are pairwise disjoint Lipschitz domains whose closures cover Ω. However, for the specific purposes of this section, we assume Ω is connected and, rather than assuming G-invariance of each Ω_i, we instead suppose that G preserves the collection $\{\Omega_i\}_{i=1}^n$ (while — as per our general postulate — also respecting the decomposition (2-1), which dictates the boundary conditions (2-2)) and acts transitively on its elements (so in particular the Ω_i are pairwise isometric). The (possibly trivial) subgroup of G which preserves Ω_1 we call H. Note that H preserves $\partial_{\mathrm{int}}\Omega_1$ in particular.

For each $p \in \partial_{\mathrm{int}}\Omega_1$, we define

$$G_p := \big\{\phi \in G : \exists U \underset{\text{open}}{\subseteq} \partial_{\mathrm{int}}\Omega_1,\ p \in U \text{ and } \phi|_U = \mathrm{id}\big\}.$$

Then G_p is a subgroup of G having order at most 2, as we now explain. Let $\phi_1, \phi_2 \in G_p$. Then we have open neighborhoods U_1, U_2 of p in $\partial_{\mathrm{int}}\Omega_1$ with U_i fixed pointwise by ϕ_i. By the Lipschitz assumption, there exists $q \in U_1 \cap U_2$ at which $\partial_{\mathrm{int}}\Omega_1$ has a well-defined outward unit conormal η_q. Then, for each i, we have $(d_q\phi_i)(\eta_q) = \epsilon_i \eta_q$ for some $\epsilon_i = \pm 1$. If an $\epsilon_i = +1$, then, since ϕ_i fixes U_i pointwise and Ω is connected, ϕ_i must be the identity on Ω (which comes essentially by arguing, e.g., as in Lemma 4.5 of [Carlotto and Li 2024]). If $\epsilon_1 = \epsilon_2 = -1$, then similarly $\phi_1 \circ \phi_2^{-1}$ is the identity on Ω, establishing our claim. Note also that the set $\{p : |G_p| = 2\}$ is open in $\partial_{\mathrm{int}}\Omega_1$ and that, for each $\chi \in H$, the map $\phi \mapsto \chi \circ \phi \circ \chi^{-1}$ defines an isomorphism from G_p to $G_{\chi(p)}$ which commutes with σ.

For each p, we next set

$$\sigma_p := \begin{cases} 0 & \text{if } |G_p| = 1, \\ 1 & \text{if } |G_p| = 2 \text{ but } \sigma(G_p) = \{+1\}, \\ -1 & \text{if } |G_p| = 2 \text{ and } \sigma(G_p) = \{+1, -1\}, \end{cases}$$

and we in turn define the subsets $\partial_+\Omega_1, \partial_-\Omega_1 \subseteq \partial_{\mathrm{int}}\Omega_1$ by letting (respectively)

$$\partial_\pm\Omega_1 := \sigma_p^{-1}(\pm 1).$$

With the aid of the foregoing observations, we see that $\partial_+\Omega_1$ and $\partial_-\Omega_1$ are open and disjoint, and each is preserved by H. We now impose the additional assumption that their closures cover $\partial_{\mathrm{int}}\Omega_1$, and finally we set $T_{\Omega_1} := T[\Omega_1, g, q, r, \partial_{\mathrm{D}}\Omega_1, \partial_{\mathrm{N}}\Omega_1, \partial_{\mathrm{R}}\Omega_1]$, where

$$\begin{aligned} \partial_{\mathrm{D}}\Omega_1 &:= \partial_-\Omega_1 \cup (\partial_{\mathrm{ext}}\Omega_1 \cap \partial_{\mathrm{D}}\Omega), \\ \partial_{\mathrm{N}}\Omega_1 &:= \partial_+\Omega_1 \cup (\partial_{\mathrm{ext}}\Omega_1 \cap \partial_{\mathrm{N}}\Omega), \\ \partial_{\mathrm{R}}\Omega_1 &:= \partial_{\mathrm{ext}}\Omega_1 \cap \partial_{\mathrm{R}}\Omega. \end{aligned}$$

Lemma 3.5 (reduction and extension of domain under symmetries). *Under the above assumptions, for every integer $i \geq 1$,*

$$\lambda_i^{G,\sigma}(T) = \lambda_i^{H,\sigma}(T_{\Omega_1}),$$

and the (H, σ)-invariant eigenfunctions of T_{Ω_1} are the restrictions to Ω_1 of the (G, σ)-invariant eigenfunctions of T.

Proof. First observe that

$$v \in \pi_{G,\sigma} H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g) \quad \Longrightarrow \quad v|_{\Omega_1} \in \pi_{H,\sigma} H^1_{\partial_{\mathrm{D}}\Omega_1}(\Omega_1, g),$$

using in particular the fact that any (G, σ)-invariant function in $H^1(\Omega, g)$ must have vanishing trace along $\partial_-\Omega_1$. Next observe that our assumptions guarantee that each (H, σ)-invariant function u on Ω_1 has a unique (G, σ)-invariant extension $\bar{u}$ to Ω. This is also true of vector fields, the action being $(\phi, X) \mapsto \sigma(\phi)\phi_* X$. Now suppose $u \in \pi_{H,\sigma} H^1_{\partial_{\mathrm{D}}\Omega_1}(\Omega_1, g)$. Obviously $\bar{u} \in L^2(\Omega, g)$, and we next check that in fact $\bar{u} \in H^1(\Omega, g)$ with $\nabla_g \bar{u} = \overline{\nabla_g u}$.

For this let X be a smooth vector field with support contained in Ω. Let $Y := \pi_{G,\sigma} X$ (meaning we average as in (2-11) but with the appropriate action for vector fields, as above). Then (writing, with slight abuse of notation, $L^2(\Omega, g)$ and $L^2(\Omega_1, g)$ also for the Hilbert spaces of L^2 vector fields on Ω and Ω_1, respectively, in metric g)

$$\begin{aligned}
\langle X, \overline{\nabla_g u}\rangle_{L^2(\Omega,g)} &= \langle Y, \overline{\nabla_g u}\rangle_{L^2(\Omega,g)} = n\langle Y|_{\Omega_1}, \nabla_g u\rangle_{L^2(\Omega_1,g)} \\
&= n\langle 1, \operatorname{div}(uY|_{\Omega_1})\rangle_{L^2(\Omega_1,g)} - n\langle u, \operatorname{div} Y|_{\Omega_1}\rangle_{L^2(\Omega_1,g)} \\
&= n\langle u|_{\partial\Omega_1}, g(\eta_g^{\Omega_1}, Y|_{\partial\Omega_1})\rangle_{L^2(\partial\Omega_1,g)} - \langle \bar{u}, \operatorname{div} Y\rangle_{L^2(\Omega,g)} \\
&= 0 - \langle \operatorname{div} X, \bar{u}\rangle_{L^2(\Omega,g)};
\end{aligned}$$

in the third line we have used the divergence theorem (see for example Theorem 4.6 of [Evans and Gariepy 2015] for a statement serving our assumptions) with $u|_{\partial\Omega_1}$ of course the trace of u and $\eta_g^{\Omega_1}$ the almost everywhere defined outward unit conormal, and in the fourth line we have used the fact that the (G, σ)-invariance of Y forces it to be (almost everywhere) orthogonal to this last conormal on $\partial_+\Omega_1$, while on the other hand, as already noted above, $u|_{\partial\Omega_1}$ vanishes on $\partial_-\Omega_1$. Thus every element of $\pi_{H,\sigma} H^1_{\partial_{\mathrm{D}}\Omega_1}(\Omega_1, g)$ extends uniquely to an element of $\pi_{G,\sigma} H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g)$. It is now straightforward to verify that, for all $t \in \mathbb{R}$, restriction to Ω_1 furnishes a bijection $E^{=t}_{G,\sigma}(T) \to E^{=t}_{H,\sigma}(T_{\Omega_1})$, which implies the claims. □

For the purposes of our later geometric applications, it is convenient to focus on two special cases, which correspond to the examples we presented in Section 2.4.

Example 3.6 (actions of order-2 groups). With respect to our general setup, let $\Omega = M$ and consider $G = \langle\phi\rangle$, where ϕ is a (nontrivial) isometric involution of M. Suppose further (which is not true in general) that the set of fixed points of the action divides M into two open regions, which we shall label Ω_1 and Ω_2. Then note that, arguing as above, one must have $\phi(\Omega_1) = \Omega_2$ (as well as $\phi(\Omega_2) = \Omega_1$). In particular H is the trivial subgroup, just consisting of the identity element. That said, there are two cases depending on the choice of twisting homomorphism $\sigma : G \to \{-1, 1\}$ we consider:

(1) If we let $\sigma(\phi) = +1$, then $\partial_+\Omega_1 = \partial_{\mathrm{int}}\Omega_1$ and $\partial_-\Omega_1 = \varnothing$, so we are considering the (nonequivariant) spectrum of T_{Ω_1} adding a Neumann boundary condition along $\partial_{\mathrm{int}}\Omega_1$.

(2) If we let $\sigma(\phi) = -1$, then $\partial_+\Omega_1 = \varnothing$ and $\partial_-\Omega_1 = \partial_{\mathrm{int}}\Omega_1$, so we are considering the (nonequivariant) spectrum of T_{Ω_1} adding a Dirichlet boundary condition along $\partial_{\mathrm{int}}\Omega_1$.

Example 3.7 (actions of self-congruences of two-sided hypersurfaces). Here we follow up on the discussion of Example 2.4 but specified to $\Omega = M$ for $N = \mathbb{B}^3$ and $G = \mathbb{P}_n$ (i.e., we postulate the ambient manifold to be the Euclidean ball, and the surface M to have prismatic symmetry). We refer the reader to the first part of Section 4 for basic recollections about this and related group action. We let Ω_1 be an open fundamental domain for this action (so that M is covered by the closures of exactly $4n$ pairwise isometric domains); it follows that again H is the trivial subgroup. Considering the sign homomorphism $\sigma : \mathbb{P}_n \to \{-1, +1\}$ defined in Example 2.4, it is readily checked that $\partial_+\Omega_1 = \partial_{\mathrm{int}}\Omega_1$ and $\partial_-\Omega_1 = \varnothing$, and so — when applied to this case — Lemma 3.5 compares (and proves equality of) the (fully-)equivariant spectrum of the problem with the spectrum of a fundamental domain, with Neumann boundary conditions added on each interior side.

3.3. ***Spectral stability.*** As it has been anticipated in the introduction, in our applications we will analyze the spectrum of free boundary minimal surfaces obtained by gluing certain constituting blocks. In that respect, we will need to derive from "geometric convergence" results some corresponding "spectral convergence" results. Suppose we have a sequence $\{(\Omega_n, g_n, q_n, r_n, \partial_{\mathrm{D}}\Omega_n, \partial_{\mathrm{N}}\Omega_n, \partial_{\mathrm{R}}\Omega_n, G_n, \sigma_n)\}$ of admissible data, as well as "limit data" $(\Omega, g, q, r, \partial_{\mathrm{D}}\Omega, \partial_{\mathrm{N}}\Omega, \partial_{\mathrm{R}}\Omega, G_\infty, \sigma_\infty)$, satisfying all our assumptions on admissible data except that G_∞ is possibly allowed to have infinite order. For instance, in our later applications G_∞ is the compact Lie group O(2). Although we originally introduced the notation $\lambda_i^{G_\infty,\sigma_\infty}(T)$, with T the bilinear form associated to the foregoing data, for G_∞ finite, the notion remains well-defined for infinite G_∞. The quantities $\mathrm{ind}_{G_\infty}^{\sigma_\infty}(T)$ and $\mathrm{nul}_{G_\infty}^{\sigma_\infty}(T)$ are likewise defined in this setting; as a special case, we can in turn define $\mathrm{ind}_{G_\infty}(T)$ and $\mathrm{nul}_{G_\infty}(T)$ for G_∞ a suitable infinite-order symmetry group of a hypersurface (as per Example 2.4). That being said, alongside $T[\Omega, g, q, r, \partial_{\mathrm{D}}\Omega, \partial_{\mathrm{N}}\Omega, \partial_{\mathrm{R}}\Omega]$, we then have the corresponding sequence $\{T_n\}$ with

$$T_n := T\big[\Omega_n, g_n, q_n, r_n, \partial_{\mathrm{D}}\Omega_n, \partial_{\mathrm{N}}\Omega_n, \partial_{\mathrm{R}}\Omega_n\big].$$

We will present some conditions on the data that ensure

$$\lim_{n\to\infty} \lambda_i^{G_n,\sigma_n}(T_n) = \lambda_i^{G_\infty,\sigma_\infty}(T) \quad \text{for all } i. \tag{3-3}$$

As we are especially interested in index and nullity, we immediately point out that (3-3) implies

$$\begin{aligned} \mathrm{ind}_{G_\infty}^{\sigma_\infty}(T) &\le \liminf_{n\to\infty} \mathrm{ind}_{G_n}^{\sigma_n}(T_n), \\ \limsup_{n\to\infty} \mathrm{nul}_{G_n}^{\sigma_n}(T_n) &\le \mathrm{nul}_{G_\infty}^{\sigma_\infty}(T), \\ \limsup_{n\to\infty} (\mathrm{ind}_{G_n}^{\sigma_n}(T_n) + \mathrm{nul}_{G_n}^{\sigma_n}(T_n)) &\le \mathrm{ind}_{G_\infty}^{\sigma_\infty}(T) + \mathrm{nul}_{G_\infty}^{\sigma_\infty}(T). \end{aligned} \tag{3-4}$$

Proposition 3.8. *Let $(\Omega, g, q, r, \partial_{\mathrm{D}}\Omega, \partial_{\mathrm{N}}\Omega, \partial_{\mathrm{R}}\Omega, G_\infty, \sigma_\infty)$ satisfy all our assumptions on admissible data except that we allow G_∞ to have infinite order; let T be the bilinear form determined by the data. Let $\{(\Omega, g_n, q_n, r_n, \partial_{\mathrm{D}}\Omega, \partial_{\mathrm{N}}\Omega, \partial_{\mathrm{R}}\Omega, G_n, \sigma_n)\}$ be a sequence of admissible data, with corresponding sequence $\{T_n\}$ of bilinear forms. Assume*

$$\sup_n \sup_\Omega (|g_n|_g + |g_n^{-1}|_g + |q_n| + |r_n|) < \infty \quad \textit{and} \quad (g_n, q_n, r_n) \xrightarrow[n\to\infty]{\textit{a.e. on } \Omega} (g, q, r).$$

Assume further that

(1) $G_n \le G_\infty$ *for all* n, *and* $\sigma_n(\phi_n) = \sigma(\phi_n)$ *for all* n *and all* $\phi_n \in G_n$;

(2) *for each* $\phi \in G_\infty$, *there exists a sequence* $\{\phi_n\}$ *such that*

  (a) $\phi_n \in G_n$ *for all* n,

  (b) $\phi_n^* \xrightarrow[n\to\infty]{} \phi^*$ *strongly as linear endomorphisms of* $L^2(\Omega, g)$,

  (c) $\sigma_n(\phi_n) = \sigma(\phi)$ *for all* n.

Then

$$\lim_{n\to\infty} \lambda_i^{G_n,\sigma_n}(T_n) = \lambda_i^{G_\infty,\sigma_\infty}(T) \quad \textit{for all } i.$$

Proof. For expository convenience, we will first focus on the case when $G_n = G_\infty$ and $\sigma_n = \sigma_\infty$ for all n, thereby implicitly assuming $(\Omega, g, q, r, \partial_{\mathrm{D}}\Omega, \partial_{\mathrm{N}}\Omega, \partial_{\mathrm{R}}\Omega, G_\infty, \sigma_\infty)$ to be admissible data (in our standard sense); we shall simply denote by G the group in question, and by σ the associated homomorphism.

Fix the index $i \ge 1$. We will start by showing that

$$\limsup_{n\to\infty} \lambda_i^{G,\sigma}(T_n) \le \lambda_i^{G,\sigma}(T). \tag{3-5}$$

For this we start with an $L^2(\Omega, g)$-orthonormal set $\{u_j\}_{j=1}^i$ such that u_j is a (G, σ)-invariant eigenfunction of T with eigenvalue $\lambda_j^{G,\sigma}(T)$. Then our assumptions on the coefficients together with the dominated convergence theorem imply that, for all $1 \le j, k \le i$,

$$\begin{aligned}
\lim_{n\to\infty} \langle u_j, u_k\rangle_{L^2(\Omega,g_n)} &= \langle u_j, u_k\rangle_{L^2(\Omega,g)},\\
\lim_{n\to\infty} \langle u_j, q_n u_k\rangle_{L^2(\Omega,g_n)} &= \langle u_j, q u_k\rangle_{L^2(\Omega,g)},\\
\lim_{n\to\infty} \int_\Omega g_n(\nabla_{g_n} u_j, \nabla_{g_n} u_k)\, d\mathscr{H}^d(g_n) &= \int_\Omega g(\nabla_g u_j, \nabla_g u_k)\, d\mathscr{H}^d(g),\\
\lim_{n\to\infty} \int_{\partial_{\mathrm{R}}\Omega} r_n u_j u_k\, d\mathscr{H}^{d-1}(g_n) &= \int_{\partial_{\mathrm{R}}\Omega} r u_j u_k\, d\mathscr{H}^{d-1}(g).
\end{aligned}$$

In conjunction with the min-max characterization (2-13) this proves (3-5). To conclude it thus suffices to prove the complementary inequality

$$\liminf_{n\to\infty} \lambda_i^{G,\sigma}(T_n) \ge \lambda_i^{G,\sigma}(T). \tag{3-6}$$

By (3-5) the sequence $\lambda_i^{G,\sigma}(T_n)$ is bounded from above uniformly in n, and by the min-max characterization (2-13) of eigenvalues along with the assumed uniform bounds on q_n and r_n and the trace inequality (2-3) it is also bounded from below. Therefore the left-hand side of (3-6) is a real number, and, by passing to a subsequence of the data if necessary (without renaming), we in fact assume without loss of generality that

$$\{\lambda_j^{G,\sigma}(T_n)\} \text{ converges to } \lambda_j^\infty \in \mathbb{R} \quad \text{for each } j \le i, \tag{3-7}$$

with λ_j^∞ the lim inf of the j-th (G, σ)-eigenvalue of the original sequence.

For each $j \leq i$ and each n, let $v_j^{(n)}$ be a (G,σ)-invariant eigenfunction of T_n with eigenvalue $\lambda_j^{G,\sigma}(T_n)$ such that, for each n, the set $\{v_j^{(n)}\}_{j=1}^i$ is $L^2(\Omega, g_n)$-orthonormal. It follows from the assumed unit $L^2(\Omega, g_n)$ bounds on the $v_j^{(n)}$, the definitions of eigenvalues and eigenfunctions, the eigenvalue bound following from (3-7), and the assumed bounds on q_n and r_n as well as g_n and g_n^{-1} that the sequence $\|v_j^{(n)}\|_{H^1(\Omega,g_n)}$ is bounded uniformly in n. (The assumptions on the metrics are needed there to ensure that the constants in the trace inequality (2-3), as applied here, can be chosen independently of n.) It then follows, in turn, using again the assumed bounds on g_n and g_n^{-1}, that $\|v_j^{(n)}\|_{H^1(\Omega,g)}$ is likewise bounded. Consequently, passing to a further subsequence if needed, for each $j \leq i$, there exists $v_j \in H^1(\Omega, g)$ which is simultaneously a limit in $L^2(\Omega, g)$ and a weak limit in $H^1(\Omega, g)$ of $v_j^{(n)}$ as $n \to \infty$. Note in particular that each v_j is (G,σ)-invariant.

The dominated convergence theorem, our assumptions on the metrics, and the $L^2(\Omega, g)$-convergence for each j of $\{v_j^{(n)}\}$ to v_j imply that $\{v_j\}_{j=1}^i$ is $L^2(\Omega, g)$-orthonormal, so in particular this finite family is linearly independent. In the same fashion, but also appealing to the assumptions on the q_n, we get, for all $1 \leq j \leq i$ and all $w \in L^2(\Omega, g)$,

$$\lim_{n\to\infty} \langle v_j^{(n)}, w\rangle_{L^2(\Omega,g_n)} = \langle v_j, w\rangle_{L^2(\Omega,g)}, \quad \lim_{n\to\infty} \langle q_n v_j^{(n)}, w\rangle_{L^2(\Omega,g_n)} = \langle q v_j, w\rangle_{L^2(\Omega,g)}.$$

Thanks to the weak convergence in $H^1(\Omega, g)$ of $\{v_j^{(n)}\}$ to v_j for each j (and again using the dominated convergence theorem, the assumptions on the metrics, and the L^2 convergence of each $\{v_j^{(n)}\}$), we further conclude that, for all $1 \leq j \leq i$ and all $w \in H^1(\Omega, g)$,

$$\lim_{n\to\infty} \int_\Omega g_n(\nabla_{g_n} v_j^{(n)}, \nabla_{g_n} w)\, d\mathscr{H}^d(g_n) = \int_\Omega g(\nabla_g v_j, \nabla_g w)\, d\mathscr{H}^d(g).$$

We use the trace inequality (2-3) in conjunction with boundedness in $H^1(\Omega, g)$ of $\{v_j^{(n)}\} \cup \{v_j\}$ and the convergence in $L^2(\Omega, g)$ for each j of $v_j^{(n)}$ to v_j to deduce that we also have $L^2(\partial\Omega, g)$-convergence of the traces. As one consequence we see that each v_j in fact belongs to $H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g)$. As another, by virtue of the assumptions on the r_n and once again the dominated convergence theorem, we obtain, for all $1 \leq j \leq i$ and $w \in H^1(\Omega, g)$,

$$\lim_{n\to\infty} \int_{\partial_{\mathrm{R}}\Omega} r_n v_j^{(n)} w\, d\mathscr{H}^{d-1}(g_n) = \int_{\partial_{\mathrm{R}}\Omega} r v_j w\, d\mathscr{H}^{d-1}(g_n).$$

From the definition of the $v_j^{(n)}$, the assumption (3-7), and the above three displayed equations, we conclude that, for all $1 \leq j \leq i$ and $w \in H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g)$, we eventually have

$$T(v_j, w) = \lim_{n\to\infty} T_n(v_j^{(n)}, w) = \lim_{n\to\infty} \lambda_j^{G,\sigma}(T_n)\langle v_j^{(n)}, w\rangle_{L^2(\Omega,g_n)} = \lambda_j^\infty \langle v_j, w\rangle_{L^2(\Omega,g)}.$$

Specifically, for the second equality above we have used the fact that $v_j^{(n)}$ is an eigenfunction of T_n; together, the inequalities then show that v_j is an eigenfunction of T. Since $\{v_j\}_{j=1}^i$ is a linearly independent subset of $\pi_{G,\sigma} H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g)$, it follows that

$$\lambda_i^\infty \geq \lambda_i^{G,\sigma}(T),$$

completing the proof in the case of a "fixed symmetry group".

However, it is actually straightforward to generalize the above argument to capture also continuity in the symmetries. The proof above goes through with mostly superficial modification, and we address the only two salient points. First, in proving (3-5) but with (G, σ) replaced on the left by (G_n, σ_n) and on the right by $(G_\infty, \sigma_\infty)$, note that each u_j, now assumed $(G_\infty, \sigma_\infty)$-invariant, is by our hypotheses also (G_n, σ_n)-invariant for each n. Second, in proving the corresponding analogue of (3-6), note that each v_j is, as the $L^2(\Omega, g)$ limit of a sequence whose n-th term is (G_n, σ_n)-invariant, by our hypotheses, itself $(G_\infty, \sigma_\infty)$-invariant. □

We now turn our attention to the related yet different problem of handling controlled changes in the domain. We switch to slightly different notation, which is again tailor-made to best fit our later applications.

Proposition 3.9. *Let $(\Omega, g, q, r, \partial_{\mathrm{D}}\Omega, \partial_{\mathrm{N}}\Omega, \partial_{\mathrm{R}}\Omega, G, \sigma)$ be admissible data, with corresponding bilinear form T. Suppose that, for any $\delta > 0$ less than the injectivity radius of (M, g), say δ_0, we are given a Lipschitz domain $\Omega_\delta \subset \Omega$ such that $(\Omega_\delta, g, q, r, \partial_{\mathrm{D}}\Omega_\delta, \partial_{\mathrm{N}}\Omega_\delta, \partial_{\mathrm{R}}\Omega_\delta, G, \sigma)$ are also admissible data (with suitable restrictions of tensors and functions tacitly understood) and whose complement $K_\delta := \Omega \setminus \Omega_\delta$ satisfies*

$$\bigcup_{p\in S} \overline{B_{f_1(\delta)}(p)} \subset K_\delta \subset \bigcup_{p\in S} \overline{B_{f_2(\delta)}(p)} \tag{3-8}$$

for some finite set of points $S \subset \overline{\Omega}$ and monotone functions $f_1, f_2 : [0, \delta_0[\to \mathbb{R}_{\geq 0}$ such that $\lim_{\delta\to 0} f_2(\delta) = 0$. Consider the sets as in (2-16) *with Ω_δ in lieu of Ω_1 as well as the associated bilinear form*

$$T^{\mathrm{D}_{\mathrm{int}}}_{\Omega_\delta} := T\big[\Omega_\delta, g, q, r, \partial^{\mathrm{D}_{\mathrm{int}}}_{\mathrm{D}}\Omega_\delta, \partial^{\mathrm{D}_{\mathrm{int}}}_{\mathrm{N}}\Omega_\delta, \partial^{\mathrm{D}_{\mathrm{int}}}_{\mathrm{R}}\Omega_\delta\big].$$

Then, for each integer $i \geq 1$,

$$\lambda_i^{G,\sigma}(T^{\mathrm{D}_{\mathrm{int}}}_{\Omega_\delta}) \geq \lambda_i^{G,\sigma}(T), \tag{3-9}$$

and we have

$$\lim_{\delta\to 0} \lambda_i^{G,\sigma}(T^{\mathrm{D}_{\mathrm{int}}}_{\Omega_\delta}) = \lambda_i^{G,\sigma}(T). \tag{3-10}$$

The conclusion simply relies on the fact that points have null $W^{1,s}$-capacity in $\mathbb{R}^n$ for $1 \leq s \leq n$ and so, in particular, have null $W^{1,2}$-capacity in $\mathbb{R}^n$ for any $n \geq 2$; for the sake of completeness, we provide a self-contained argument focusing on the case of surfaces ($d = 2$), where a logarithmic cutoff trick is required and omit the simpler modifications for $d \geq 3$.

Proof. Given any $u_\delta, v_\delta \in H^1_{\partial^{\mathrm{D}_{\mathrm{int}}}_{\mathrm{D}}\Omega_\delta}(\Omega_\delta)$, postulated to be (G, σ)-invariant, it is standard to note that their extensions by 0, say $\bar{u}_\delta, \bar{v}_\delta$ respectively, belong to $H^1_{\partial_{\mathrm{D}}\Omega}(\Omega)$, that such functions are themselves (G, σ)-invariant, and, for any $\delta \in (0, \delta_0)$ we have $\langle u_\delta, v_\delta\rangle_{L^2(\Omega_\delta, g)} = \langle \bar{u}_\delta, \bar{v}_\delta\rangle_{L^2(\Omega, g)}$ and $T^{\mathrm{D}_{\mathrm{int}}}_{\Omega_\delta}(u_\delta, u_\delta) = T(\bar{u}_\delta, \bar{u}_\delta)$. Hence, it follows at once from the variational characterization of eigenvalues, (2-13), that, for each integer $i \geq 1$, we have indeed $\lambda_i^{G,\sigma}(T^{\mathrm{D}_{\mathrm{int}}}_{\Omega_\delta}) \geq \lambda_i^{G,\sigma}(T)$, which is the first claim. Appealing again to the domain monotonicity, it actually suffices to check (3-10) in the case when K_δ is in fact a union of metric balls, namely when we have equality in (3-8), for $f_1 = f_2$. To simplify the notation we can (without loss of generality, up to reparametrization) assume in fact $f_2(\delta) = \delta$ for any δ in the assumed domain. That said,

given any $\bar{u}, \bar{v} \in H^1_{\partial_{\mathrm{D}}\Omega}(\Omega)$, (G,σ)-invariant, and $\delta > 0$ (small as in the statement), one can simply define $u_\delta = \bar{u}\varphi_\delta$ and $v_\delta = \bar{v}\varphi_\delta$, where (for $r := d_g(p,q)$ and $p \in S$) we set

$$\varphi_\delta(q) = \begin{cases} 0 & \text{if } r \le \delta^{3/4}, \\ 3 - 4\log r/\log\delta & \text{if } \delta^{3/4} \le r \le \delta^{1/2}, \\ 1 & \text{otherwise.} \end{cases}$$

It is then clear that $u_\delta, v_\delta \in H^1_{\partial^{\mathrm{D}_{\mathrm{int}}}_{\mathrm{D}}\Omega_\delta}(\Omega_\delta)$, that such functions are (G,σ)-invariant, and, in addition,

$$\lim_{\delta\to 0} T^{\mathrm{D}_{\mathrm{int}}}_{\Omega_\delta}(u_\delta, u_\delta) = T(\bar{u},\bar{u}), \quad \lim_{\delta\to 0}\langle u_\delta, v_\delta\rangle_{L^2(\Omega_\delta,g)} = \langle \bar{u},\bar{v}\rangle_{L^2(\Omega,g)}.$$

Hence, again appealing to (2-13), we must conclude

$$\limsup_{\delta\to 0} \lambda_i^{G,\sigma}(T^{\mathrm{D}_{\mathrm{int}}}_{\Omega_\delta}) \le \lambda_i^{G,\sigma}(T), \tag{3-11}$$

whence, combining this inequality with the one above, the conclusion follows. □

Corollary 3.10. *Given the setting and the assumptions of Proposition 3.9, we have*

$$\lim_{\delta\to 0} \mathrm{ind}^\sigma_G(T^{\mathrm{D}_{\mathrm{int}}}_{\Omega_\delta}) = \mathrm{ind}^\sigma_G(T).$$

3.4. ***Conformal change in dimension 2.*** In this section we suppose, in addition to the assumptions above, that $d = \dim M = 2$ and that we are given a smooth, strictly positive, G-invariant function ρ on $\overline{\Omega}$. Note that the above bilinear form T of (2-4) is invariant under scaling, namely under the simultaneous transformations $g \mapsto \rho^2 g$, $q \mapsto \rho^{-2}q$, and $r \mapsto \rho^{-1}r$:

$$T\big[\Omega, \rho^2 g, \rho^{-2}q, \rho^{-1}r, \partial_{\mathrm{D}}\Omega, \partial_{\mathrm{N}}\Omega, \partial_{\mathrm{R}}\Omega\big] = T\big[\Omega, g, q, r, \partial_{\mathrm{D}}\Omega, \partial_{\mathrm{N}}\Omega, \partial_{\mathrm{R}}\Omega\big],$$

with the corresponding domains $H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, \rho^2 g)$ and $H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g)$ agreeing as sets of functions and having equivalent norms. This claim needs a clarification: the standard H^1 norms of $H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, \rho^2 g)$ and $H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g)$ are only equivalent up to constants that depend on the extremal (inf and sup) values of the conformal factor ρ.

In general, the eigenvalues (as defined in Section 2.3) will be affected by the conformal scaling, and yet the index and nullity are nonetheless invariant when this operation is performed.

Proposition 3.11 (invariance of index and nullity under conformal change in dimension 2). *With assumptions as in the preceding paragraph,*

$$\mathrm{ind}^\sigma_G(T, \rho^2 g) = \mathrm{ind}^\sigma_G(T, g) \quad \textit{and} \quad \mathrm{nul}^\sigma_G(T, \rho^2 g) = \mathrm{nul}^\sigma_G(T, g).$$

Proof. By definition, we have that $u \in E^{=0}_{G,\sigma}(T, g)$ if and only if u is (G,σ)-invariant and $T(u, v) = 0$ for all $v \in H^1_{\partial_{\mathrm{D}}\Omega}(\Omega, g)$ (and likewise if each g is replaced by $\rho^2 g$), so the nullity equality is clear. For the index, because we can reverse the roles of g and $\rho^2 g$ by replacing ρ with ρ^{-1}, it suffices to check that the claim holds with $\ge$ in place of $=$. This follows at once from the min-max characterization (2-13) applied to the (G,σ)-eigenvalues of $(T, \rho^2 g)$, by considering the "competitor" subspace $E^{<0}_{G,\sigma}(T, g)$ in the minimization problem therein, for $i = \mathrm{ind}^\sigma_G(T, g)$. □

4. Free boundary minimal surfaces in the ball: a first application

From now on, we specialize our study to the case when $\overline{\Omega} = M$ is a properly embedded free boundary minimal surface, henceforth denoted by Σ, of the closed unit ball $\mathbb{B}^3 := \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 \leq 1\}$ in Euclidean space $(\mathbb{R}^3, g^{\mathbb{R}^3})$. Observe that, by the maximum principle, every embedded free boundary minimal surface is properly embedded.

As anticipated in the introduction, our task here will be to obtain quantitative estimates on the Morse index of free boundary minimal surfaces; hence our Schrödinger operator is the Jacobi (or stability) operator on Σ acting on functions subject to the Robin condition

$$du(\eta^{\Sigma}_{g^{\mathbb{R}^3}}) = u \quad \text{on } \partial\Sigma, \tag{4-1}$$

namely: $q = |A^{\Sigma}|^2$, the squared norm of the second fundamental form of Σ, and $\partial_{\mathrm{D}}\Sigma = \partial_{\mathrm{N}}\Sigma = \varnothing$, $\partial_{\mathrm{R}}\Sigma = \partial\Sigma$, $r = 1$. Correspondingly, as our bilinear form T we will consider the index (or stability or Jacobi) form of Σ, which we will denote by Q^{Σ}. We define the index and nullity of Σ in the usual way, setting

$$\mathrm{ind}(\Sigma) := \mathrm{ind}(Q^{\Sigma}) \quad \text{and} \quad \mathrm{nul}(\Sigma) := \mathrm{nul}(Q^{\Sigma}),$$

and we likewise define the G-equivariant index and nullity of Σ, $\mathrm{ind}_G(\Sigma)$ and $\mathrm{nul}_G(\Sigma)$, in the sense of (2-15), when given a group $G < \mathrm{O}(3)$ of symmetries of Σ one considers the associated sign homomorphism. More generally, we will also study the (G, σ)-index and (G, σ)-nullity of Σ, $\mathrm{ind}^{\sigma}_G(\Sigma)$ and $\mathrm{nul}^{\sigma}_G(\Sigma)$, when given a group G and, further, a homomorphism $\sigma : G \to \mathrm{O}(1)$ (thus, in either case, these expressions are to be understood by replacing Σ by Q^{Σ}).

It has already been mentioned above how general lower bounds for the index, linear in the topological data (genus and number of boundary components), have been obtained in [Ambrozio et al. 2018b] and by [Sargent 2017] in the special case when the ambient manifold is a convex body in Euclidean $\mathbb{R}^3$. We begin this section by presenting an alternative lower bound (Proposition 4.2 below) in terms of symmetries, which, though much less general in nature, nevertheless yields sharper lower bounds for many of the known examples (in terms of the coefficients describing the linear growth rate as a function of the topological data). Before proceeding, we pause to explain some notation we will find convenient.

Cylindrical coordinates and wedges. We shall describe points in Euclidean $\mathbb{R}^3$, endowed with standard Cartesian coordinates (x, y, z) and also in terms of cylindrical coordinates (r, θ, z), so that the point with cylindrical coordinates (r_0, θ_0, z_0) has Cartesian coordinates $(x, y, z) = (r_0 \cos\theta_0, r_0 \sin\theta_0, z_0)$. However we wish to stress that, for our purposes, it will be convenient to allow arbitrary real values for both r and θ; thus the triples (r, θ, z) and $(-r, \theta + \pi, z)$ describe the same point in Euclidean space. Given real numbers $\alpha \leq \beta$, we also define the closed wedge

$$W^{\beta}_{\alpha} := \{(r\cos\theta, r\sin\theta, z) : r \geq 0,\ \theta \in [\alpha, \beta],\ z \in \mathbb{R}\}, \tag{4-2}$$

with the half-plane W^{α}_{α} accommodated as a degenerate wedge. In particular, our convention implies

$$\{\theta = \alpha\} = W^{\alpha}_{\alpha} \cup W^{\alpha+\pi}_{\alpha+\pi}.$$

Notation for symmetries. Given a plane $\Pi \subset \mathbb{R}^3$ through the origin, we write $\underline{\mathsf{R}}_\Pi \in \mathrm{O}(3)$ for reflection through Π. Similarly, given a directed line $\xi \subset \mathbb{R}^3$ through the origin and an angle $\theta \in \mathbb{R}$, we write R_ξ^θ for rotation about ξ through angle α in the usual right-handed sense. Typically we will be interested not exclusively in such a rotation R_ξ^θ but rather in the cyclic subgroup it generates, with the result that it will never really be important to associate a direction to ξ. Given symmetries $\mathsf{T}_1, \dots, \mathsf{T}_n \in \mathrm{O}(3)$, we write $\langle \mathsf{T}_1, \dots, \mathsf{T}_n \rangle$ for the subgroup they generate.

The order-2 groups generated by reflections through planes will figure repeatedly in the sequel (beginning with the following proposition), so for succinctness of notation, given a plane $\Pi \subset \mathbb{R}^3$ through the origin, we agree to set $\boldsymbol{\Pi} := \langle \underline{\mathsf{R}}_\Pi \rangle$. In such a context, consistent with the general convention we defined above, we will employ the apex $+$ (resp. $-$) to denote functions that are even (resp. odd) with respect to the reflection through Π. Similarly (but less frequently), if ξ is a line through the origin in $\mathbb{R}^3$, we will write $\boldsymbol{\xi}$ for the order-2 group generated by reflection $\underline{\mathsf{R}}_\xi$ through ξ (equivalently rotation through angle π in either sense about ξ).

We also pause to name the following three subgroups of O(3), which will be realized as subgroups of the symmetry groups of the examples we study below and which partly pertain to the statement of the next proposition: for each integer $k \geq 1$, we set

$$\begin{aligned}
\mathbb{Y}_k &:= \langle \underline{\mathsf{R}}_{\{\theta=-\pi/(2k)\}}, \underline{\mathsf{R}}_{\{\theta=\pi/(2k)\}} \rangle && \text{(pyramidal group of order } 2k\text{)}, \\
\mathbb{P}_k &:= \langle \underline{\mathsf{R}}_{\{\theta=-\pi/(2k)\}}, \underline{\mathsf{R}}_{\{\theta=\pi/(2k)\}}, \underline{\mathsf{R}}_{\{z=0\}} \rangle && \text{(prismatic group of order } 4k\text{)}, \\
\mathbb{A}_k &:= \langle \underline{\mathsf{R}}_{\{\theta=\pi/(2k)\}}, \mathsf{R}^\pi_{\{y=z=0\}} \rangle && \text{(antiprismatic group of order } 4k\text{)}.
\end{aligned} \tag{4-3}$$

Note in particular that we have $\mathbb{Y}_k = \mathbb{P}_k \cap \mathbb{A}_k$.

Remark 4.1. The above three groups are so named because they are the (maximal) symmetry groups of, respectively, a right pyramid, prism, or antiprism over a regular k-gon. See, e.g., Section 2 of [Carlotto et al. 2022b] for pictures and further details, but we caution that the above definition of the subgroup $\mathbb{P}_k$ differs slightly from that given in [Carlotto et al. 2022b]: the two subgroups are conjugate to one another via rotation through angle $\pi/(2k)$ about the z-axis.

With this terminology and notation in place, we proceed with the aforementioned lower index bound, which illustrates the Montiel–Ros methodology as developed in Section 3 and is interesting in its own right.

Proposition 4.2 (index lower bounds under pyramidal and prismatic symmetry; see [Choe 1990; Kapouleas and Wiygul 2020]). *Let Σ be a connected, embedded free boundary minimal surface in $\mathbb{B}^3$. Assume that Σ is not a disc or critical catenoid, that Σ is invariant under reflection through a plane Π_1, and that Σ is also invariant under rotation through an angle $\alpha \in]0, 2\pi[$ about a line $\xi \subset \Pi_1$. Then α is a rational multiple of 2π, there is a largest integer $k \geq 2$ such that rotation about ξ through angle $2\pi/k$ is also a symmetry of Σ, and*

(a) $\mathrm{ind}(\Sigma) \geq 2k - 1$,

(b) $\mathrm{ind}^-_{\boldsymbol{\Pi}_1}(\Sigma) \geq k - 1$, *and*

(c) *if Σ is additionally invariant under reflection through a plane $\Pi_\perp$ orthogonal to ξ, then in fact* $\mathrm{ind}^+_{\boldsymbol{\Pi}_\perp}(\Sigma) \geq 2k - 1$.

Note that the symmetries assumed in the preamble of Proposition 4.2 generate, up to conjugacy in O(3), the group $\mathbb{Y}_k$ from (4-3), while one instead obtains (again up to conjugacy) the group $\mathbb{P}_k$ by adjoining the additional symmetry assumed in item (c).

The proof below is an abstraction and transplantation to the free boundary setting of some index lower bounds obtained in the course of [Kapouleas and Wiygul 2020] and drawing on ideas from [Montiel and Ros 1991]. The estimates ultimately depend on a lower bound on the number of nodal domains of a suitable Jacobi field, which was also the basis for earlier index estimates (of complete minimal surfaces in $\mathbb{R}^3$ and closed minimal surfaces in $\mathbb{S}^3$) established in [Choe 1990].

Proof. By excluding the discs and critical catenoids we ensure that Σ is not $\mathbb{S}^1$-invariant about ξ, implying the claim on α and the existence of the rotational symmetry about ξ through angle of the form $2\pi/k$, as follows. First, if the cyclic subgroup generated by rotation about ξ through angle α were not finite, then it would be dense in the SO(2) subgroup of rotations about ξ, but the symmetry group of Σ is closed in O(3); yet, as already observed, our assumptions ensure that Σ has no SO(2) symmetry subgroup. Thus α must be a rational multiple of 2π, as claimed. Now let β be the least angle in $]0, 2\pi[$ through which rotation about ξ is generated by the assumed rotational symmetry through angle α, and let k be the least positive integer such that $k\beta \geq 2\pi$. Then rotation through angle $k\beta - 2\pi$, which lies in $[0, \beta[$, is also generated by the assumed rotational symmetry. The presumed minimality of β then forces $\beta = 2\pi/k$.

By composing the assumed symmetries, it follows that Σ is also invariant under reflection through each of the $k-1$ planes $\Pi_2, \dots, \Pi_k$ containing ξ and there meeting Π_1 at angle an integer multiple of π/k. Now suppose $\Pi \in \{\Pi_i\}_{i=1}^k$. We necessarily have $\Pi \cap \Sigma \neq \varnothing$ (for example since Π separates $\mathbb{B}^3$ into two components and is a plane of symmetry for Σ, which is assumed to be connected). Because Π is a plane of symmetry and Σ is embedded, these two surfaces must intersect either orthogonally or tangentially, but in the latter case Σ must be a disc, which possibility we have excluded by assumption; consequently, the intersection is orthogonal. Moreover, by the symmetries each of the $2k$ components $W_1, \dots, W_{2k}$ of $\mathbb{B}^3 \setminus \bigcup_{i=1}^k \Pi_i$ then has nontrivial intersection $\Omega_i := \Sigma \cap W_i$ with Σ. Without loss of generality, let us agree to label the domains under consideration in counterclockwise order such that $\Omega_1, \dots, \Omega_k$ all lie on the same side of Π_1.

Note that the members of the family $\{\Omega_i\}_{i=1}^{2k}$ are pairwise isometric and each is connected. (Indeed, Σ is itself connected, so any two points in any single Ω_i can be joined by some path in Σ, but this path can leave Ω_i only through the latter's intersection with planes of symmetry, so we can always produce a path connecting the two points that is entirely contained in Ω_i by repeated reflection and replacement, if necessary.) Furthermore, each Ω_i has Lipschitz boundary contained in $\mathbb{S}^2 \cup \bigcup_{i=1}^k \Pi_i$ because the intersection of Σ with either $\mathbb{S}^2$ and any of the planes $\Pi_1, \dots, \Pi_k$ is orthogonal (thus transverse) and exactly k of the Ω_i lie on each side of Π_1.

Next, letting κ_ξ be a choice of (scalar-valued) Jacobi field on Σ induced by the rotations about ξ and again using the fact that Σ is not rotationally symmetric (and so, in particular, not planar either), we conclude that κ_ξ vanishes on $\Sigma \cap \bigcup_{i=1}^k \Pi_i$ (because of the aforementioned orthogonality) but does not vanish identically on any Ω_i. As a result, imposing, for each i, the Robin condition (4-1) on $\mathbb{S}^2 \cap \partial\Omega_i$ and the Dirichlet condition on $\partial\Omega_i \cap \bigcup_{i=1}^k \Pi_i$, the corresponding nullity of Ω_i is at least 1. An appeal to

item (i) of Corollary 3.2 (for claims (a) and (c)) and to item (i) of Proposition 3.1 (for (b)) now completes the proof. Specifically:

- for (a), we consider the partition of Σ into the $2k$ domains $\Omega_1, \dots, \Omega_{2k}$ and take G to be the trivial group;
- for (b), we take $G = \langle \underline{R}_{\Pi_1} \rangle$ to be the group with two elements (as in Example 2.3), the homomorphism determined by $\sigma(\underline{R}_{\Pi_1}) = -1$ (thereby imposing *odd* symmetry) and, correspondingly, we consider the partition of Σ into k domains obtained by equivariant pairing, i.e., by taking $\Omega_{i+1} \cup \Omega_{2k-i}$ for $i = 0, \dots, k-1$;
- for (c), we consider the partition of Σ into the $2k$ domains $\Omega_1, \dots, \Omega_{2k}$, take $G = \langle \underline{R}_{\Pi_\perp} \rangle$ to be the group with two elements, and the homomorphism determined by $\sigma(\underline{R}_{\Pi_\perp}) = +1$ (thereby imposing *even* symmetry).

Thereby the proof is complete. □

5. Effective index estimates for two sequences of examples

5.1. ***Review of the construction and lower index bounds.*** As we have already alluded to in the introduction, in [Carlotto et al. 2022b] two families of embedded free boundary minimal surfaces in $\mathbb{B}^3$ were constructed by desingularizing (in the spirit of [Kapouleas 1997]) the configurations $-\mathbb{k}_0 \cup \mathbb{k}_0$ and $-\mathbb{k}_0 \cup \mathbb{B}^2 \cup \mathbb{k}_0$, where $\mathbb{k}_0$ is the intersection with $\mathbb{B}^3$ of a certain catenoid having axis of symmetry $\{x = y = 0\}$ and meeting $\partial\mathbb{B}^3$ (not orthogonally) along the equator $\partial\mathbb{B}^2$ and orthogonally along one additional circle of latitude at height $h > 0$.

Proposition 5.1 (existence and basic properties of $\mathbb{k}_0$). *There exists a minimal annulus $\mathbb{k}_0$ which is properly embedded in $\mathbb{B}^3$ and intersects the unit sphere $\partial\mathbb{B}^3$ exactly along the equator $\partial_0\mathbb{k}_0 := \partial\mathbb{B}^3 \cap \{z = 0\}$ and orthogonally along a circle of latitude at height $z = h \approx 0.87028$, which we denote by*

$$\partial_\perp \mathbb{k}_0 := \partial\mathbb{k}_0 \setminus \partial_0\mathbb{k}_0.$$

Moreover, $\mathbb{k}_0$ coincides with the surface of revolution of the graph of $r : [0, h] \to]0, 1[$ given by $r(\zeta) = (1/a)\cosh(a\zeta - s)$ for suitable $a \approx 2.3328$ and $s \approx 1.4907$.

Proof. The existence of $\mathbb{k}_0$ is proven in [Carlotto et al. 2022b, Lemma 3.3]. For the numerical values of a, h, and s, we refer to [loc. cit., Remark 3.9]. □

That being said, these are (somewhat simplified) versions of the main existence results we proved in [loc. cit.].

Theorem 5.2 (desingularizations of $-\mathbb{k}_0 \cup \mathbb{k}_0$ [Carlotto et al. 2022b]). *For each sufficiently large integer n, there exists in $\mathbb{B}^3$ a properly embedded free boundary minimal surface $\Xi_n^{-\mathbb{k}_0 \cup \mathbb{k}_0}$ that has genus 0, exactly $n+2$ boundary components, and is invariant under the prismatic group $\mathbb{P}_n$ from* (4-3). *Moreover, $\Xi_n^{-\mathbb{k}_0 \cup \mathbb{k}_0}$ converges to $-\mathbb{k}_0 \cup \mathbb{k}_0$ in the sense of varifolds, with unit multiplicity, and smoothly away from the equator, as $n \to \infty$.*

Theorem 5.3 (desingularizations of $-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0$ [Carlotto et al. 2022b]). *For each sufficiently large integer m, there exists in $\mathbb{B}^3$ a properly embedded free boundary minimal surface $\Sigma_m^{-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0}$ that has genus m, exactly three boundary components and is invariant under the antiprismatic group $\mathbb{A}_{m+1}$ from* (4-3). *Moreover, $\Sigma_m^{-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0}$ converges to $-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0$ in the sense of varifolds, with unit multiplicity, and smoothly away from the equator, as $m \to \infty$.*

Proposition 5.4 (lower bounds by symmetry on the index of the examples of [Carlotto et al. 2022b]). *There exist $n_0, m_0 > 0$ such that we have the following index estimates for all integers $n > n_0$ and $m > m_0$:*

$$\operatorname{ind}^+_{\{z=0\}}(\Xi_n^{-\mathbb{K}_0 \cup \mathbb{K}_0}) \geq 2n - 1 \quad \textit{and} \quad \operatorname{ind}(\Sigma_m^{-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0}) \geq 2m + 1.$$

Proof. As stated in Theorem 5.2, $\Xi_n^{-\mathbb{K}_0 \cup \mathbb{K}_0}$ is invariant under the action of the prismatic group $\mathbb{P}_n$, which is generated by the reflections through the vertical planes $\{\theta = -\pi/(2n)\}$ and $\{\theta = \pi/(2n)\}$ and through the horizontal plane $\{z = 0\}$. As a composition of the first two reflections, $\mathbb{P}_n$ also contains the rotation by angle $2\pi/n$ about the vertical axis $\xi_0 = \{r = 0\}$. Applying Proposition 4.2 (c) with $k = n$, $\xi = \xi_0$, $\Pi_1 = \{\theta = \pi/(2n)\}$, and $\Pi_\perp = \{z = 0\}$, we obtain

$$\operatorname{ind}^+_{\{z=0\}}(\Xi_n^{-\mathbb{K}_0 \cup \mathbb{K}_0}) \geq 2n - 1.$$

Similarly, Theorem 5.3 states that $\Sigma_m^{-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0}$ is invariant under the action of the antiprismatic group $\mathbb{A}_{m+1}$, which contains the reflection through the vertical plane $\{\theta = \pi/(2(m+1))\}$ and also the rotation by angle $2\pi/(m+1)$ about the vertical axis ξ_0. Applying Proposition 4.2 (a) then yields

$$\operatorname{ind}(\Sigma_m^{-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0}) \geq 2m + 1. \qquad \square$$

In terms of topological data, the previous proposition (compared to [Ambrozio et al. 2018b]) provides a coefficient 2 for the growth rate of the Morse index of $\Xi_n^{-\mathbb{K}_0 \cup \mathbb{K}_0}$ (resp. $\Sigma_m^{-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0}$) with respect to the number of boundary components (resp. of the genus), modulo an additive term. In fact, the lower bound on the Morse index of $\Xi_n^{-\mathbb{K}_0 \cup \mathbb{K}_0}$ can be further improved via the following observation, which pertains to the *odd* contributions to the index instead (again with respect to reflections across the $\{z = 0\}$ plane in $\mathbb{R}^3$); incidentally this is also an example of an application of Proposition 3.1 to a collection of domains that are *not* pairwise isometric.

Proposition 5.5. *There exists $n_0 > 0$ such that we have the following index estimates for all integers $n > n_0$:*

$$\operatorname{ind}^-_{\{z=0\}}(\Xi_n^{-\mathbb{K}_0 \cup \mathbb{K}_0}) \geq 3.$$

Proof. Let Π_1 denote a vertical plane of symmetry, passing through the origin, of the surface $\Xi_n^{-\mathbb{K}_0 \cup \mathbb{K}_0}$ (which, we recall, has prismatic symmetry $\mathbb{P}_n$), let ξ be the line obtained as the intersection of such a plane with $\{z = 0\}$, and let finally $\Pi_2 = \xi^\perp$ be the vertical plane, again passing through the origin, that is orthogonal to Π_1. Consider on $\Xi_n^{-\mathbb{K}_0 \cup \mathbb{K}_0}$ the function $\kappa_\xi = K_\xi \cdot \nu$, where K_ξ is the Killing vector field associated to rotations around ξ (oriented either way) and ν is a choice of the unit normal to the surface in question. Clearly, the flow of K_ξ generates a curve of free boundary minimal surfaces around $\Xi_n^{-\mathbb{K}_0 \cup \mathbb{K}_0}$, hence the function κ_ξ lies in the kernel of the Jacobi operator of $\Xi_n^{-\mathbb{K}_0 \cup \mathbb{K}_0}$ and satisfies the

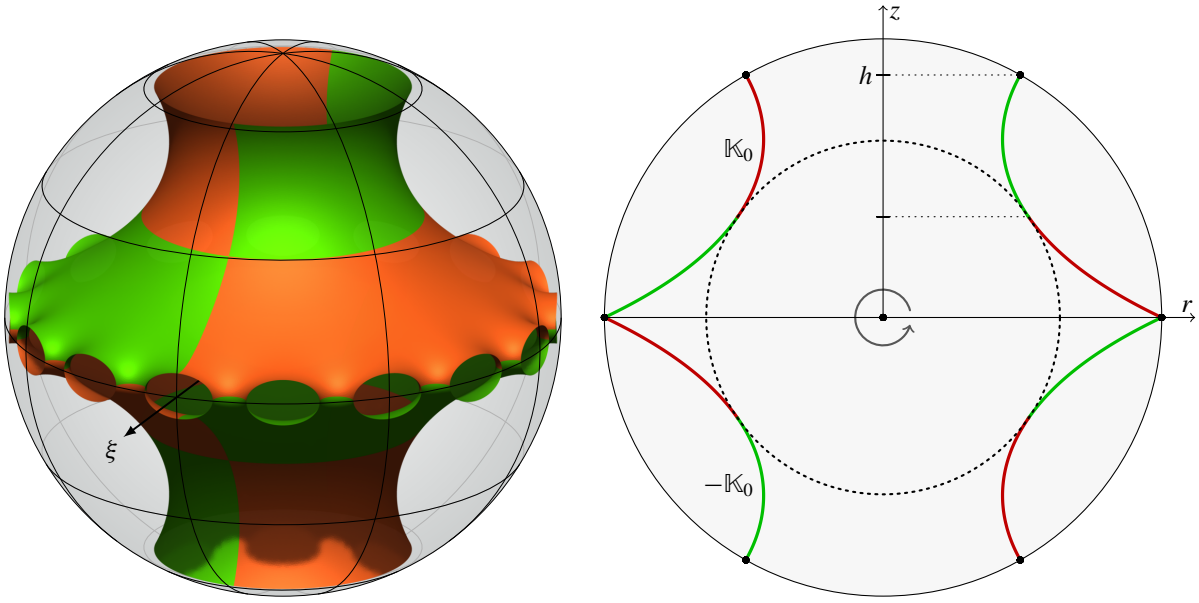

Figure 2. Nodal domains of the function induced by rotations around the symmetry axis ξ.

natural Robin boundary condition along the free boundary. Furthermore, concerning its nodal set, we first note it contains the curves

$$\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0} \cap \{z = 0\} \quad \text{and} \quad \Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0} \cap \Pi_1.$$

We also claim that, for any sufficiently large n, the function κ_ξ changes sign along the connected arc

$$\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0} \cap \Pi_2^+ \cap \{z \geq z_0\}, \tag{5-1}$$

where Π_2^+ denotes either of the half-planes determined by Π_1 on Π_2 and $z_0 > 0$ is any sufficiently small value (as we are about to describe, stressing that we can choose it independently of n). Since one has smooth convergence of $\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}$ to $-\mathbb{K}_0 \cup \mathbb{K}_0$ as $n \to \infty$ away from the equator, it suffices to verify an analogous claim for $\mathbb{K}_0$. In fact, it then follows from an explicit calculation that the function induced by rotations around the symmetry axis ξ (the analogue of κ_ξ on $\mathbb{K}_0$) has opposite signs on the two endpoints of the arc $\mathbb{K}_0 \cap \Pi_2^+$ (see Figure 2, right image), and so — assuming without loss of generality it is negative on the equatorial point — by continuity there exists $\bar{z}_0 > 0$ such that the same function is also strictly negative at all points of $\mathbb{K}_0 \cap \Pi_2^+$ at height $z_0 \in [0, \bar{z}_0]$. In particular, we can indeed choose one such value $z_0 \in (0, \bar{z}_0)$ once and for all.

Hence, appealing to the aforementioned smooth convergence, by the intermediate value theorem for any sufficiently large n, there must be a point along the arc (5-1) where κ_ξ vanishes. Now, standard results about the structure of the nodal sets of eigenfunctions of Schrödinger operators ensure that such a zero is not isolated, but is either a regular point of a smooth curve or a branch point out of which finitely many smooth arcs emanate. In either case, combining all facts above we must conclude that on $\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0} \cap \{z \geq 0\}$ the function κ_ξ has at least four nodal domains, and thus an application of Proposition 3.1 with $t = 0$, $G = \langle \underline{\mathrm{R}}_\Pi \rangle$ for $\Pi = \{z = 0\}$, and $\sigma(\underline{\mathrm{R}}_\Pi) = -1$ ensures the conclusion. □

Remark 5.6. Note that the very same argument would lead, when applied with no equivariance constraint at all (i.e., when G is the trivial group) to the conclusion that, for any sufficiently large n, the index of $\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}$ is bounded from below by 7, which however is a lot worse than the bound provided by combining Proposition 5.4 with Proposition 5.5. Furthermore, we note that one can show that the function κ_ξ has *exactly* 8 nodal domains and not more, as visualized in Figure 2.

Remark 5.7. Concerning the sharpness of the estimate given in Proposition 5.5, we note that numerical simulations of $\mathbb{K}_0$ with fixed lower boundary $\partial_0\mathbb{K}_0$ and upper boundary $\partial_\perp\mathbb{K}_0$ constrained to the unit sphere indicate that it has in fact index equal to 3. Roughly speaking, one negative direction comes from "pinching" the catenoidal neck and the other two negative directions correspond to "translations" of $\partial_\perp\mathbb{K}_0$ on the northern hemisphere.

The rest of this section is aimed at obtaining *upper* bounds on the Morse index of our examples, which is a more delicate task and one that relies crucially not only on the symmetries of the surfaces in question but also on the way they were actually constructed (which we encode in suitable convergence results).

5.2. ***Equivariant index and nullity of the models.*** For upper bounds we will exploit the regionwise convergence of the two families to the models glued together in their construction. Therefore we first study the index and nullity on these models.

Equivariant index and nullity of $\mathbb{K}_0$. We begin with a summary of the properties of the minimal annulus $\mathbb{K}_0$ we will need. Let $\partial_0\mathbb{K}_0 = \partial\mathbb{K}_0 \cap \{z = 0\}$ and $\partial_\perp\mathbb{K}_0 = \partial\mathbb{K}_0 \setminus \partial_0\mathbb{K}_0$ be as in Proposition 5.1, so that $\partial_\perp\mathbb{K}_0$ is the boundary component along which $\mathbb{K}_0$ meets the sphere $\partial\mathbb{B}^3$ orthogonally. Referring to (2-4), we define

$$Q_{\mathrm{N}}^{\mathbb{K}_0} := T\big[\mathbb{K}_0,\, g^{\mathbb{K}_0},\, q := |A^{\mathbb{K}_0}|^2,\, r := 1,\, \partial_{\mathrm{D}}\mathbb{K}_0 := \varnothing,\, \partial_{\mathrm{N}}\mathbb{K}_0 := \partial_0\mathbb{K}_0,\, \partial_{\mathrm{R}}\mathbb{K}_0 := \partial_\perp\mathbb{K}_0\big]$$

(where we abuse notation in that by $\mathbb{K}_0$ we really mean its topological interior) to be the Jacobi form of $\mathbb{K}_0$ subject to the natural geometric Robin condition (4-1) on $\partial_\perp\mathbb{K}_0$ and to the Neumann condition on $\partial_0\mathbb{K}_0$. Clearly, for each $k \geq 1$, the pyramidal group $\mathbb{Y}_k$ from (4-3) preserves $\mathbb{K}_0$ and each of its boundary components individually.

Lemma 5.8 ($\mathbb{Y}_k$-equivariant index and nullity of $\mathbb{K}_0$). *With notation as above, for each sufficiently large integer* k,

$$\mathrm{ind}_{\mathbb{Y}_k}(Q_{\mathrm{N}}^{\mathbb{K}_0}) = 1 \quad \textit{and} \quad \mathrm{nul}_{\mathbb{Y}_k}(Q_{\mathrm{N}}^{\mathbb{K}_0}) = 0.$$

Proof. We shall start by recalling [Carlotto et al. 2022b, Lemma 4.4], which states that when imposing the Dirichlet condition on $\partial_0\mathbb{K}_0$ and the Robin condition on $\partial_\perp\mathbb{K}_0$, then the Jacobi operator acting on $\mathbb{Y}_k$-equivariant functions on $\mathbb{K}_0$ is invertible provided that k is sufficiently large, which means that the equivariant nullity vanishes in this case. Considering that the coordinate function $u = z$ on $\mathbb{K}_0$, which is harmonic, satisfies the Dirichlet condition on $\partial_0\mathbb{K}_0$ and the Robin condition on $\partial_\perp\mathbb{K}_0$, it is also evident that the equivariant index is at least 1 in this case (see [loc. cit., Lemma 7.2]). This implies that when instead the Neumann condition is imposed on $\partial_0\mathbb{K}_0$, the equivariant index is again at least 1. Below we prove that it is exactly 1 and the equivariant nullity is exactly 0 in the Neumann case by showing that the second eigenvalue is strictly positive. (We note here, incidentally, that this information also proves that a posteriori the equivariant index is also exactly 1 in the case that a Dirichlet condition is imposed on $\partial_0\mathbb{K}_0$.)

Let $a, h, s > 0$ and $r(\zeta) = (1/a)\cosh(a\zeta - s)$ be as in Proposition 5.1. In particular, we have $(r')^2 + 1 = \cosh^2(a\zeta - s)$. Thus, when $\mathbb{K}_0$ is parametrized as a surface of revolution in terms of the coordinates (θ, ζ) with profile function $r(\zeta)$, the metric $g_{\mathbb{K}_0}$ and the squared norm of the second fundamental form $A_{\mathbb{K}_0}$ on $\mathbb{K}_0$ are given by

$$g_{\mathbb{K}_0} = ((r')^2 + 1)\, d\zeta^2 + r^2\, d\theta^2, \quad |A_{\mathbb{K}_0}|^2 = \frac{(-r'')^2}{((r')^2+1)^3} + \frac{1}{((r')^2+1)^2 r^2} = \frac{a^2 + a^{-2}}{\cosh^4(a\zeta - s)}.$$

The outward unit conormal along $\partial_\perp \mathbb{K}_0 = \mathbb{K}_0 \cap \{\zeta = h\}$ is given by

$$\eta_{\mathbb{K}_0} = \frac{1}{\sqrt{(r')^2(h)+1}}\partial_\zeta = \frac{1}{\cosh(ah - s)}\partial_\zeta = \frac{1}{ar(h)}\partial_\zeta.$$

Assume, for the sake of a contradiction, that $\lambda_2 = \lambda_2^{\mathbb{Y}_k, \mathrm{sgn}} \leq 0$, where we are considering the spectrum of the Jacobi operator of $\mathbb{K}_0$ acting on $\mathbb{Y}_k$-equivariant functions (see Example 2.4), and subject to the boundary conditions described above. Then, by first invoking the Courant nodal domain theorem as in the proof of [Carlotto et al. 2022b, Lemma 4.4] we may assume that the associated eigenfunction u_2 is rotationally symmetric provided that k is sufficiently large, i.e., u_2 only depends on ζ and not on θ.

That said, let u be a function on $\mathbb{K}_0$ which is rotationally symmetric, i.e., constant in θ. Then

$$\Delta_{\mathbb{K}_0} u = \frac{1}{\cosh^2(a\zeta - s)}\frac{\partial^2 u}{\partial \zeta^2},$$

and we shall consider the Jacobi operator $J = \Delta_{\mathbb{K}_0} + |A_{\mathbb{K}_0}|^2$ and the eigenvalue problem

$$\begin{cases} Ju = -\lambda u, & \\ u'(0, \cdot) = 0 & \text{(Neumann condition on } \partial_0 \mathbb{K}_0), \\ u'(h, \cdot) = \cosh(ah - s)u(h, \cdot) & \text{(Robin condition on } \partial_\perp \mathbb{K}_0). \end{cases}$$

Since u_2 must change sign, there exists $z_0 \in]0, h[$ such that $u_2(z_0) = 0$. Multiplying the eigenvalue equation

$$\frac{\partial^2 u_2}{\partial \zeta^2} + \frac{a^2 + a^{-2}}{\cosh^2(a\zeta - s)} u_2 = -\lambda_2 u_2 \cosh^2(a\zeta - s) \tag{5-2}$$

with u_2 and integrating from $\zeta = 0$ to $\zeta = z_0$, we obtain

$$\int_0^{z_0} -\lambda_2 u_2^2 \cosh^2(a\zeta - s)\, d\zeta = -\int_0^{z_0} |u_2'|^2\, d\zeta + \int_0^{z_0} \frac{a^2 + a^{-2}}{\cosh^2(a\zeta - s)} u_2^2\, d\zeta.$$

Since $u(z_0) = 0$, we can obtain the Poincaré-type inequality

$$\int_0^{z_0} |u_2(\zeta)|^2\, d\zeta = \int_0^{z_0} \left| \int_{z_0}^{\zeta} u_2'(t)\, dt \right|^2 d\zeta \leq \int_0^{z_0} (z_0 - \zeta) \int_\zeta^{z_0} |u_2'(t)|^2\, dt\, d\zeta \leq \frac{z_0^2}{2} \int_0^{z_0} |u_2'(\zeta)|^2\, d\zeta.$$

Hence,

$$\int_0^{z_0} -\lambda_2 u_2^2 \cosh^2(a\zeta - s)\, d\zeta \leq \int_0^{z_0} \left(\frac{a^2 + a^{-2}}{\cosh^2(a\zeta - s)} - \frac{2}{z_0^2} \right) u_2^2\, d\zeta.$$

The right-hand side is negative if

$$z_0 < \sqrt{\frac{2}{a^2 + a^{-2}}} \approx 0.5962,$$

and so, in this case, we conclude $\lambda_2 > 0$, a contradiction.

Integrating the eigenvalue equation (5-2) instead from $\zeta = z_0$ to $\zeta = h$ and recalling the Robin condition $u'(h) = \cosh(ah - s)u(h)$ along $\partial_\perp \mathbb{K}_0$, we obtain the alternative estimate

$$\begin{aligned}\int_{z_0}^{h} -\lambda_2 u_2^2 \cosh^2(a\zeta - s)\, d\zeta &= |u_2(h)|^2 \cosh(ah - s) - \int_{z_0}^{h} |u_2'|^2\, d\zeta + \int_{z_0}^{h} \frac{a^2 + a^{-2}}{\cosh^2(a\zeta - s)} u_2^2\, d\zeta \\ &\le ((h - z_0)\cosh(ah - s) - 1) \int_{z_0}^{h} |u_2'|^2\, d\zeta + \int_{z_0}^{h} \frac{a^2 + a^{-2}}{\cosh^2(a\zeta - s)} u_2^2\, d\zeta \\ &\le \left(a^2 + a^{-2} + \frac{2}{(h - z_0)^2}((h - z_0)\cosh(ah - s) - 1)\right) \int_{z_0}^{h} u_2^2\, d\zeta,\end{aligned}$$

provided that $(h - z_0)\cosh(ah - s) - 1 < 0$. Now the right-hand side is negative if $z_0 > 0.4443$.

Since the intervals $[0, 0.5962]$ and $[0.4443, h]$ intersect, we anyway obtain a contradiction. Thus, we confirm the claim $\lambda_2 > 0$, as desired. □

Observing (as we have already done in the previous proof) that any eigenfunction "generating" the index in Lemma 5.8 is rotationally invariant, we have the following obvious corollary (which in fact can conversely be used to prove the lemma, with the aid of Proposition 3.8). In the statement, $G^{\mathbb{K}_0}$ denotes the subgroup of O(3) preserving $\mathbb{K}_0$. Note that $G^{\mathbb{K}_0}$ consists of rotations about the z-axis and reflections through planes containing the z-axis. In particular $G^{\mathbb{K}_0}$ is isomorphic to O(2), and each element of $G^{\mathbb{K}_0}$ preserves either choice of unit normal of $\mathbb{K}_0$.

Corollary 5.9 (fully equivariant index and nullity of $\mathbb{K}_0$). *With notation as above and recalling the comments immediately preceding Proposition 3.8, we have*

$$\operatorname{ind}_{G^{\mathbb{K}_0}}(Q_{\mathrm{N}}^{\mathbb{K}_0}) = 1 \quad \textit{and} \quad \operatorname{nul}_{G^{\mathbb{K}_0}}(Q_{\mathrm{N}}^{\mathbb{K}_0}) = 0.$$

Equivariant index and nullity of $\mathbb{B}^2$. The analysis for the flat disc $\mathbb{B}^2$ (featured in the construction of just one of the families) is trivial, and the conclusions are as follows; in the statement we write $Q_N^{\mathbb{B}^2}$ for the index form of $\mathbb{B}^2$ as a minimal surface with boundary in $(\mathbb{R}^3, g^{\mathbb{R}^3})$ subject to the Neumann boundary condition, namely

$$Q_{\mathrm{N}}^{\mathbb{B}^2} := T\big[\mathbb{B}^2,\, g^{\mathbb{B}^2},\, 0,\, 0,\, \varnothing,\, \partial_{\mathrm{N}}\mathbb{B}^2 := \partial\mathbb{B}^2,\, \varnothing\big].$$

Lemma 5.10 ($(\mathbb{A}_{m+1}$-equivariant) index and nullity of $\mathbb{B}^2$). *With notation as above,*

$$\operatorname{ind}(Q_N^{\mathbb{B}^2}) = 0 \quad \textit{and} \quad \operatorname{nul}(Q_N^{\mathbb{B}^2}) = 1.$$

Moreover, for each integer $m \ge 0$, *the antiprismatic group* $\mathbb{A}_{m+1}$ *preserves* $\mathbb{B}^2$ *and*

$$\operatorname{ind}_{\mathbb{A}_{m+1}}(Q_N^{\mathbb{B}^2}) = \operatorname{nul}_{\mathbb{A}_{m+1}}(Q_N^{\mathbb{B}^2}) = 0.$$

Proof. The first line of equalities is clear, since the Jacobi operator on $\mathbb{B}^2$ is simply the standard Laplacian, whose Neumann kernel is spanned by the constants (to rule out index one can for instance just appeal to the Hopf boundary point lemma). The invariance of $\mathbb{B}^2$ under each $\mathbb{A}_{m+1}$ is obvious, and the proof is then completed by the observation that the constants are not $\mathbb{A}_{m+1}$-equivariant (for any $m \ge 0$). □

From Lemma 5.10 we immediately obtain, analogously to Corollary 5.9 from Lemma 5.8, the following corollary. In the statement O(2) refers to the group of intrinsic isometries of $\mathbb{B}^2$ (extended to isometries of $\mathbb{R}^2$), rather than to some subgroup of O(3), and we write 1 and det for the trivial and determinant homomorphisms $\mathrm{O}(2) \to \mathrm{O}(1)$, respectively. The (O(2), 1)-invariant functions on $\mathbb{B}^2$ are thus the radial functions, while the space of (O(2), det)-invariant functions is trivial.

Corollary 5.11 (indices and nullities of $\mathbb{B}^2$ under O(2) actions). *With notation as above, we have*

$$\mathrm{ind}^{1}_{\mathrm{O}(2)}(Q^{\mathbb{B}^2}_{\mathrm{N}}) = 0, \quad \mathrm{nul}^{1}_{\mathrm{O}(2)}(Q^{\mathbb{B}^2}_{\mathrm{N}}) = 1, \quad \mathrm{ind}^{\det}_{\mathrm{O}(2)}(Q^{\mathbb{B}^2}_{\mathrm{N}}) = \mathrm{nul}^{\det}_{\mathrm{O}(2)}(Q^{\mathbb{B}^2}_{\mathrm{N}}) = 0.$$

Equivariant index and nullity of $\mathbb{M}^{\Xi}$ and $\mathbb{M}^{\Sigma}$. We recall how, away from the equator $\mathbb{S}^1$, the surfaces $\Xi_n^{-\mathbb{K}_0 \cup \mathbb{K}_0}$ and $\Sigma_m^{-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0}$ are constructed as graphs over (subsets of) $-\mathbb{K}_0 \cup \mathbb{K}_0$ and $-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0$. In the vicinity of $\mathbb{S}^1$, the surfaces are instead modeled on certain singly periodic minimal surfaces that belong to a family discovered by Karcher [1988] and generalize the classical singly periodic minimal surfaces of Scherk [1835]. We now summarize the key properties of such models to the extent we will need later.

Proposition 5.12 (desingularizing models). *There exist in $\mathbb{R}^3$ complete, connected, properly embedded minimal surfaces $\mathbb{M}^{\Xi}$ and $\mathbb{M}^{\Sigma}$ having the following properties, which uniquely determine the surfaces up to congruence*:

(i) *$\mathbb{M}^{\Xi}$ and $\mathbb{M}^{\Sigma}$ are periodic in the y direction with period 2π, and the corresponding quotient surfaces have genus zero.*

(ii) *$\mathbb{M}^{\Xi}$ and $\mathbb{M}^{\Sigma}$ are invariant under $\underline{\mathrm{R}}_{\{x=0\}}$, $\underline{\mathrm{R}}_{\{y=\pi/2\}}$, and $\underline{\mathrm{R}}_{\{y=-\pi/2\}}$.*

(iii) *$\mathbb{M}^{\Xi}$ is invariant under $\underline{\mathrm{R}}_{\{z=0\}}$ and $\mathbb{M}^{\Sigma}$ under $\underline{\mathrm{R}}_{\{y=z=0\}}$.*

(iv) *$\mathbb{M}^{\Xi}$ has four ends and $\mathbb{M}^{\Sigma}$ has six ends, all asymptotically planar.*

(v) *Each of $\mathbb{M}^{\Xi}$ and $\mathbb{M}^{\Sigma}$ has an end contained in $\{x \le 0\} \cap \{z \ge 0\}$ whose asymptotic plane intersects $\{z = 0\}$ at the same angle $\omega_0 > 0$ at which $\mathbb{K}_0$ intersects $\mathbb{B}^2$, and $\mathbb{M}^{\Sigma}$ has additionally $\{z = 0\}$ as an asymptotic plane.*

(vi) *$\mathbb{M}^{\Xi}_{\mathrm{fb}} := \mathbb{M}^{\Xi} \cap \{x \le 0\} \cap \{|y| \le \pi/2\}$ and $\mathbb{M}^{\Sigma}_{\mathrm{fb}} := \mathbb{M}^{\Sigma} \cap \{x \le 0\} \cap \{|y| \le \pi/2\}$ are connected free boundary minimal surfaces in the half-slab $\{x \le 0\} \cap \{|y| \le \pi/2\}$, with $\mathbb{M}^{\Xi}_{\mathrm{fb}}$ invariant under $\underline{\mathrm{R}}_{\{z=0\}}$ and $\mathbb{M}^{\Sigma}_{\mathrm{fb}}$ invariant under $\underline{\mathrm{R}}_{\{y=z=0\}}$* (see Figure 3).

(vii) *Each of $\mathbb{M}^{\Xi}_{\mathrm{fb}} \setminus \{z = 0\}$ and $\mathbb{M}^{\Sigma}_{\mathrm{fb}} \setminus \{y = z = 0\}$ has exactly two connected components.*

(viii) *$\mathbb{M}^{\Xi}$ has no umbilics, while the set of umbilic points of $\mathbb{M}^{\Sigma}$ is $\{(0, n\pi, 0) : n \in \mathbb{Z}\}$.*

(ix) *The Gauss map ν^{Ξ} of $\mathbb{M}^{\Xi}$ restricted to the closure of either component of $\mathbb{M}^{\Xi}_{\mathrm{fb}} \setminus \{z = 0\}$ is a bijection onto a solid spherical triangle with all sides geodesic segments of length $\pi/2$ (in other words, a quarter hemisphere), less a point in the interior of one side.*

(x) *The Gauss map ν^{Σ} of $\mathbb{M}^{\Sigma}$ restricted to the closure of either component of $\mathbb{M}^{\Sigma}_{\mathrm{fb}} \setminus \{y = z = 0\}$ is a bijection onto a spherical lune of dihedral angle $\pi/2$ (in other words, a half-hemisphere), less one vertex and a point in the interior of one side.*

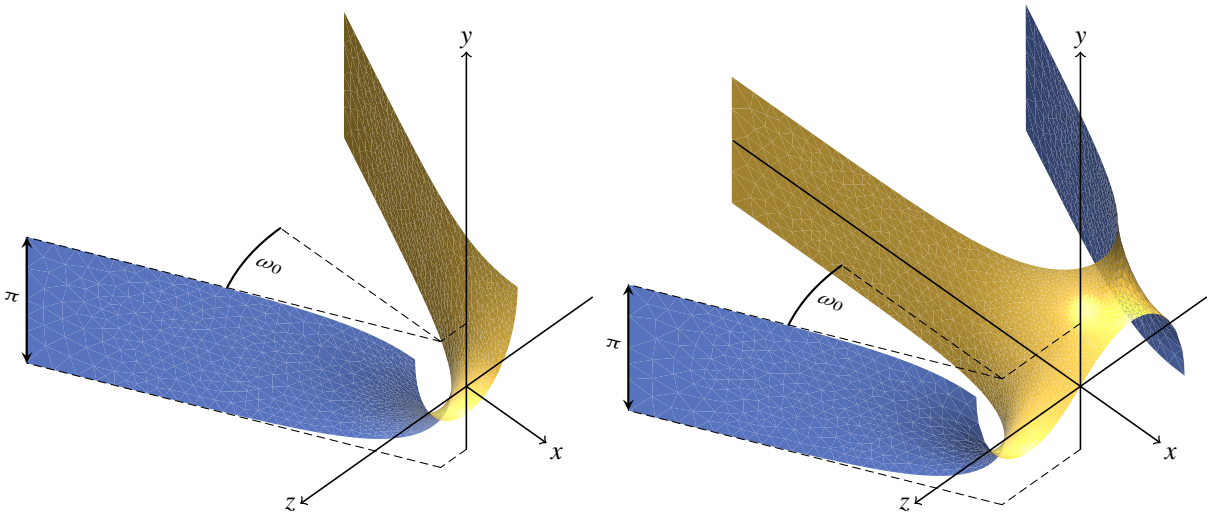

Figure 3. The minimal surfaces $\mathbb{M}^{\Xi}_{\mathrm{fb}}$ (left) and $\mathbb{M}^{\Sigma}_{\mathrm{fb}}$ (right) as defined in Proposition 5.12 (vi).

We refer the reader to Section 3 and Appendix A of [Carlotto et al. 2022b] for further details and a fine analysis of the properties of both surfaces in question. The free boundary minimal surfaces $\mathbb{M}^{\Xi}_{\mathrm{fb}}$ and $\mathbb{M}^{\Sigma}_{\mathrm{fb}}$ are visualized in Figure 3.

Now we examine the index and nullity of $\mathbb{M}^{\Xi}_{\mathrm{fb}}$ and $\mathbb{M}^{\Sigma}_{\mathrm{fb}}$ as free boundary minimal surfaces in the half-slab $\{x \le 0\} \cap \{|y| \le \pi/2\}$. Because the boundary of such a domain is piecewise planar, the corresponding Robin condition associated with the index forms of these surfaces is in fact homogeneous (Neumann).

Let us prove an ancillary result. We will observe (in the proof of Lemma 5.16, to follow shortly) that, by virtue of the behavior of the Gauss maps described in Proposition 5.12, the analysis of the index and nullity of $\mathbb{M}^{\Xi}_{\mathrm{fb}}$ and $\mathbb{M}^{\Sigma}_{\mathrm{fb}}$ reduces to the following index and nullity computations for boundary value problems on suitable Lipschitz domains of $\mathbb{S}^2$.

Lemma 5.13 (index and nullity of $\Delta_{g^{\mathbb{S}^2}} + 2$ on images of Gauss maps of $\mathbb{M}^{\Xi}_{\mathrm{fb}}$ and $\mathbb{M}^{\Sigma}_{\mathrm{fb}}$). *Set*

$$\Omega^{\Xi}_{\mathbb{S}^2} := \mathbb{S}^2 \cap \{x > 0\} \cap \{y > 0\} \cap \{z > 0\},$$
$$\Omega^{\Sigma}_{\mathbb{S}^2} := \mathbb{S}^2 \cap \{x > 0\} \cap \{y > 0\}.$$

Then we have the following indices and nullities, where the final row holds for any $\zeta \in]-1, 1[$ *and, throughout,* T *is the bilinear form* (2-4) *with* Ω *as indicated,* $g = g^{\mathbb{S}^2}$ *is the round metric,* $q = 2$ *(and so is associated to the Schrödinger operator* $\Delta_{g^{\mathbb{S}^2}} + 2$*),* $\partial_{\mathrm{R}}\Omega = \varnothing$*,* $\partial_{\mathrm{D}}\Omega$ *is as indicated, and* $\partial_{\mathrm{N}}\Omega = \partial\Omega \setminus \overline{\partial_{\mathrm{D}}\Omega}$*:*

Ω	$\partial_{\mathrm{D}}\Omega$	$\mathrm{ind}(T)$	$\mathrm{nul}(T)$
$\Omega^{\Xi}_{\mathbb{S}^2}$	$\varnothing$	1	0
	$\{z = 0\}$	0	1
$\Omega^{\Sigma}_{\mathbb{S}^2}$	$\varnothing$	1	1
	$\{x = 0\}$	0	1
	$\{x = 0\} \cap \{z > \zeta\}$	1	0

(5-3)

Proof. By Lemma 3.5, we can fill in the first four rows by identifying the index and nullity of $\Delta_{g^{\mathbb{S}^2}} + 2$ on the entire sphere subject to appropriate symmetries, the relevant spherical harmonics being simply the restrictions of affine functions on $\mathbb{R}^3$. Lemma 3.5 is not directly applicable to the final row, but by the min-max characterization (2-13) of eigenvalues, the i-th eigenvalue for the bilinear form specified in that row must lie between the i-th eigenvalues of the forms specified in the two preceding rows ($\geq$ that of the third row and $\leq$ that of the fourth); moreover, the unique continuation principle implies that both inequalities must be strict ($>$ and $<$). The entries of the final row now follow, concluding the proof. □

We shall fix components of

$$\mathbb{M}^{\Xi}_{\mathrm{fb}} \setminus \{z = 0\} \quad \text{and} \quad \mathbb{M}^{\Sigma}_{\mathrm{fb}} \setminus \{y = z = 0\}$$

once and for all and write Ω^{Ξ} and Ω^{Σ} for their respective interiors: it follows from Proposition 5.12 that $\nu^{\Xi}|_{\Omega^{\Xi}}$ and $\nu^{\Sigma}|_{\Omega^{\Sigma}}$ are diffeomorphisms onto their images, which we can and will identify with the triangle $\Omega^{\Xi}_{\mathbb{S}^2}$ and lune $\Omega^{\Sigma}_{\mathbb{S}^2}$ of Lemma 5.13, respectively, and in particular

$$\begin{aligned}
\{x = 0\} \cap \partial\Omega^{\Xi}_{\mathbb{S}^2} &= \nu^{\Xi}(\{x = 0\} \cap \partial\Omega^{\Xi}),\\
\{y = 0\} \cap \partial\Omega^{\Xi}_{\mathbb{S}^2} &= \overline{\nu^{\Xi}(\{y = \pm\pi/2\} \cap \partial\Omega^{\Xi})},\\
\{z = 0\} \cap \partial\Omega^{\Xi}_{\mathbb{S}^2} &= \nu^{\Xi}(\{z = 0\} \cap \partial\Omega^{\Xi}),
\end{aligned}$$

and

$$\begin{aligned}
\{x = 0\} \cap \partial\Omega^{\Sigma}_{\mathbb{S}^2} &= \overline{\nu^{\Sigma}((\{x = 0\} \cup \{y = z = 0\}) \cap \partial\Omega^{\Sigma})},\\
\{y = 0\} \cap \partial\Omega^{\Sigma}_{\mathbb{S}^2} &= \overline{\nu^{\Sigma}(\{y = \pm\pi/2\} \cap \partial\Omega^{\Sigma})}.
\end{aligned}$$

In what follows, recalling, e.g., that the index of a minimal surface, when finite, can be computed by exhaustion (see [Fischer-Colbrie 1985]), we conveniently introduce this notation, which pertains to certain truncations of $\mathbb{M}^{\Xi}$, $\mathbb{M}^{\Sigma}$, $\mathbb{M}^{\Xi}_{\mathrm{fb}}$, and $\mathbb{M}^{\Sigma}_{\mathrm{fb}}$. To do so, we first fix $R_1 > 0$ large enough that $\mathbb{M}^{\Xi} \setminus \{x^2 + z^2 = R_1^2\}$ consists of five connected components: one component C in $\{x^2 + z^2 < R_1^2\}$ and four components W_1, W_2, W_3, W_4 in the complement, each of which is a graph over (a subset of) an asymptotic half-plane (see Figure 4). For each W_i let $\tau^{(i)}$ be a unit vector parallel to the asymptotic half-plane of W_i, perpendicular to the y-axis (the axis of periodicity), and directed away from ∂W_i toward the corresponding end, namely (up to relabeling)

$$\begin{aligned}
\tau^{(1)} &= (\cos\omega_0,\, 0,\, \sin\omega_0) = -\tau^{(3)},\\
\tau^{(2)} &= (-\cos\omega_0,\, 0,\, \sin\omega_0) = -\tau^{(4)},
\end{aligned}$$

where we recall that $\omega_0 > 0$ is the angle at which $\mathbb{K}_0$ intersects $\mathbb{B}^2$. Now, given $s > R_1$, we define the truncations

$$\begin{aligned}
&W_i(s) := W_i \cap \{\tau^{(i)} \cdot (x, y, z) \leq s\},\\
&\mathbb{M}^{\Xi}(s) := \overline{C} \cup \textstyle\bigcup_{i=1}^{4} W_i(s), && \mathbb{M}^{\Sigma}(s) \text{ analogously (for six ends)},\\
&\mathbb{M}^{\Xi}_{-}(s) := \mathbb{M}^{\Xi}(s) \cap \{x \leq 0\}, && \mathbb{M}^{\Sigma}_{-}(s) := \mathbb{M}^{\Sigma}(s) \cap \{x \leq 0\},\\
&\mathbb{M}^{\Xi}_{\mathrm{fb}}(s) := \mathbb{M}^{\Xi}(s) \cap \mathbb{M}^{\Xi}_{\mathrm{fb}}, && \mathbb{M}^{\Sigma}_{\mathrm{fb}}(s) := \mathbb{M}^{\Sigma}(s) \cap \mathbb{M}^{\Sigma}_{\mathrm{fb}}.
\end{aligned} \tag{5-4}$$

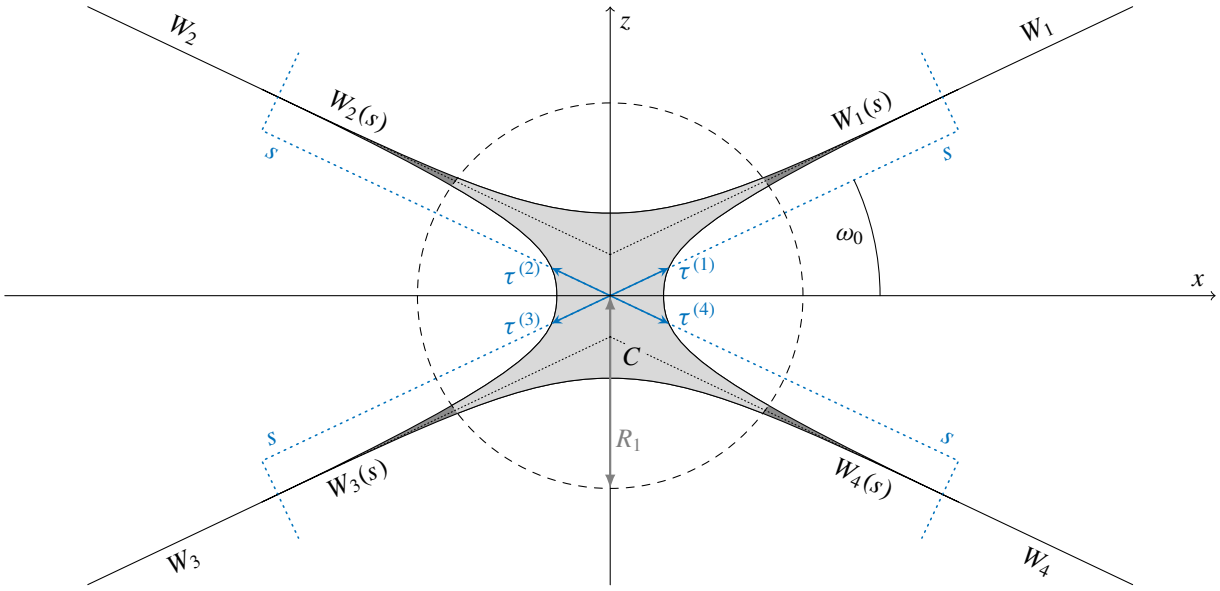

Figure 4. A view of $\mathbb{M}^{\Xi}(s)$.

For each $\epsilon, \epsilon' > 0$, we then set similarly $\mathbb{M}^{\Sigma}_{\mathrm{fb}}(\epsilon^{-1}, \epsilon') := \mathbb{M}^{\Sigma}_{\mathrm{fb}}(\epsilon^{-1}) \cap \{x^2 + y^2 + z^2 > \epsilon'\}$ and

$$\Omega^{\Xi}(\epsilon) := \Omega^{\Xi} \cap \mathbb{M}^{\Xi}_{\mathrm{fb}}(\epsilon^{-1}),$$
$$\Omega^{\Sigma}(\epsilon, \epsilon') := \Omega^{\Sigma} \cap \mathbb{M}^{\Sigma}_{\mathrm{fb}}(\epsilon^{-1}, \epsilon'),$$

truncating Ω^{Ξ} and Ω^{Σ} at (affine) distance ϵ^{-1} and excising from Ω^{Σ} a disc with radius $\sqrt{\epsilon'}$ and center at the umbilic $(0, 0, 0)$. We then in turn define $\Omega^{\Xi}_{\mathbb{S}^2}(\epsilon) := \nu^{\Xi}(\Omega^{\Xi}(\epsilon)) \subset \Omega^{\Xi}_{\mathbb{S}^2}$ as well as $\Omega^{\Sigma}_{\mathbb{S}^2}(\epsilon, \epsilon') := \nu^{\Sigma}(\Omega^{\Sigma}(\epsilon, \epsilon')) \subset \Omega^{\Sigma}_{\mathbb{S}^2}$. As a direct consequence of Lemma 5.13 and Proposition 3.9 we get what follows.

Corollary 5.14. *In the setting above, consider for any $\epsilon, \epsilon' > 0$ the Schrödinger operator $\Delta_{g^{\mathbb{S}^2}} + 2$ on the domains given by $\Omega^{\Xi}_{\mathbb{S}^2}(\epsilon)$ and $\Omega^{\Sigma}_{\mathbb{S}^2}(\epsilon, \epsilon')$, respectively, and subject to any of the boundary conditions specified in the table* (5-3)*, where the boundary is contained in $\partial\Omega^{\Sigma}_{\mathbb{S}^2}$ and $\partial\Omega^{\Xi}_{\mathbb{S}^2}$, respectively, and subject to Dirichlet conditions elsewhere. In other words, let T^{Ξ} be either bilinear form corresponding to the top two rows of* (5-3)*, let T^{Σ} be any bilinear form corresponding to the bottom three rows of* (5-3)*, and consider also the bilinear forms*

$$T^{\Xi}_{\epsilon} := (T^{\Xi})^{\mathrm{D_{int}}}_{\Omega^{\Xi}_{\mathbb{S}^2}(\epsilon)} = T\big[\Omega^{\Xi}_{\mathbb{S}^2}(\epsilon), g^{\mathbb{S}^2}, 2, 0, \partial_{\mathrm{D}}\Omega^{\Xi}_{\mathbb{S}^2} \cup (\partial\Omega^{\Xi}_{\mathbb{S}^2}(\epsilon) \setminus \partial\Omega^{\Xi}_{\mathbb{S}^2}), \partial_{\mathrm{N}}\Omega^{\Xi}_{\mathbb{S}^2}, \varnothing\big],$$
$$T^{\Sigma}_{\epsilon,\epsilon'} := (T^{\Sigma})^{\mathrm{D_{int}}}_{\Omega^{\Sigma}_{\mathbb{S}^2}(\epsilon,\epsilon')} = T\big[\Omega^{\Sigma}_{\mathbb{S}^2}(\epsilon, \epsilon'), g^{\mathbb{S}^2}, 2, 0, \partial_{\mathrm{D}}\Omega^{\Sigma}_{\mathbb{S}^2} \cup (\partial\Omega^{\Sigma}_{\mathbb{S}^2}(\epsilon, \epsilon') \setminus \partial\Omega^{\Sigma}_{\mathbb{S}^2}), \partial_{\mathrm{N}}\Omega^{\Sigma}_{\mathbb{S}^2}, \varnothing\big]$$

using the notation (2-17)*. Then there exists $\epsilon_0 > 0$ such that, for all $0 < \epsilon, \epsilon' < \epsilon_0$,*

$$\mathrm{ind}(T^{\Xi}_{\epsilon}) = \mathrm{ind}(T^{\Xi}) \quad \textit{and} \quad \mathrm{ind}(T^{\Sigma}_{\epsilon,\epsilon'}) = \mathrm{ind}(T^{\Sigma}).$$

In particular, we can derive the following geometric conclusions.

Corollary 5.15 (index of $\mathbb{M}^{\Xi}_{\mathrm{fb}}$ and $\mathbb{M}^{\Sigma}_{\mathrm{fb}}$). *We have the following even and odd indices for* $\mathbb{M}^{\Xi}_{\mathrm{fb}}$ *and* $\mathbb{M}^{\Sigma}_{\mathrm{fb}}$:

S	G	$\mathrm{ind}^+_G(S)$	$\mathrm{ind}^-_G(S)$
$\mathbb{M}^{\Xi}_{\mathrm{fb}}$	$\{z=0\}$	1	0
$\mathbb{M}^{\Sigma}_{\mathrm{fb}}$	$\{y=z=0\}$	1	1

Proof. We will verify (as a sample) the even index asserted in the second row of the table; the other claims are checked in the same fashion. The Gauss map of a minimal surface in $\mathbb{R}^3$ is (anti)conformal away from its umbilics, with conformal factor (one half of) the pointwise square of the norm of its second fundamental form, so by Proposition 3.11, for each $\epsilon, \epsilon' > 0$, the index of ${\Omega^{\Sigma}}_{\mathbb{S}^2}(\epsilon, \epsilon')$ with the foregoing boundary conditions (as in Corollary 5.14, according to the third row of the table in Lemma 5.13) agrees also with the index of $\Omega^{\Sigma}(\epsilon, \epsilon')$ subject to the corresponding boundary conditions. By Lemma 3.5, this last index agrees with the $\{y=z=0\}$-even index of $\mathbb{M}^{\Sigma}_{\mathrm{fb}}(\epsilon^{-1}, \epsilon')$ subject to the Dirichlet condition along the excisions and the Neumann condition everywhere else. Hence, thanks to Corollary 5.14, such a value of the index is equal to 1 for any sufficiently small ϵ, ϵ'. We now conclude, first letting $\epsilon' \to 0$ and appealing to Proposition 3.9 to control the effect of the excision near $(0,0,0)$, and then appealing to the aforementioned characterization of the Morse index via exhaustions, that $\mathbb{M}^{\Sigma}_{\mathrm{fb}}$ indeed has $\{y=z=0\}$-index 1. □

For use in the following subsection, we fix a smooth cutoff function $\Psi : [0,\infty[\to [0,1]$ that is constantly 1 on $\{x \le 1\}$ and constantly 0 on $\{x \ge 2\}$, and we define on $\mathbb{M}^{\Xi}$ and $\mathbb{M}^{\Sigma}$ the functions and metrics

$$\begin{aligned} \psi^{\Xi} &:= (\Psi \circ |x|)|_{\mathbb{M}^{\Xi}}, \quad \rho^{\Xi} := \sqrt{\psi^{\Xi} + \tfrac{1}{2}|A^{\mathbb{M}^{\Xi}}|^2(1-\psi^{\Xi})}, \quad h^{\Xi} := (\rho^{\Xi})^2 g^{\mathbb{M}^{\Xi}}, \\ \psi^{\Sigma} &:= (\Psi \circ |x|)|_{\mathbb{M}^{\Sigma}}, \quad \rho^{\Sigma} := \sqrt{\psi^{\Sigma} + \tfrac{1}{2}|A^{\mathbb{M}^{\Sigma}}|^2(1-\psi^{\Sigma})}, \quad h^{\Sigma} := (\rho^{\Sigma})^2 g^{\mathbb{M}^{\Sigma}}. \end{aligned} \tag{5-5}$$

Note that ρ^{Ξ} is invariant under $\underline{\mathrm{R}}_{\{z=0\}}$, ρ^{Σ} under $\underline{\mathrm{R}}_{\{y=z=0\}}$, and both are invariant under $\underline{\mathrm{R}}_{\{x=0\}}$, $\underline{\mathrm{R}}_{\{y=-\pi/2\}}$, and $\underline{\mathrm{R}}_{\{y=\pi/2\}}$. It is natural to associate to $\mathbb{M}^{\Xi}_{\mathrm{fb}}$, regarded as a free boundary minimal surface in the slab $\{x \le 0\} \cap \{|y| \le \pi/2\}$, the stability form $Q^{\mathbb{M}^{\Xi}_{\mathrm{fb}}}$, defined at least on smooth functions of compact support by

$$Q^{\mathbb{M}^{\Xi}_{\mathrm{fb}}}(u,v) := \int_{\mathbb{M}^{\Xi}_{\mathrm{fb}}} g^{\mathbb{M}^{\Xi}}(\nabla_{g^{\mathbb{M}^{\Xi}}} u, \nabla_{g^{\mathbb{M}^{\Xi}}} v)\, d\mathscr{H}^2(g^{\mathbb{M}^{\Xi}}) - \int_{\mathbb{M}^{\Xi}_{\mathrm{fb}}} |A^{\mathbb{M}^{\Xi}}|^2_{g^{\mathbb{M}^{\Xi}}} uv\, d\mathscr{H}^2(g^{\mathbb{M}^{\Xi}}).$$

From the identity

$$Q^{\mathbb{M}^{\Xi}_{\mathrm{fb}}}(u,v) = \int_{\mathbb{M}^{\Xi}_{\mathrm{fb}}} h^{\Xi}(\nabla_{h^{\Xi}} u, \nabla_{h^{\Xi}} v)\, d\mathscr{H}^2(h^{\Xi}) - \int_{\mathbb{M}^{\Xi}_{\mathrm{fb}}} |A^{\mathbb{M}^{\Xi}}|^2_{h^{\Xi}} uv\, d\mathscr{H}^2(h^{\Xi})$$

and the manifest boundedness of $|A^{\mathbb{M}^{\Xi}}|^2_{h^{\Xi}} = (\rho^{\Xi})^{-2}|A^{\mathbb{M}^{\Xi}}|^2_{g^{\mathbb{M}^{\Xi}}}$, we see that $Q^{\mathbb{M}^{\Xi}_{\mathrm{fb}}}$ is in fact well-defined on $H^1(\mathbb{M}^{\Xi}_{\mathrm{fb}}, h^{\Xi})$. Likewise, the analogously defined $Q^{\mathbb{M}^{\Sigma}_{\mathrm{fb}}}$ is well-defined on $H^1(\mathbb{M}^{\Sigma}_{\mathrm{fb}}, h^{\Sigma})$.

We now point out that we can identify the interiors of $\mathbb{M}^{\Xi}_{\mathrm{fb}}$ and $\mathbb{M}^{\Sigma}_{\mathrm{fb}}$ under the metrics h^{Ξ} and h^{Σ}, respectively, as Lipschitz domains in the setting of Section 2. Concretely, we first consider the Riemannian quotients $\widetilde{\mathbb{M}}^{\Xi}$ and $\widetilde{\mathbb{M}}^{\Sigma}$ of $(\mathbb{M}^{\Xi}, h^{\Xi})$ and $(\mathbb{M}^{\Sigma}, h^{\Sigma})$ under a fundamental period. Then $\widetilde{\mathbb{M}}^{\Xi}$ is diffeomorphic to $\mathbb{S}^2$ with four points removed and $\widetilde{\mathbb{M}}^{\Sigma}$ is diffeomorphic to $\mathbb{S}^2$ with six points removed. By virtue of (5-5) and the behavior of the Gauss maps outlined in Proposition 5.12, we can in fact choose the last

two diffeomorphisms so that they are isometries on neighborhoods of the punctures. In this way we obtain smooth Riemannian compactifications. By composing the defining projection of each tower onto its quotient by a fundamental period with the corresponding embedding into the compactification, we identify (via isometric embedding) the interior of $\mathbb{M}^{\Xi}_{\mathrm{fb}}$ under h^{Ξ} and the interior of $\mathbb{M}^{\Sigma}_{\mathrm{fb}}$ under h^{Σ} with Lipschitz domains $\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}$ and $\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}$ in the two respective compactifications, and we likewise identify $\partial\mathbb{M}^{\Xi}_{\mathrm{fb}}$ and $\partial\mathbb{M}^{\Sigma}_{\mathrm{fb}}$ with subsets of $\partial\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}$ and $\partial\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}$, respectively. Of course, the role of the "ambient manifold" for such Lipschitz domains is played by the Riemannian manifolds $(\mathbb{S}^2, h^{\Xi})$ and $(\mathbb{S}^2, h^{\Sigma})$, respectively; here, with slight abuse of notation, we have tacitly extended the metrics in question across the four and six punctures respectively.

Next, recalling the definition of T from (2-4), we define the bilinear form

$$Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}} := T\big[\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}},\, h^{\Xi},\, q = (\rho^{\Xi})^{-2}|A^{\mathbb{M}^{\Xi}}|^2_{g^{\mathbb{M}^{\Xi}}},\, r = 0,\, \partial_{\mathrm{D}}\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}} = \varnothing,\, \partial_{\mathrm{N}}\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}} = \partial\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}},\, \partial_{\mathrm{R}}\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}} = \varnothing\big],$$

where (as we shall do generally in the sequel for functions defined on $\mathbb{M}^{\Xi}$ or $\mathbb{M}^{\Sigma}$, without further comment) for the potential we tacitly interpret the right-hand side as a function on $\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}$; we define $Q^{\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}}$ in analogous fashion. We then have (see Section 3.4) the equalities

$$Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}} = Q^{\mathbb{M}^{\Xi}_{\mathrm{fb}}} \text{ on } H^1(\mathbb{M}^{\Xi}_{\mathrm{fb}}, h^{\Xi}) \quad \text{and} \quad Q^{\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}} = Q^{\mathbb{M}^{\Sigma}_{\mathrm{fb}}} \text{ on } H^1(\mathbb{M}^{\Sigma}_{\mathrm{fb}}, h^{\Sigma}). \tag{5-6}$$

Lemma 5.16 (index and nullity of $Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}}$ and $Q^{\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}}$). *With definitions as in the preceding paragraph, we have the following indices and nullities*:

S	G	$\mathrm{ind}^+_G(Q^S)$	$\mathrm{nul}^+_G(Q^S)$	$\mathrm{ind}^-_G(Q^S)$	$\mathrm{nul}^-_G(Q^S)$
$\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}$	$\{z = 0\}$	1	0	0	1
$\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}$	$\{y = z = 0\}$	1	1	1	0

Proof. The first row follows from a direct application of Proposition 3.11 in conjunction with the first two rows of the table in Lemma 5.13. Indeed, in this case there are no umbilic points in play (for, recall, $\mathbb{M}^{\Xi}$ has no umbilic points) and the Gauss map furnishes an (anti)conformal map from the compactified quotient onto $\mathbb{S}^2$. For $\mathbb{M}^{\Sigma}$, however, the corresponding conformal factor degenerates at the umbilic at $(0, 0, 0)$, as all of its translates. Nevertheless, aided by Lemma 3.5 and Corollary 3.10, we can verify the indices in the second row in much the same fashion, applying Proposition 3.11 on suitable subdomains (obtained by removing smaller and smaller neighborhoods of the origin).

For the nullities, however, we employ an ad hoc argument since one cannot expect an analogue of the aforementioned Corollary 3.10 to hold true in general. That said, we observe first that the translations in the z direction induce a nontrivial, smooth, bounded, $(\{y = z = 0\}, +)$-invariant (scalar-valued) Jacobi field on $\mathbb{M}^{\Sigma}$ which readily implies that it defines an element of $H^1(\mathbb{M}^{\Sigma}_{\mathrm{fb}}, h^{\Sigma})$. This shows, in view of (5-6), that the nullities in question are at least the values indicated in the table. On the other hand (appealing to Lemma 3.5 for the regularity), each element, say $u : \widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}} \to \mathbb{R}$, of the eigenspace with eigenvalue zero corresponding to the nullities in question is smooth and bounded. If we restrict it to $\Omega^{\Sigma} \subset \widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}$ and consider the precomposition with the inverse of the Gauss map (which, let us recall, yields an (anti)conformal diffeomorphism $\nu^{\mathbb{M}^{\Sigma}} : \Omega^{\Sigma} \to \Omega^{\Sigma}_{\mathbb{S}^2}$), then the resulting function $u_0 := u \circ (\nu^{\mathbb{M}^{\Sigma}})^{-1}$

satisfies $(\Delta_{g^{\mathbb{S}^2}} + 2)u_0 = 0$, and so we get an element contributing to $\mathrm{nul}(T)$, where T is as encoded in the third (resp. fifth) row of the table (5-3) when starting from the $(\{y = z = 0\}, +)$-invariant (resp. $(\{y = z = 0\}, -)$-invariant) problem on $\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}$. It is clear that one thereby gets injective maps of vector spaces, and so from Lemma 5.13

$$\mathrm{nul}^{+}_{G}(Q^{\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}}) \leq 1, \quad \mathrm{nul}^{-}_{G}(Q^{\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}}) \leq 0,$$

which in particular implies that such maps are, a posteriori, linear isomorphisms, and thus completes the proof. □

When we wish to consider the sets $\mathbb{M}^{\Xi}_{\mathrm{fb}}(s)$ and $\mathbb{M}^{\Sigma}_{\mathrm{fb}}(s)$ endowed with the metrics h^{Ξ} and h^{Σ}, respectively, we shall denote them by $\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}(s)$ and $\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}(s)$. Recalling the notation of Section 2.5, we further define

$$Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}(s)}_{\mathrm{D}} := (Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}})^{\mathrm{D}_{\mathrm{int}}}_{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}(s)} \quad \text{and} \quad Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}(s)}_{\mathrm{N}} := (Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}})^{\mathrm{N}_{\mathrm{int}}}_{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}(s)}. \tag{5-7}$$

In short, we are adjoining respectively Dirichlet or Neumann boundary conditions along the cuts.

Lemma 5.17 (spectra of $Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}(s)}$ and $Q^{\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}(s)}$). *For each integer $i \geq 1$,*

$$\lim_{s\to\infty} \lambda_i^{\{z=0\},\pm}(Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}(s)}_{\mathrm{D}}) = \lim_{s\to\infty} \lambda_i^{\{z=0\},\pm}(Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}(s)}_{\mathrm{N}}) = \lambda_i^{\{z=0\},\pm}(Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}}),$$
$$\lim_{s\to\infty} \lambda_i^{\{y=z=0\},\pm}(Q^{\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}(s)}_{\mathrm{D}}) = \lim_{s\to\infty} \lambda_i^{\{y=z=0\},\pm}(Q^{\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}(s)}_{\mathrm{N}}) = \lambda_i^{\{y=z=0\},\pm}(Q^{\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}})$$

for any consistent choice of $+$ or $-$ on both sides of each equality.

Proof. We will write down the proof of the two equalities in the first line for the $+$ choice, as the remaining cases can be proved in the same way. First note that Proposition 3.9 gives us

$$\lim_{s\to\infty} \lambda_i^{\{z=0\},+}(Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}(s)}_{\mathrm{D}}) = \lambda_i^{\{z=0\},+}(Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}}).$$

Using the min-max characterization (2-13) of eigenvalues, we then also get

$$\limsup_{s\to\infty} \lambda_i^{\{z=0\},+}(Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}(s)}_{\mathrm{N}}) \leq \limsup_{s\to\infty} \lambda_i^{\{z=0\},+}(Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}(s)}_{\mathrm{D}}) = \lambda_i^{\{z=0\},+}(Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}}).$$

The key step now toward the goal of establishing

$$\liminf_{s\to\infty} \lambda_i^{\{z=0\},+}(Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}(s)}_{\mathrm{N}}) \geq \lambda_i^{\{z=0\},+}(Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}})$$

(which completes the proof) is to construct a family of (appropriately symmetric) linear extension operators $E_s : H^1(\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}(s)) \to H^1(\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}})$ uniformly bounded in s, assuming $s \geq s_0$ for some universal $s_0 > 0$. With these extensions in hand it is straightforward, for example, to adapt the argument for (3-6) in the proof of Proposition 3.8.

We now outline the construction of the E_s extension operators. By the imposed symmetry (in the case under discussion even reflection through $\{z = 0\}$) and by taking s large enough, it suffices to specify the extension on a single end W, a graph over a subset of the corresponding asymptotic plane Π (with τ the corresponding defining vector, recalling the notation preceding (5-4)). Let $\varpi : W \to \Pi$ be the associated projection. By partitioning the given function using appropriately chosen smooth cutoff functions (fixed independently of s), it in fact suffices to consider the extension problem for a function

$v \in H^1(W \cap \mathbb{M}^{\Xi}_{\mathrm{fb}}(s), h^{\Xi})$ such that the support of $\varpi^* v$ is compactly contained in the rectangle (expressed in the notation of (5-4))

$$\{0 < \tau \cdot (x, y, z) \leq s\} \cap \{-\pi \leq 2y \leq \pi\}.$$

We can extend $\varpi^* v$ via even reflection through the s side of the above rectangle, thereby obtaining an extension of v to an element of $H^1(W, h^{\Xi})$. The asymptotic convergence of W to Π, the monotonic decay of ρ^{Ξ} along W toward ∞, and the conformal invariance (in the current two-dimensional setting) of the Dirichlet energy ensure that this extension has the desired properties. □

5.3. ***Deconstruction of the surfaces and regionwise geometric convergence.*** We first take a moment to briefly review the constructions of the surfaces from [Carlotto et al. 2022b]. First (see [loc. cit., Section 3]), an approximate minimal surface in $\mathbb{B}^3$, called the initial surface, whose boundary is contained in $\partial\mathbb{B}^3$ and which meets $\partial\mathbb{B}^3$ exactly orthogonally, is fashioned by hand, via suitable interpolations, from the models ($\mathbb{K}_0$, $\mathbb{M}^{\Xi}$, or $\mathbb{M}^{\Sigma}$, and for $\Sigma_m^{-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0}$ also $\mathbb{B}^2$). Second (see [loc. cit., Section 5]), the final exact solution is identified as the normal graph of a small function over the approximate solution. For what pertains to this second step we wish only to highlight that the assignment of graph to function is made using not the usual Euclidean metric $g^{\mathbb{R}^3}$ but instead an O(3)-invariant metric (fixed once and for all, independently of the data n or m) conformally Euclidean and called the auxiliary metric. On a neighborhood of the origin this metric agrees exactly with the Euclidean one, while on a neighborhood of $\partial\mathbb{B}^3 = \mathbb{S}^2$ it agrees exactly with the cylindrical metric on $\mathbb{S}^2 \times \mathbb{R}$; this last property and the orthogonality of the intersection of the initial surface with $\partial\mathbb{B}^3$ ensure that the boundary of the resulting graph is also in $\partial\mathbb{B}^3$. We will write $\widehat{\Xi}_n^{-\mathbb{K}_0 \cup \mathbb{K}_0}$ and $\widehat{\Sigma}_m^{-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0}$ for the initial surfaces and $\varpi_n^{\Xi} : \Xi_n^{-\mathbb{K}_0 \cup \mathbb{K}_0} \to \widehat{\Xi}_n^{-\mathbb{K}_0 \cup \mathbb{K}_0}$ and $\varpi_m^{\Sigma} : \Sigma_m^{-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0} \to \widehat{\Sigma}_m^{-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0}$ for the nearest-point projections under the above auxiliary metric.

Turning to the first step, actually (because of the presence of a cokernel) one constructs for each given n or m not just a single initial surface but a (continuous) one-parameter family of them. In the construction this parameter is treated as an unknown and is determined only in the second step, simultaneously with the defining function for the final surface. Here, however, we can take the construction for granted and accordingly speak of a single initial surface, whose defining parameter value is some definite (though not explicit) function of n or m as appropriate. Nevertheless we must explain that this parameter enters the construction at the level of the building blocks, except for $\mathbb{B}^2$, which is unaffected, as follows. First, the catenoidal annulus $\mathbb{K}_0$ is just one in a family $\mathbb{K}_\epsilon$ (see the beginning of Section 3.1 in [loc. cit.]) of such annuli, all rotationally symmetric about the z-axis, depending smoothly on ϵ. The details are not critical here, but each $\mathbb{K}_\epsilon$ is the intersection with $\mathbb{B}^3$ of a complete catenoid with axis the z-axis, and $\mathbb{K}_\epsilon$ meets $\mathbb{S}^2$ at two circles of latitude, the upper one a circle of orthogonal intersection and the lower one the circle at height $z = \epsilon$. Similarly, from $\mathbb{M}^{\Xi}$ and $\mathbb{M}^{\Sigma}$ we define, by explicit graphical deformation, families which here we will call $\mathbb{M}^{\Xi}_\delta$ and $\mathbb{M}^{\Sigma}_\delta$ (see the beginning of Section 3.2 of [loc. cit.]). These deformations are the identity on the "cores" of $\mathbb{M}^{\Xi}$ and $\mathbb{M}^{\Sigma}$ and smoothly transition to translations on the ends, in the z-direction, up or down depending on the end, and through a displacement determined by δ. Importantly, all the $\mathbb{M}^{\Xi}_\delta$ and $\mathbb{M}^{\Sigma}_\delta$ have the same symmetries as $\mathbb{M}^{\Xi}$ and $\mathbb{M}^{\Sigma}$, respectively. Now the datum n determines building blocks $\mathbb{M}^{\Xi}_{\delta^{\Xi}(n)}$ and $\mathbb{K}_{\epsilon^{\Xi}(n)}$, while the datum m determines building blocks $\mathbb{M}^{\Sigma}_{\delta^{\Sigma}(m)}$, $\mathbb{K}_{\epsilon^{\Sigma}(m)}$, and $\mathbb{B}^2$.

We next define maps Φ_n^{Ξ} and Φ_m^{Σ} [loc. cit., (3.37)] from neighborhoods of $(1/n)\mathbb{M}^{\Xi}_{\delta^{\Xi}(n)} \cap \{x \leq 0\}$ and $(1/(m+1))\mathbb{M}^{\Sigma}_{\delta^{\Sigma}(m)} \cap \{x \leq 0\}$, respectively, into $\mathbb{B}^3$, so as to "wrap" the cores of these surfaces around the equator $\mathbb{S}^1$ approximately isometrically but to take their asymptotic half-planes (in $\{x \leq 0\}$) onto $\pm\mathbb{K}_{\epsilon^{\Xi}(n)}$ in the first case and onto $\pm\mathbb{K}_{\epsilon^{\Sigma}(m)}$ and $\mathbb{B}^2$ in the second. Thus, just referring to the family $\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}$ for the sake of brevity, we truncate the surface $\mathbb{M}^{\Xi}_{\delta^{\Xi}(n)}$ by intersecting with $\{x \geq -n^{3/4}\}$, and then apply Φ_n^{Ξ} to the truncated surface scaled-down by a factor of $1/n$. The image is embedded (for n large enough) and contained in the ball, and is in fact contained in a tubular neighborhood of $\mathbb{S}^1$ with radius of order $n^{-1/4}$.

Near the two truncation boundary components, the surface is a small graph over either $\pm\mathbb{K}_{\epsilon^{\Xi}(n)}$. We smoothly cut off the defining function in a $(1/n)$-neighborhood of the boundary to make the surface exactly catenoidal there and then extend using these annuli on the other side of the truncation boundary all the way to $\partial\mathbb{B}^3$. The result is our initial surface $\widehat{\Xi}_n^{-\mathbb{K}_0\cup\mathbb{K}_0}$. The initial surface $\widehat{\Sigma}_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}$ is constructed analogously, now also smoothly transitioning from the middle truncation boundary to coincide with $\mathbb{B}^2$ on a neighborhood of the origin. In what follows we will distill those objects and ancillary results that are needed for the spectral convergence theorems we will prove in Section 5.4.

Decompositions. Recalling (5-4) for the definition of the below domains, our construction in [Carlotto et al. 2022b] provides, in particular, smooth maps

$$\varphi^{M_n^{\Xi}} : \mathbb{M}_-^{\Xi}(n^{5/8}) \to \Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0},$$
$$\varphi^{M_m^{\Sigma}} : \mathbb{M}_-^{\Sigma}((m+1)^{5/8}) \to \Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0},$$

which are smooth coverings of their images. For all $0 < s \leq \sqrt{n}$, or, respectively, $0 < s \leq \sqrt{m+1}$, we in turn define

$$M_n^{\Xi}(s) := \varphi^{M_n^{\Xi}}(\mathbb{M}_-^{\Xi}(s)) \subset \Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0},$$
$$M_m^{\Sigma}(s) := \varphi^{M_m^{\Sigma}}(\mathbb{M}_-^{\Sigma}(s)) \subset \Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}.$$

In practice, in addition to the upper bound required on s, we will be interested only in s greater than a universal constant set by $\mathbb{M}^{\Xi}$ and $\mathbb{M}^{\Sigma}$: we want to truncate far enough out (in the domain) that near and beyond the truncation boundary the surface is already the graph of a small function over the asymptotic planes. In a typical application to follow we will take s large in absolute terms and then take n or m large with respect to s, so we will not always repeat either restriction. When they do hold, $\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0} \setminus M_n^{\Xi}(s)$ consists of two connected components and $\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0} \setminus M_m^{\Sigma}(s)$ consists of three, and we define

$$K_n^{\Xi}(s) := \text{the closure of the component of } \Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0} \setminus M_n^{\Xi}(s) \text{ on which } z \text{ is maximized},$$
$$K_m^{\Sigma}(s) := \text{the closure of the component of } \Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0} \setminus M_m^{\Sigma}(s) \text{ on which } z \text{ is maximized},$$
$$B_m^{\Sigma}(s) := \text{the closure of the component of } \Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0} \setminus M_m^{\Sigma}(s) \text{ that contains the origin}.$$

Observe that each $M_n^{\Xi}(s)$ is invariant under $\underline{\mathrm{R}}_{\{z=0\}}$, that the interiors of $M_n^{\Xi}(s)$, $K_n^{\Xi}(s)$, and $\underline{\mathrm{R}}_{\{z=0\}}K_n^{\Xi}(s)$ are pairwise disjoint, and that the last three regions cover $\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}$. In particular, considering the interior of such sets, one thereby determines a candidate partition for the application of Proposition 3.1. Similarly, $M_m^{\Sigma}(s)$ and $B_m^{\Sigma}(s)$ are invariant under $\underline{\mathrm{R}}_{\{y=z=0\}}$; the interiors of $M_m^{\Sigma}(s)$, $B_m^{\Sigma}(s)$, $K_m^{\Sigma}(s)$, and $\underline{\mathrm{R}}_{\{y=z=0\}}K_m^{\Sigma}(s)$ are pairwise disjoint, and these four surfaces also cover $\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}$.

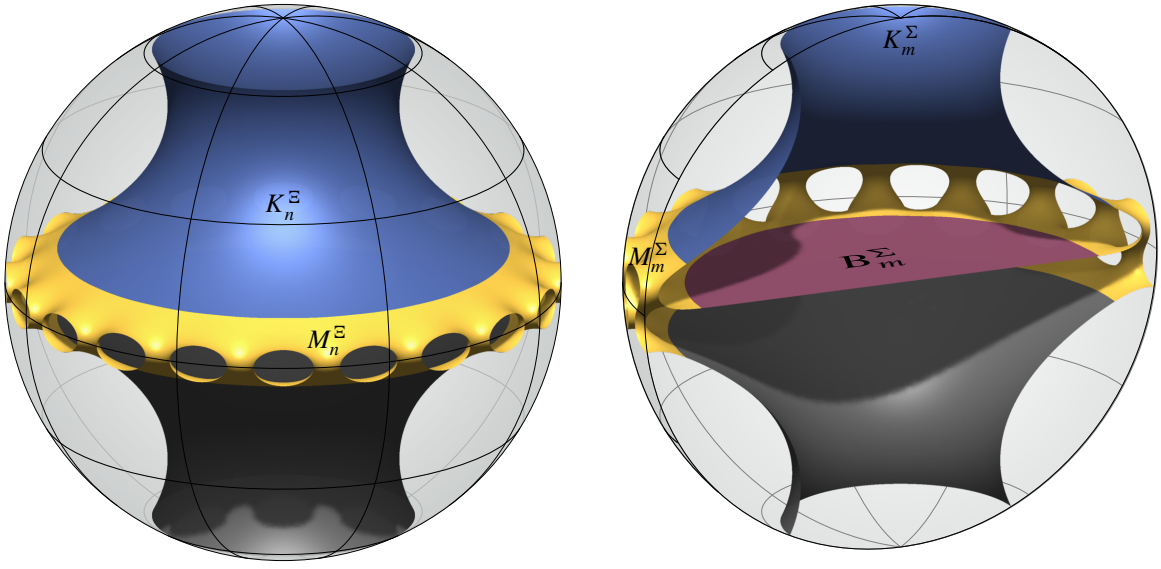

Figure 5. Decomposition of $\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}$ (left) and $\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}$ (right, cutaway view).

We agree to distinguish the choices $s=\sqrt{n}$ and $s=\sqrt{m+1}$ by omission of the parameter value:

$$\begin{aligned} M_n^\Xi &:= M_n^\Xi(\sqrt{n}), & K_n^\Xi &:= K_n^\Xi(\sqrt{n}),\\ M_m^\Sigma &:= M_m^\Sigma(\sqrt{m+1}), & K_m^\Sigma &:= K_m^\Sigma(\sqrt{m+1}), \quad B_m^\Sigma := B_m^\Sigma(\sqrt{m+1}), \end{aligned}$$

as visualized in Figure 5. We also define the dilated truncations (see Figure 6)

$$\begin{aligned} M_{\mathrm{fb},n}^\Xi(s) &:= n(M_n^\Xi(s)\cap W_{-\pi/(2n)}^{\pi/(2n)}) = n\varphi^{M_n^\Xi}(\mathbb{M}_{\mathrm{fb}}^\Xi(s)), & M_{\mathrm{fb},n}^\Xi &:= M_{\mathrm{fb},n}^\Xi(\sqrt{n}),\\ M_{\mathrm{fb},m}^\Sigma(s) &:= (m+1)(M_m^\Sigma(s)\cap W_{-\pi/(2(m+1))}^{\pi/(2(m+1))}) = (m+1)\varphi^{M_m^\Sigma}(\mathbb{M}_{\mathrm{fb}}^\Sigma(s)), & M_{\mathrm{fb},m}^\Sigma &:= M_{\mathrm{fb},m}^\Sigma(\sqrt{m+1}), \end{aligned}$$

where the notation for wedges has been given in (4-2), and finally introduce the transition regions

$$\Lambda_n^\Xi(s) := \overline{M_{\mathrm{fb},n}^\Xi \setminus M_{\mathrm{fb},n}^\Xi(s)}, \quad \Lambda_m^\Sigma(s) := \overline{M_{\mathrm{fb},m}^\Sigma \setminus M_{\mathrm{fb},m}^\Sigma(s)}.$$

Geometric estimates. Before proceeding, we declare the following abbreviated notation for the metrics and second fundamental forms on $M_{\mathrm{fb},n}^\Xi$ and $M_{\mathrm{fb},m}^\Sigma$ (induced by their inclusions in $(\mathbb{R}^3, g^{\mathbb{R}^3})$):

$$g_n^\Xi := g^{M_{\mathrm{fb},n}^\Xi}, \quad g_m^\Sigma := g^{M_{\mathrm{fb},m}^\Sigma}, \quad A_n^\Xi := A^{M_{\mathrm{fb},n}^\Xi}, \quad A_m^\Sigma := A^{M_{\mathrm{fb},m}^\Sigma}.$$

In analogy with (5-5), we first write ψ_n^Ξ, ψ_m^Σ for the unique functions on $M_{\mathrm{fb},n}^\Xi$, $M_{\mathrm{fb},m}^\Sigma$ such that

$$\psi^\Xi = (n\circ\varphi^{M_n^\Xi})^*\psi_n^\Xi, \quad \psi^\Sigma = ((m+1)\circ\varphi^{M_m^\Sigma})^*\psi_m^\Sigma,$$

and then in turn define

$$\begin{aligned} \rho_n^\Xi &:= \sqrt{\psi_n^\Xi + \tfrac{1}{2}|A_n^\Xi|_{g_n}^2(1-\psi_n^\Xi) + e^{-2n}}, & h_n^\Xi &:= (\rho_n^\Xi)^2 g_n^\Xi,\\ \rho_m^\Sigma &:= \sqrt{\psi_m^\Sigma + \tfrac{1}{2}|A_m^\Sigma|_{g_m^\Sigma}^2(1-\psi_m^\Sigma) + e^{-2m}}, & h_m^\Sigma &:= (\rho_m^\Sigma)^2 g_m^\Sigma. \end{aligned} \tag{5-8}$$

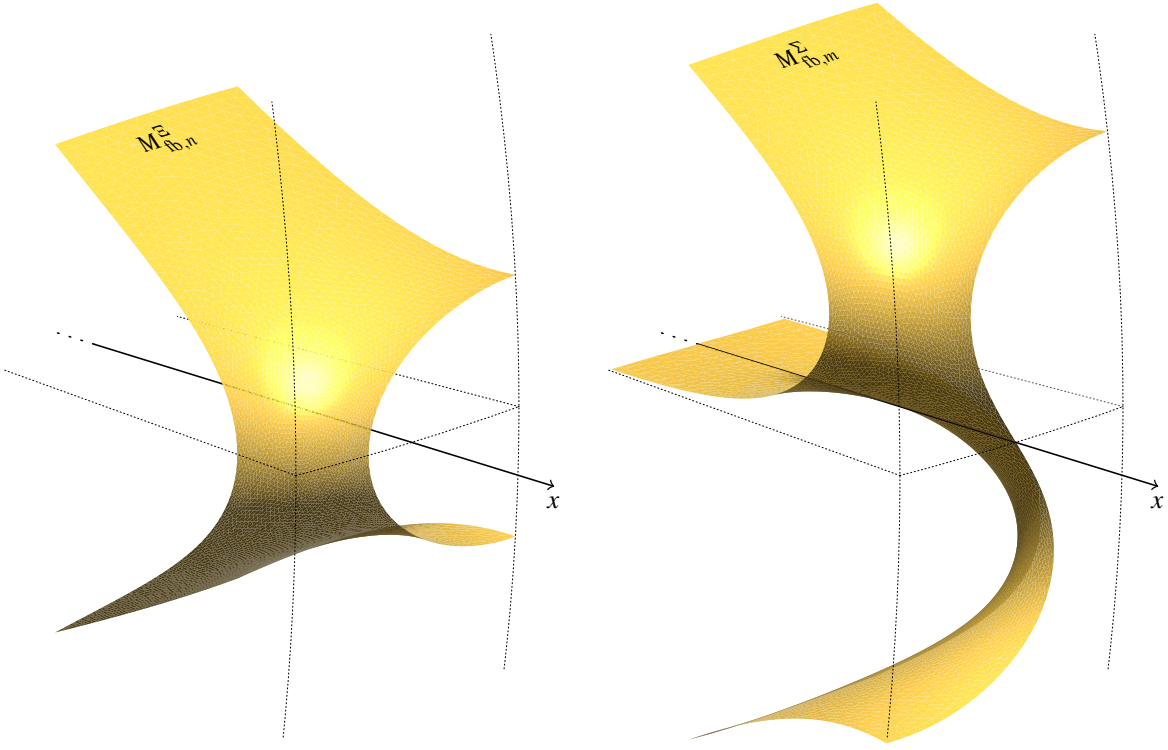

Figure 6. The dilated truncations $M^{\Xi}_{\mathrm{fb},n}$ (left) and $M^{\Sigma}_{\mathrm{fb},m}$ (right).

The terms e^{-2n} and e^{-2m} above are included to ensure the conformal factors vanish nowhere. For the sake of brevity, and consistent with the notation adopted in the previous subsections, we set

$$\widehat{M}^{\Xi}_{\mathrm{fb},n} := (M^{\Xi}_{\mathrm{fb},n}, h^{\Xi}_n), \quad \widehat{M}^{\Sigma}_{\mathrm{fb},m} := (M^{\Sigma}_{\mathrm{fb},m}, h^{\Sigma}_m), \quad \widehat{M}^{\Xi}_{\mathrm{fb},n}(s) := (M^{\Xi}_{\mathrm{fb},n}(s), h^{\Xi}_n), \quad \widehat{M}^{\Sigma}_{\mathrm{fb},m}(s) := (M^{\Sigma}_{\mathrm{fb},m}(s), h^{\Sigma}_m),$$

so that $\widehat{M}^{\Xi}_{\mathrm{fb},n}$ and $\widehat{M}^{\Sigma}_{\mathrm{fb},m}$ and their truncations $\widehat{M}^{\Xi}_{\mathrm{fb},n}(s) \subset M^{\Xi}_{\mathrm{fb},n}$ and $M^{\Sigma}_{\mathrm{fb},m}(s) \subset M^{\Sigma}_{\mathrm{fb},m}$ are always understood as being equipped with the conformal metrics h^{Ξ}_n and h^{Σ}_m, rather than g^{Ξ}_n and g^{Σ}_m.

Lemma 5.18 (convergence of $M^{\Xi}_{\mathrm{fb},n}(s)$ and $M^{\Sigma}_{\mathrm{fb},m}(s)$). *For every $s > 0$, there exists $m_s > 0$ such that, for every integer $m > m_s$,*

(i) *the region $M^{\Sigma}_{\mathrm{fb},m}(s)$ is defined and is the diffeomorphic image under $(m+1)\varphi^{M^{\Sigma}_m}$ of $\mathbb{M}^{\Sigma}_{\mathrm{fb}}(s)$,*

(ii) $(m+1)\varphi^{M^{\Sigma}_m}(\mathbb{M}^{\Sigma}_{\mathrm{fb}}(s) \cap \{x = 0\}) = M^{\Sigma}_{\mathrm{fb},m}(s) \cap (m+1)\mathbb{S}^2$,

(iii) $\varphi^{M^{\Sigma}_m}$ *commutes with* $\underline{\mathrm{R}}_{\{z=0\}}$, *and*

(iv) $M^{\Sigma}_m(s) = (m+1)^{-1}\mathbb{A}_{m+1}M^{\Sigma}_{\mathrm{fb},m}(s)$ *is a surface with smooth boundary.*

Moreover, for every $s > 0$ and $\alpha \in]0,1[$,

(v) $((m+1)\circ\varphi^{M^{\Sigma}_m})^* g^{\Sigma}_m \xrightarrow[m\to\infty]{C^{1,\alpha}(\mathbb{M}^{\Sigma}_{\mathrm{fb}}(s), g^{\mathbb{M}^{\Sigma}})} g^{\mathbb{M}^{\Sigma}}$ *and*

(vi) $((m+1)\circ\varphi^{M^{\Sigma}_m})^* A^{\Sigma}_m \xrightarrow[m\to\infty]{C^{0,\alpha}(\mathbb{M}^{\Sigma}_{\mathrm{fb}}(s), g^{\mathbb{M}^{\Sigma}})} A^{\mathbb{M}^{\Sigma}}$.

All the above statements have analogues for $\Xi^{-\mathbb{K}_0\cup\mathbb{K}_0}_n$ in place of $\Sigma^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}_m$, mutatis mutandis.

The first four claims are immediate from the definitions, while the convergence assertions are ensured, in the case of $\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}$, by the following estimates from [Carlotto et al. 2022b], the case of $\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}$ being completely analogous. Namely, the estimate [loc. cit., (5.20)] provides $C^{2,\alpha}$ bounds for the defining function of $\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}$ as a graph over the corresponding initial surface, and so controls the projection map ϖ_m^Σ from $\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}$ to the initial surface. The same estimate [loc. cit., (5.20)] also bounds the parameter value for the initial surface from the one-parameter family that is selected to produce the final one. On the other hand, [loc. cit., Proposition 3.18] provides estimates on the initial surface in terms of the datum g as well as the value of the continuous parameter. (As an aid to extracting the required information, we point out that the map $\varpi_{M_{m,\xi}}$ in [loc. cit., (3.43)] is essentially (that is, up to some quotienting and the exact extent of the domains) the inverse of the map $\varpi_{m-1}^\Sigma \circ \varphi^{M_{m-1}^\Sigma}$ of the present article.)

Let us consider the other portions of our surfaces. By construction $\varpi_n^\Xi(K_n^\Xi)$ and $\varpi_m^\Sigma(K_m^\Sigma)$ (subsets of the initial surfaces) are graphs (under the Euclidean metric $g^{\mathbb{R}^3}$) over subsets of $\mathbb{K}_{\epsilon^\Xi(n)}$ and $\mathbb{K}_{\epsilon^\Sigma(m)}$, and $\varpi_m^\Sigma(B_m^\Sigma)$ is a graph over $\mathbb{B}^2$. Thus, by composition with a further projection, we obtain injective maps

$$\varpi_n^\Xi(K_n^\Xi) \to \mathbb{K}_{\epsilon^\Xi(n)}, \quad \varpi_m^\Sigma(K_m^\Sigma) \to \mathbb{K}_{\epsilon^\Sigma(m)}, \quad \text{and} \quad B_m^\Sigma \to \mathbb{B}^2.$$

Moreover, the image of each of these three maps is O(2) invariant: the image of the third is a disc with radius tending to 1 as $m \to \infty$, the image of the second is a catenoidal annulus with upper boundary circle coinciding with that of $\mathbb{K}_{\epsilon^\Sigma(m)}$ and lower boundary circle tending to that of $\mathbb{K}_{\epsilon^\Sigma(m)}$ as $m \to \infty$; the image of the first admits an analogous description.

In particular, by composing further with dilations of scale factor tending to 1, we obtain diffeomorphisms

$$\varphi^{B_m^\Sigma} : \mathbb{B}^2 \to B_m^\Sigma;$$

similarly reparametrizing in the radial direction one also obtains diffeomorphisms

$$\varphi^{K_n^\Xi} : \mathbb{K}_0 \to K_n^\Xi, \quad \varphi^{K_m^\Sigma} : \mathbb{K}_0 \to K_m^\Sigma.$$

The inverses of these maps may be regarded as small perturbations (for n and m large) of the nearest-point projection onto $\mathbb{R}^2 \subset \mathbb{B}^2$ or onto the complete catenoid containing $\mathbb{K}_0$, as appropriate. Somewhat more formally, by reference to [loc. cit.] (specifically Proposition 3.18 and estimate (5.20) therein), much as in the proof of Lemma 5.18, we confirm the following properties of K_n^Ξ, K_m^Σ, and B_m^Σ.

Lemma 5.19 (convergence of K_n^Ξ and K_m^Σ). *There exists $m_0 > 0$ such that, for each integer $m > m_0$,*

(i) *$\varphi^{K_m^\Sigma}$ is defined and a diffeomorphism from $\mathbb{K}_0$ onto K_m^Σ,*

(ii) *$\varphi^{K_m^\Sigma}$ commutes with each element of $\mathbb{Y}_{m+1}$, and*

(iii) *$\varphi^{K_m^\Sigma}$ takes the upper boundary component of $\mathbb{K}_0$ to the upper boundary component of K_m^Σ.*

Moreover, for every $\alpha \in]0, 1[$,

(vi) $(\varphi^{K_m^\Sigma})^* g^{\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}}|_{K_m^\Sigma} \xrightarrow[m\to\infty]{C^{1,\alpha}(\mathbb{K}_0, g^{\mathbb{K}_0})} g^{\mathbb{K}_0}$ *and*

(vii) $(\varphi^{K_m^\Sigma})^* A^{\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}}|_{K_m^\Sigma} \xrightarrow[m\to\infty]{C^{0,\alpha}(\mathbb{K}_0, g^{\mathbb{K}_0})} A^{\mathbb{K}_0}$.

All the above statements have analogues for $\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}$ in place of $\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}$, mutatis mutandis.

Lemma 5.20 (convergence of B_g^Σ). *There exists $m_0 > 0$ such that, for each integer $m > m_0$,*

(i) *$\varphi^{B_m^\Sigma}$ is defined and a diffeomorphism from $\mathbb{B}^2$ onto B_m^Σ and*

(ii) *$\varphi^{B_m^\Sigma}$ commutes with each element of $\mathbb{A}_{m+1}$.*

Moreover, for each $\alpha \in]0, 1[$,

(iii) $(\varphi^{B_m^\Sigma})^* g^{\Sigma_m^{-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0}}|_{B_m^\Sigma} \xrightarrow[m\to\infty]{C^{1,\alpha}(\mathbb{B}^2, g^{\mathbb{B}^2})} g^{\mathbb{B}^2}$ *and*

(iv) $(\varphi^{B_m^\Sigma})^* A^{\Sigma_m^{-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0}}|_{B_m^\Sigma} \xrightarrow[m\to\infty]{C^{0,\alpha}(\mathbb{B}^2, g^{\mathbb{B}^2})} 0.$

Last we focus on the transition regions. Let us agree to write t_n^Ξ and t_m^Σ for the distance functions on $n\mathbb{K}_{\epsilon^\Xi(n)}$ and $(m+1)\mathbb{K}_{\epsilon^\Sigma(m)}$ from their respective lower boundary circles. By construction (assuming s large enough in absolute terms) $n\varpi_n^\Xi(n^{-1}\Lambda_n^\Xi(s))$ has two connected components, one a graph over the catenoidal annular wedge

$$\{s \le t_n^\Xi \le \sqrt{n}\} \cap W^{\pi/(2n)}_{-\pi/(2n)} \subset n\mathbb{K}_{\epsilon^\Xi(n)}$$

and the other the reflection of this last one through $\{z = 0\}$, while $(m+1)\varpi_m^\Sigma((m+1)^{-1}\Lambda_n^\Sigma(s))$ has three connected components, one a graph over the planar annular wedge

$$\{s \le (m+1) - r \le \sqrt{m+1}\} \cap W^{\pi/(2(m+1))}_{-\pi/(2(m+1))} \cap (m+1)\mathbb{B}^2,$$

another a graph over the catenoidal annular wedge

$$\{s \le t_m^\Sigma \le \sqrt{m+1}\} \cap W^{\pi/(2(m+1))}_{-\pi/(2(m+1))} \subset (m+1)\mathbb{K}_{\epsilon^\Sigma(m)},$$

and the third the reflection of this last one through $\{y = z = 0\}$.

Projecting onto these rotationally invariant sets and parametrizing them by arc length t in the "radial" direction and $\vartheta := n\theta$, or, respectively, $\vartheta := (m+1)\theta$ in the angular direction (with θ restricted to the appropriate interval containing 0), we obtain injective maps

$$\varphi^{\Lambda_n^\Xi(s),\mathbb{K}} : [s, \sqrt{n}] \times \left[\frac{\pi}{2}, \frac{\pi}{2}\right] \to \Lambda_n^\Xi, \quad \varphi^{\Lambda_m^\Sigma(s),\mathbb{K}}, \varphi^{\Lambda_m^\Sigma(s),\mathbb{B}^2} : [s, \sqrt{m+1}] \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \to \Lambda_m^\Sigma,$$

whose images are components of $\Lambda_n^\Xi(s)$ and $\Lambda_m^\Sigma(s)$ that generate the latter regions under $\{z = 0\}$ and $\{y = z = 0\}$, respectively.

Lemma 5.21 (estimates on $\Lambda_n^\Xi(s)$ and $\Lambda_m^\Sigma(s)$). *Let $\alpha \in]0, 1[$. There exists $s_0 > 0$ such that, for each $s > s_0$, there exists $m_s > 0$ such that, for every integer $m > m_s$,*

(i) *$(\varphi^{\Lambda_m^\Sigma(s),\mathbb{K}})^* |A_m^\Sigma|^2_{g_m^\Sigma}(t, \vartheta) = a_1(t)m^{-2} + a_2(t, \vartheta)e^{-t/4}$ for some smooth functions a_1 and a_2 having $C^{0,\alpha}(dt^2 + d\vartheta^2)$ norm bounded independently of m and s,*

(ii) *$(\varphi^{\Lambda_m^\Sigma(s),\mathbb{B}^2})^* |A_m^\Sigma|^2_{g_m^\Sigma}(t, \vartheta) = a_3(t, \vartheta)e^{-t/4}$ for some smooth function a_3 having $C^{0,\alpha}(dt^2 + d\vartheta^2)$ norm bounded independently of m and s,*

(iii) *$(\varphi^{\Lambda_m^\Sigma(s),\mathbb{K}})^* g_m^\Sigma = dt^2 + (1 + m^{-1}tf^1(t))\, d\vartheta^2 + f_{uv}^1(t, \vartheta)e^{-t/4}\, du\, dv$ for some smooth functions f^1, f_{uv}^1 having $C^{1,\alpha}(dt^2 + d\vartheta^2)$ norm bounded independently of m and s,*

(iv) *$(\varphi^{\Lambda_m^\Sigma(s),\mathbb{B}^2})^* g_m^\Sigma = dt^2 + (1 + m^{-1}tf^2(t))\, d\vartheta^2 + f_{uv}^2(t, \vartheta)e^{-t/4}\, du\, dv$ for some smooth functions f^2 and f_{uv}^2 having $C^{1,\alpha}(dt^2 + d\vartheta^2)$ norm bounded independently of m and s,*

(v) $\Delta_{(\varphi^{\Lambda^\Sigma_m(s),\mathbb{K}})^* g^\Sigma_m} = \partial_t^2 + m^{-1} c_1^t(t)\,\partial_t + (1+m^{-1/2} b_1^{\vartheta\vartheta}(t))\,\partial_\vartheta^2 + e^{-t/4}(b_2^{uv}(t,\vartheta)\,\partial_u\,\partial_v + c_2^u(t,\vartheta)\,\partial_u)$ *for some smooth functions* $b_1^{\vartheta\vartheta}$, b_2^{uv}, c_1^t, c_2^u *having* $C^{0,\alpha}(dt^2+d\vartheta^2)$ *norm bounded independently of* m *and* s, *and*

(vi) $\Delta_{(\varphi^{\Lambda^\Sigma_m(s),\mathbb{B}^2})^* g^\Sigma_m} = \partial_t^2 + m^{-1} c_3^t(t)\,\partial_t + (1+m^{-1/2} b_3^{\vartheta\vartheta}(t))\,\partial_\vartheta^2 + e^{-t/4}(b_4^{uv}(t,\vartheta)\,\partial_u\,\partial_v + c_4^u(t,\vartheta)\,\partial_u)$ *for some smooth functions* $b_3^{\vartheta\vartheta}$, b_4^{uv}, c_3^t, c_4^i *having* $C^{0,\alpha}(dt^2+d\vartheta^2)$ *norm bounded independently of* m *and* s.

It is understood that, in items (iii), (iv), (v), (vi), *one sums over* $u, v \in \{t, \vartheta\}$.

Furthermore,

(vii) $\lim_{s\to\infty}\lim_{m\to\infty} \mathscr{H}^2(h^\Sigma_m)(\Lambda^\Sigma_m(s)) = 0$.

The same claims hold for $\Lambda^\Xi_n(s)$, *mutatis mutandis.*

Proof. Again the estimates are ultimately justified by reference to the construction in [Carlotto et al. 2022b], most specifically (5.20) and Proposition 3.18 therein. That said, we also note how (v) follows easily from (iii), as does (vi) from (iv); furthermore, it is clear that the justification of (ii) is analogous to (in fact simpler than) (i), and (iv) is analogous to (iii). As a result, we briefly explain the ideas behind the elementary computations required for the proof, in the case of $\Lambda^\Sigma_m(s)$, with regard to items (i) and (iii).

The projection of this region onto the blown-up initial surface $(m+1)\widehat{\Sigma}^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}_m$ is itself constructed as a graph over $(m+1)\mathbb{K}_{\epsilon^\Sigma(m)}$ or $\mathbb{B}^2$. Estimate [loc. cit., (5.20)] ensures that $m\epsilon^\Sigma(m)$ is bounded uniformly in m. The defining function of the above graph is obtained by "transferring" the defining functions of the corresponding ends of $\mathbb{M}^\Sigma$ over their asymptotic planes. These defining functions decay exponentially in the distance along the planes. In turn $\Lambda^\Sigma_m(s)$ is a graph over this portion of the initial surface with defining function that is also guaranteed (by [loc. cit., (5.20)]) to decay exponentially, though a priori at a slower rate; we have chosen $\frac{1}{4}$ somewhat arbitrarily. This accounts for all exponential factors appearing in the estimates.

The m-dependent terms in the estimates for the metric (and Laplacian) arise simply from the choice of (t,ϑ) coordinates on disc and catenoidal models. The m^{-2} term in the first item arises from scaling the second fundamental form of the "asymptotic" catenoid to this component (while the corresponding term for the disc vanishes). With the estimates for the second fundamental form in place, the final item — the area estimate — follows (recalling the definitions (5-8)) from the bound

$$\int_{-\pi/2}^{\pi/2}\int_s^{\sqrt{m+1}} (a_1 m^{-2} + a_2 e^{-t/4})\,dt\,d\vartheta \le C(m^{-3/2} + e^{-s/4})$$

and the analogous estimate concerning the disk-type component instead. □

5.4. ***Regionwise spectral convergence.*** For each region S among M^Ξ_n, M^Σ_m, K^Ξ_n, K^Σ_n, and B^Σ_m (depicted in Figure 5), we write Q^S_{N} for the Jacobi form of S as a minimal surface in $\mathbb{B}^3$ with boundary, subject to the Robin condition (4-1) where ∂S meets $\partial\mathbb{B}^3$ and subject to the Neumann condition elsewhere: recalling (2-17), we set

$$Q^S_{\mathrm{N}} := \begin{cases} (Q^{\Xi^{-\mathbb{K}_0\cup\mathbb{K}_0}_n})^{\mathrm{N_{int}}}_S & \text{for } S \subset \Xi^{-\mathbb{K}_0\cup\mathbb{K}_0}_n, \\ (Q^{\Sigma^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}_m})^{\mathrm{N_{int}}}_S & \text{for } S \subset \Sigma^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}_m \end{cases}$$

(where on the right-hand side we slightly abuse notation in that in place of S we really mean its interior). Similarly, for S either $M^{\Xi}_{\mathrm{fb},n}$ or $M^{\Sigma}_{\mathrm{fb},m}$, we write Q^{S}_{N} for the Jacobi form of S as a minimal surface in either $n\mathbb{B}^3$ or $(m+1)\mathbb{B}^3$, subject to the Robin condition either $du(\eta) = n^{-1}u$ or $du(\eta) = (m+1)^{-1}u$ where ∂S meets either $n\mathbb{S}^2$ or $(m+1)\mathbb{S}^2$, respectively, and subject to the Neumann condition elsewhere. Keeping in mind the statement of Proposition 3.1, we stress that the adjunction of Neumann conditions in the "interior" boundaries is motivated by our task of deriving *upper* bounds on the Morse index of our examples. Recalling the notation $\widehat{M}^{\Xi}_{\mathrm{fb},n}$ and $\widehat{M}^{\Sigma}_{\mathrm{fb},n}$, we remark that the bilinear forms Q^{S}_{N} and $Q^{\hat{S}}_{\mathrm{N}}$ agree by definition for each S as above, but whenever we refer to the eigenvalues, eigenfunctions, index, and nullity of the latter we shall always mean those defined with respect to the h^{Ξ}_n or h^{Σ}_m metric.

In the notation of (2-4), we have in particular (see Proposition 3.11)

$$\begin{aligned} Q^{M^{\Xi}_{\mathrm{fb},n}}_{\mathrm{N}} &= T\big[M^{\Xi}_{\mathrm{fb},n},\, g^{\Xi}_n,\, q^{\Xi}_n = |A^{\Xi}_n|^2_{g^{\Xi}_n},\, r^{\Xi}_n = n^{-1},\\ &\qquad \partial_{\mathrm{D}} M^{\Xi}_{\mathrm{fb},n} = \varnothing,\, \partial_{\mathrm{N}} M^{\Xi}_{\mathrm{fb},n} = \partial M^{\Xi}_{\mathrm{fb},n} \setminus n\mathbb{S}^2,\, \partial_{\mathrm{R}} M^{\Xi}_{\mathrm{fb},n} = \partial M^{\Xi}_{\mathrm{fb},n} \setminus \overline{\partial_{\mathrm{N}} M^{\Xi}_{\mathrm{fb},n}}\big]\\ &= T\big[M^{\Xi}_{\mathrm{fb},n},\, h^{\Xi}_n,\, (\rho^{\Xi}_n)^{-2} q^{\Xi}_n,\, (\rho^{\Xi}_n)^{-1} n^{-1},\, \varnothing,\, \partial M^{\Xi}_{\mathrm{fb},n} \setminus n\mathbb{S}^2,\, \partial M^{\Xi}_{\mathrm{fb},n} \setminus \overline{\partial_{\mathrm{N}} M^{\Xi}_{\mathrm{fb},n}}\big] = Q^{\widehat{M}^{\Xi}_{\mathrm{fb},n}}_{\mathrm{N}} \end{aligned} \tag{5-9}$$

and similarly for $Q^{M^{\Sigma}_{\mathrm{fb},m}}_{\mathrm{N}} = Q^{\widehat{M}^{\Sigma}_{\mathrm{fb},m}}_{\mathrm{N}}$. Observe further (see Lemma 3.5 and Proposition 3.11) that

$$\begin{aligned} \mathrm{ind}_{\mathbb{P}_n}(Q^{M^{\Xi}_n}_{\mathrm{N}}) &= \mathrm{ind}^{+}_{\{z=0\}}(Q^{\widehat{M}^{\Xi}_{\mathrm{fb},n}}_{\mathrm{N}}), \qquad \mathrm{ind}_{\mathbb{Y}_n}(Q^{M^{\Xi}_n}_{\mathrm{N}}) = \mathrm{ind}(Q^{\widehat{M}^{\Xi}_{\mathrm{fb},n}}_{\mathrm{N}}),\\ \mathrm{ind}_{\mathbb{A}_{m+1}}(Q^{M^{\Sigma}_m}_{\mathrm{N}}) &= \mathrm{ind}^{-}_{\{y=z=0\}}(Q^{\widehat{M}^{\Sigma}_{\mathrm{fb},m}}_{\mathrm{N}}), \quad \mathrm{ind}_{\mathbb{Y}_{m+1}}(Q^{M^{\Sigma}_m}_{\mathrm{N}}) = \mathrm{ind}(Q^{\widehat{M}^{\Sigma}_{\mathrm{fb},m}}_{\mathrm{N}}), \end{aligned}$$

and likewise for the corresponding nullities.

Lemma 5.22 (equivariant index and nullity on K^{Ξ}_n, K^{Σ}_m, and B^{Σ}_m). *There exist $n_0, m_0 > 0$ such that we have the following indices and nullities for all integers $n > n_0$ and $m > m_0$:*

S	G	$\mathrm{ind}_G(Q^S_{\mathrm{N}})$	$\mathrm{nul}_G(Q^S_{\mathrm{N}})$
K^{Ξ}_n	$\mathbb{Y}_n$	1	0
K^{Σ}_m	$\mathbb{Y}_{m+1}$	1	0
B^{Σ}_m	$\mathbb{A}_{m+1}$	0	0

Additionally, still assuming $m > m_0$, we have the upper bound

$$\mathrm{ind}_{\mathbb{Y}_{m+1}}(Q^{B^{\Sigma}_m}_{\mathrm{N}}) + \mathrm{nul}_{\mathbb{Y}_{m+1}}(Q^{B^{\Sigma}_m}_{\mathrm{N}}) \leq 1.$$

Proof. We use the convergence described in Lemmas 5.19 and 5.20 along with Proposition 3.8 to compare the low eigenvalues of the regions in question with those of their limiting models, as recorded in Lemmas 5.8 and 5.10. □

While we have cut the surfaces $\Xi^{-\mathbb{K}_0 \cup \mathbb{K}_0}_n$ and $\Sigma^{-\mathbb{K}_0 \cup \mathbb{B}^2 \cup \mathbb{K}_0}_m$ in such a way that the resulting regions K^{Ξ}_n and K^{Σ}_m converge uniformly to $\mathbb{K}_0$ and likewise B^{Σ}_m to $\mathbb{B}^2$, thereby securing the preceding lemma in a straightforward fashion, the cases of M^{Ξ}_n and M^{Σ}_m are more subtle. Our approach here (especially the proof of eigenfunction bounds in Lemma 5.25 and their application to Lemma 5.26) draws inspiration from the analysis Kapouleas makes of the invertibility of the Jacobi operator on "extended standard regions" in many gluing constructions; for a specific example, concerning Scherk towers glued to catenoids, we refer the reader to the proof of [Kapouleas 1997, Lemma 7.4].

To proceed, recalling (2-17), for each $s > 0$ and each integer n (sufficiently large in terms of s), we define

$$Q_{\mathrm{D}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}(s)} := (Q_{\mathrm{N}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}})^{\mathrm{D_{int}}}_{\widehat{M}^{\Xi}_{\mathrm{fb},n}(s)} \quad \text{and} \quad Q_{\mathrm{N}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}(s)} := (Q_{\mathrm{N}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}})^{\mathrm{N_{int}}}_{\widehat{M}^{\Xi}_{\mathrm{fb},n}(s)}$$

and analogously for $\widehat{M}^{\Sigma}_{\mathrm{fb},m}(s)$ in place of $\widehat{M}^{\Xi}_{\mathrm{fb},n}(s)$.

Lemma 5.23 (spectral convergence for $\widehat{M}^{\Xi}_{\mathrm{fb},n}(s)$ and $\widehat{M}^{\Sigma}_{\mathrm{fb},m}(s)$). *With the above notation, we have*

$$\lambda_i^{\{z=0\},\pm}(Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}}) = \lim_{s\to\infty}\lim_{n\to\infty} \lambda_i^{\{z=0\},\pm}(Q_{\mathrm{D}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}(s)}) = \lim_{s\to\infty}\lim_{n\to\infty} \lambda_i^{\{z=0\},\pm}(Q_{\mathrm{N}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}(s)})$$

for each integer $i \geq 1$ and each common choice of sign $\pm$ on both sides of each equation. The analogous statements hold, mutatis mutandis, for $\widehat{M}^{\Sigma}_{\mathrm{fb},m}$ in place of $\widehat{M}^{\Xi}_{\mathrm{fb},n}$.

Proof. Fix i. By Lemma 5.18 and Proposition 3.8, for each $s > 0$, we have

$$\lim_{n\to\infty} \lambda_i^{\{z=0\},+}(Q_{\mathrm{D}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}(s)}) = \lambda_i^{\{z=0\},+}(Q_{\mathrm{D}}^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}(s)}),$$
$$\lim_{n\to\infty} \lambda_i^{\{z=0\},+}(Q_{\mathrm{N}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}(s)}) = \lambda_i^{\{z=0\},+}(Q_{\mathrm{N}}^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}(s)}).$$

An application of Lemma 5.17 completes the proof in this case, and the proofs of the remaining three cases are structurally identical to this one. □

Lemma 5.24 (eigenvalue upper bounds on $\widehat{M}^{\Xi}_{\mathrm{fb},n}$ and $\widehat{M}^{\Sigma}_{\mathrm{fb},m}$). *With the above notation, we have*

$$\limsup_{n\to\infty} \lambda_i^{\{z=0\},\pm}(Q_{\mathrm{N}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}}) \leq \lambda_i^{\{z=0\},\pm}(Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}}),$$
$$\limsup_{m\to\infty} \lambda_i^{\{y=z=0\},\pm}(Q_{\mathrm{N}}^{\widehat{M}^{\Sigma}_{\mathrm{fb},m}}) \leq \lambda_i^{\{y=z=0\},\pm}(Q^{\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}})$$

for each integer $i \geq 1$ and each common choice of sign $\pm$ on both sides of each equation.

Proof. We give the proof for the $+$ choice on both sides of the top equation, the proofs for the remaining three cases being identical in structure to this one. Fix $i \geq 1$. By (2-13), considering extensions by zero of functions corresponding to the right-hand side below to obtain valid test functions corresponding to the left, we get at once the inequality

$$\lambda_i^{\{z=0\},+}(Q_{\mathrm{N}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}}) \leq \lambda_i^{\{z=0\},+}(Q_{\mathrm{D}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}(s)})$$

for all $s > 0$ and all n sufficiently large (in terms of s) such that $\widehat{M}^{\Xi}_{\mathrm{fb},n}(s)$ is defined. We then finish by applying Lemma 5.23. □

Lemma 5.25 (uniform bounds on eigenvalues and eigenfunctions of $Q_{\mathrm{N}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}}$ and $Q_{\mathrm{N}}^{\widehat{M}^{\Sigma}_{\mathrm{fb},m}}$). *For each integer $i \geq 1$, there exist $C_i, k_i > 0$ such that, for each integer $k > k_i$ and whenever $\lambda_i^{(k)}$ is the i-th eigenvalue of $Q_{\mathrm{N}}^{\widehat{M}^{\Xi}_{\mathrm{fb},k}}$ or $Q_{\mathrm{N}}^{\widehat{M}^{\Sigma}_{\mathrm{fb},k}}$ and $v_i^{(k)}$ is any corresponding eigenfunction of unit L^2 norm (under either h_k^{Ξ} or h_k^{Σ} as appropriate), we have the bounds*

$$\max\{|\lambda_i^{(k)}|, \|v_i^{(k)}\|_{H^1}, \|v_i^{(k)}\|_{C^0}\} \leq C_i$$

(*where the H^1 norm is defined via either h_n^{Ξ} or h_m^{Σ} as applicable and we emphasize that C_i does not depend on k*).

Proof. We will give the proof for $\widehat{M}^{\Xi}_{\mathrm{fb},n}$, that for $\widehat{M}^{\Sigma}_{\mathrm{fb},m}$ being identical in structure. Fix $i \geq 1$, and let $\lambda^{(n)}$ and $v^{(n)}$ be as in the statement for each integer n (suppressing the fixed index i); it is our task to show that, by assuming n large enough in terms of just i, we can ensure the asserted bounds on $\lambda^{(n)}$ and $v^{(n)}$. In particular our assumptions include the normalization $\|v^{(n)}\|_{L^2(M^{\Xi}_{\mathrm{fb},n}, h^{\Xi}_n)} = 1$.

Lemma 5.24 provides an upper bound on $\lambda^{(n)}$ independent of n. We deduce a lower bound on $\lambda^{(n)}$ as follows. Keeping in mind the min-max characterization (2-13), we observe that in the ratio

$$\frac{\langle u, q_n u\rangle_{L^2(M^{\Xi}_{\mathrm{fb},n}, h^{\Xi}_n)} + \langle u|_{\partial_{\mathrm{R}} M^{\Xi}_{\mathrm{fb},n}}, r_n u|_{\partial_{\mathrm{R}} M^{\Xi}_{\mathrm{fb},n}}\rangle_{L^2(\partial_{\mathrm{R}} M^{\Xi}_{\mathrm{fb},n}, h^{\Xi}_n)}}{\|u\|^2_{L^2(M^{\Xi}_{\mathrm{fb},n}, h^{\Xi}_n)}}, \tag{5-10}$$

with

$$r_n := (\rho_n^{\Xi})^{-1}|_{\partial_{\mathrm{R}} M^{\Xi}_{\mathrm{fb},n}} n^{-1} = (1 + e^{-2n})^{-1/2} n^{-1}, \quad q_n := (\rho_n^{\Xi})^{-2} |A_n^{\Xi}|^2_{g_n^{\Xi}},$$

we have not only a uniform upper bound on r_n, but also, by inspecting (5-8) and bearing in mind the convergence described in Lemma 5.18 as well as the boundedness (with decay) of the second fundamental form of $\mathbb{M}^{\Xi}$,

$$\sup_n \|q_n\|_{C^0(M^{\Xi}_{\mathrm{fb},n})} < \infty.$$

In addition, the convergence in Lemma 5.18 further ensures that the constants appearing in (2-3), with $(\Omega, g) = (M^{\Xi}_{\mathrm{fb},n}, h^{\Xi}_n)$ and $\partial_{\mathrm{R}}\Omega$ in place of $\partial\Omega$, can be chosen uniformly in n: thus, employing such a trace inequality and exploiting the foregoing uniform bounds we secure the promised uniform lower bound on $\lambda^{(n)}$.

In turn, from the definitions of eigenvalues and eigenfunctions and the normalization of $v^{(n)}$, we have

$$\|\nabla_{h_n^{\Xi}} v^{(n)}\|^2_{L^2(M^{\Xi}_{\mathrm{fb},n}, h^{\Xi}_n)} = \lambda^{(n)} + \langle v^{(n)}, q_n v^{(n)}\rangle_{L^2(M^{\Xi}_{\mathrm{fb},n}, h^{\Xi}_n)} + \langle v^{(n)}|_{\partial_{\mathrm{R}} M^{\Xi}_{\mathrm{fb},n}}, r_n v^{(n)}|_{\partial_{\mathrm{R}} M^{\Xi}_{\mathrm{fb},n}}\rangle_{L^2(\partial_{\mathrm{R}} M^{\Xi}_{\mathrm{fb},n}, h^{\Xi}_n)}.$$

The uniform bound on $\|v^{(n)}\|_{H^1(M^{\Xi}_{\mathrm{fb},n}, h^{\Xi}_n)}$ now follows, in view of the above equality, from the upper bound on $\lambda^{(n)}$ as well as again the above uniform bounds on q_n and r_n.

It remains to establish the uniform C^0 bound. To start, by Lemma 3.5 and standard elliptic regularity, $v^{(n)}$ is smooth up to the boundary: indeed, it satisfies

$$\begin{cases} (\Delta_{h_n^{\Xi}} + (\rho_n^{\Xi})^{-2}|A^{\Xi}_{n\, g_n^{\Xi}}|^2 + \lambda^{(n)}) v^{(n)} = 0 & \text{in } M^{\Xi}_{\mathrm{fb},n}, \\ h_n^{\Xi}(\eta_n^{\Xi}, \nabla_{h_n^{\Xi}} v^{(n)}) = (1 + e^{-2n})^{-1/2} n^{-1} v^{(n)} & \text{on } \partial_{\mathrm{R}} M^{\Xi}_{\mathrm{fb},n}, \\ h_n^{\Xi}(\eta_n^{\Xi}, \nabla_{h_n^{\Xi}} v^{(n)}) = 0 & \text{on } \partial_{\mathrm{N}} M^{\Xi}_{\mathrm{fb},n}, \end{cases} \tag{5-11}$$

with η_n^{Ξ} the outward h_n^{Ξ} unit conormal to $M^{\Xi}_{\mathrm{fb},n}$. As established above, we have bounds independent of n on $|\lambda^{(n)}|$ and the q_n and r_n functions. By Lemma 5.18 (and the uniform geometry of $\mathbb{M}^{\Xi}$), we also have uniform control over the geometry of $(M^{\Xi}_{\mathrm{fb},n}(s), h_n^{\Xi})$ for each $s > 0$ and all n sufficiently large in terms of s.

Standard elliptic regularity therefore ensures that, for every $s > 0$, there exist $n_s > 0$ and $\gamma(s) > 0$ such that

$$\|v^{(n)}|_{M^{\Xi}_{\mathrm{fb},n}(s)}\|_{C^0(M^{\Xi}_{\mathrm{fb},n}(s), h^{\Xi}_n)} \leq \gamma(s) \quad \text{for every integer } n > n_s. \tag{5-12}$$

Since we do not have uniform control on the geometry of $(M^{\Xi}_{\mathrm{fb},n} = M^{\Xi}_{\mathrm{fb},n}(\sqrt{n}), h^{\Xi}_n)$, we do not obtain a global bound independent of n in the same fashion. Instead the proof will be completed by securing a C^0 bound for $v^{(n)}$, independent of n, on $\Lambda^{\Xi}_n(s)$ for some $s > 0$ to be determined. In the remainder of the proof, $\gamma(s)$ will continue to denote the above constant, depending on s, while C will denote a strictly positive constant whose value may change from instance to instance but can always be selected independently of s and n.

To proceed we multiply both sides of the PDE in (5-11) by $(\rho^{\Xi}_n)^2$ to get

$$(\Delta_{g^{\Xi}_n} + |A^{\Xi}_{n\, g^{\Xi}_n}|^2 + \lambda^{(n)}(\rho^{\Xi}_n)^2)v^{(n)} = 0, \tag{5-13}$$

and we aim to bound $v^{(n)}$ on $\Lambda^{\Xi}_n(s)$ on the basis of this equation, with unknown but controlled (as we explain momentarily) Dirichlet data on the portion of $\partial\Lambda^{\Xi}_n(s)$ contained in the interior of $M^{\Xi}_{\mathrm{fb},n}$ and with homogeneous Neumann data on the rest of the boundary. By the symmetries it suffices to establish the estimate on just the component of $\Lambda^{\Xi}_n(s)$ that is a graph over a subset of $n\mathbb{K}_0$. (For $\Lambda^{\Sigma}_m(s)$ one must also consider the component which is a graph over a subset of $(m+1)\mathbb{B}^2$, but this case does not differ in substance from the one we treat now.)

Recall the map

$$\varphi^{\Lambda^{\Xi}_n(s),\mathbb{K}} : [s, \sqrt{n}] \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \to \Lambda_n(s)$$

introduced above Lemma 5.21, and continue to write (t, ϑ) for the standard coordinates on its domain. For the remainder of this proof we abbreviate $\varphi^{\Lambda^{\Xi}_n(s),\mathbb{K}}$ to $\varphi_{n,s}$ and its domain to $R_{n,s}$. Setting $w^{(n)} := \varphi^*_{n,s} v^{(n)}$, we pull back (5-13) to get

$$\Delta_{\varphi^*_{n,s} g^{\Xi}_n} w^{(n)} = -w^{(n)} \varphi^*_{n,s}(|A^{\Xi}_n|^2_{g^{\Xi}_n} + \lambda^{(n)}(\rho^{\Xi}_n)^2).$$

From the uniform bound on $\lambda^{(n)}$, the expression for the conformal factor in (5-8), and item (i) of Lemma 5.21, we in turn obtain

$$\Delta_{\varphi^*_{n,s} g^{\Xi}_n} w^{(n)} = (c_{n,s} e^{-t/4} + d_{n,s} n^{-2}) w^{(n)} \tag{5-14}$$

for some smooth functions $c_{n,s},\ d_{n,s}$ having $C^{0,\alpha}(dt^2 + d\vartheta^2)$ norms uniformly bounded in n and s, with $\alpha \in]0, 1[$ now fixed for the rest of the proof. (Here and below when referring to items of Lemma 5.21 we have in mind of course the corresponding statements for $\Lambda^{\Xi}_n(s)$ in place of $\Lambda^{\Sigma}_m(s)$.)

Noting that we have (5-14) for all sufficiently large s, it now follows from the C^0 bound (5-12) and standard interior Schauder estimates (using also item (iii) of Lemma 5.21) that

$$\|w^{(n)}(s, \cdot)\|_{C^{2,\alpha}(d\vartheta^2)} \le C\gamma(s+1) \quad \text{for every integer } n > n_{s+1}. \tag{5-15}$$

Since $v^{(n)}$ satisfies the homogeneous Neumann condition along $\partial M^{\Xi}_{\mathrm{fb},n}$, with the aid of item (iii) of Lemma 5.21, we have

$$(\partial_t w^{(n)})(\sqrt{n}, \vartheta) = e_{n,s} e^{-\sqrt{n}/4} (\partial_\vartheta w^{(n)})(\sqrt{n}, \vartheta), \tag{5-16}$$

$$(\partial_\vartheta w^{(n)})(\cdot, \pm\pi/2) = 0 \tag{5-17}$$

for some smooth function $e_{n,s}$ having $C^{1,\alpha}(dt^2 + d\vartheta^2)$ norm bounded independently of n and s. (For (5-17) we simply use the fact that $\varphi_{n,s}$ has been constructed by composing and restricting maps which commute with the symmetries of the construction, including the reflections through planes corresponding to $\vartheta = \pm\pi/2$.)

Appealing again to standard Schauder estimates, now also up to the boundary, we can conclude from (5-14)–(5-17) that

$$\|w^{(n)}\|_{C^{2,\alpha}(dt^2+d\vartheta^2)} \leq C(\gamma(s+1) + \|w^{(n)}\|_{C^0}) \tag{5-18}$$

for n and s sufficiently large in terms of the bounds assumed on the functions $c_{n,s}$, $d_{n,s}$, and $e_{n,s}$, as well as constants, which can be chosen uniformly, that appear in local Schauder estimates on $R_{n,s}$. If we exploit (5-18) in (5-16), we get

$$\|(\partial_t w^{(n)})(\sqrt{n}, \cdot)\|_{C^{1,\alpha}(d\vartheta^2)} \leq Ce^{-\sqrt{n}/4}(\gamma(s+1) + \|w^{(n)}\|_{C^0}), \tag{5-19}$$

once again for n and s assumed large enough in terms of absolute constants.

We next decompose $w^{(n)}$ into

$$w_0^{(n)} := \frac{1}{\pi}\int_{-\pi/2}^{\pi/2} w^{(n)}(\,\cdot\,, \vartheta)\, d\vartheta, \quad w_\perp^{(n)} := w^{(n)} - w_0^{(n)}.$$

From (5-14), (5-18), and item (v) of Lemma 5.21, we obtain

$$\partial_t^2 w_0^{(n)} = a_{n,s}^0 e^{-t/4} + b_{n,s}^0 n^{-2} + c_{n,s}^0 n^{-1}\,\partial_t w_0^{(n)}, \tag{5-20}$$

with

$$\frac{\|a_{n,s}^0\|_{C^0} + \|b_{n,s}^0\|_{C^0}}{\gamma(s+1) + \|w^{(n)}\|_{C^0}} + \|c_{n,s}^0\|_{C^0} \leq C$$

and

$$\|\Delta_{dt^2+d\vartheta^2} w_\perp^{(n)}\|_{C^0} \leq C(e^{-s/4} + n^{-1/2})(\gamma(s+1) + \|w^{(n)}\|_{C^0}). \tag{5-21}$$

For (5-20) we have in particular integrated (5-14) in ϑ, making use of the ϑ-invariance (see item (v) of Lemma 5.21) of the coefficients of the $n^{-1}\,\partial_t$ and $n^{-1/2}\,\partial_\vartheta^2$ terms and observing that the $n^{-1/2}\,\partial_\vartheta^2$ term integrates to zero because of (5-17); for (5-21) we have made use of the fact that

$$\|\Delta_{dt^2+d\vartheta^2} w_\perp^{(n)}\|_{C^0} \leq 2\|\Delta_{dt^2+d\vartheta^2} w^{(n)}\|_{C^0}$$

and then appealed to (5-14).

To complete the analysis we will need some basic estimates for

$$\Delta_{dt^2+d\vartheta^2} = \partial_t^2 + \partial_\vartheta^2$$

on $R_{n,s}$. For any bounded (real-valued) function f on $R_{n,s}$ and for each nonnegative integer κ, let us define on $[s, \sqrt{n}]$ the Fourier coefficients f_κ by

$$f_\kappa(t) := \begin{cases} \frac{1}{\pi}\int_{-\pi/2}^{\pi/2} f(t,\vartheta)\, d\vartheta & \text{for } \kappa = 0, \\ \frac{2}{\pi}\int_{-\pi/2}^{\pi/2} f(t,\vartheta)\cos\kappa(\vartheta - \pi/2)\, d\vartheta & \text{for } \kappa > 0. \end{cases}$$

Then the Fourier coefficients of any $u \in C^2(R_{n,s}, dt^2 + d\vartheta^2)$ satisfying $(\partial_\vartheta u) = 0$ at $\vartheta = \pm\pi/2$ admit the representations

$$\begin{aligned} u_0(t) &= u_0(s) + (\partial_t u_0)(\sqrt{n}) \cdot (t-s) + \int_s^t \int_{\sqrt{n}}^\tau \partial_t^2 u_0(\sigma)\, d\sigma\, d\tau \\ &= u_0(s) + (\partial_t u_0)(\sqrt{n}) \cdot (t-s) + \int_s^t \int_{\sqrt{n}}^\tau (\Delta_{dt^2+d\vartheta^2} u)_0(\sigma)\, d\sigma\, d\tau, \end{aligned} \tag{5-22}$$

$$\begin{aligned} u_{\kappa\neq 0}(t) = {} & \frac{u_\kappa(s)}{\cosh \kappa(\sqrt{n}-s)} \cosh \kappa(t-\sqrt{n}) + \frac{(\partial_t u_\kappa)(\sqrt{n})}{\kappa \cosh \kappa(\sqrt{n}-s)} \sinh \kappa(t-s) \\ & - \frac{\cosh \kappa(t-\sqrt{n})}{\kappa \cosh \kappa(\sqrt{n}-s)} \int_s^t (\Delta_{dt^2+d\vartheta^2} u)_\kappa(\tau) \sinh \kappa(\tau - s)\, d\tau \\ & - \frac{\sinh \kappa(t-s)}{\kappa \cosh \kappa(\sqrt{n}-s)} \int_t^{\sqrt{n}} (\Delta_{dt^2+d\vartheta^2} u)_\kappa(\tau) \cosh \kappa(\tau - \sqrt{n})\, d\tau. \end{aligned} \tag{5-23}$$

In particular (5-23) implies, for any $\kappa \geq 1$, the inequality

$$|u_\kappa(t)| \leq |u_\kappa(s)| + \frac{1}{\kappa} |(\partial_t u_\kappa)(\sqrt{n})| + \frac{1}{\kappa^2} \|(\Delta_{dt^2+d\vartheta^2} u)_\kappa\|_{C^0}. \tag{5-24}$$

Since u is C^2, the Fourier series $\sum_{\kappa=0}^\infty u_\kappa(t) \cos \kappa(\vartheta - \pi/2)$ converges (at least) pointwise to $u(t, \vartheta)$; furthermore (again appealing to the C^2 assumption in order to control the first two terms of (5-24)) we obtain the implication

$$\int_{-\pi/2}^{\pi/2} u(\cdot, \vartheta)\, d\vartheta = 0 \quad \Longrightarrow \quad \|u\|_{C^0} \leq C(\|u(s,\cdot)\|_{C^2(d\vartheta^2)} + \|(\partial_t u)(\sqrt{n}, \cdot)\|_{C^1(d\vartheta^2)} + \|\Delta_{dt^2+d\vartheta^2} u\|_{C^0}).$$

This last estimate in conjunction with (5-21), (5-15), and (5-19) yields

$$\|w_\perp^{(n)}\|_{C^0} \leq C(\gamma(s+1) + (e^{-\sqrt{n}/4} + e^{-s/4} + n^{-1/2}) \|w^{(n)}\|_{C^0}). \tag{5-25}$$

On the other hand, differentiating (5-22) with respect to t and applying (5-20) and (5-19), we find

$$\|\partial_t w_0^{(n)}\|_{C^0} \leq C(\gamma(s+1) + \|w^{(n)}\|_{C^0})(e^{-\sqrt{n}/4} + e^{-s/4} + n^{-3/2}) + Cn^{-1/2} \|\partial_t w_0^{(n)}\|_{C^0} \tag{5-26}$$

and therefore, by absorption,

$$\|\partial_t w_0^{(n)}\|_{C^0} \leq C(\gamma(s+1) + \|w^{(n)}\|_{C^0})(e^{-s/4} + n^{-3/2}) \tag{5-27}$$

for n sufficiently large in terms of s and the constants appearing in the above estimate. Feeding (5-27) into (5-20) and applying the result, along with (5-15) and (5-19), in (5-22), we get

$$\|w_0^{(n)}\|_{C^0} \leq C(\gamma(s+1) + (\sqrt{n} e^{-\sqrt{n}/4} + e^{-s/4} + n^{-1}) \|w^{(n)}\|_{C^0}). \tag{5-28}$$

Finally, since $\|w^{(n)}\|_{C^0} \leq \|w_0^{(n)}\|_{C^0} + \|w_\perp^{(n)}\|_{C^0}$, estimates (5-28) and (5-25) jointly imply the desired bound on the C^0 norm of $w^{(n)}$ provided we first choose s and then, in turn, n sufficiently large, in terms of the absolute constants appearing in the two estimates, to be able to absorb the $\|w^{(n)}\|_{C^0}$ terms appearing on their right-hand sides. This ends the proof. □

Lemma 5.26 (eigenvalue lower bounds on $\widehat{M}^{\Xi}_{\mathrm{fb},n}$ and $\widehat{M}^{\Sigma}_{\mathrm{fb},m}$). *For each integer $i \geq 1$,*

$$\liminf_{n\to\infty} \lambda_i^{\{z=0\},\pm}(Q_{\mathrm{N}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}}) \geq \lambda_i^{\{z=0\},\pm}(Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}}),$$
$$\liminf_{m\to\infty} \lambda_i^{\{y=z=0\},\pm}(Q_{\mathrm{N}}^{\widehat{M}^{\Sigma}_{\mathrm{fb},m}}) \geq \lambda_i^{\{y=z=0\},\pm}(Q^{\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}})$$

for each common choice of sign $\pm$ on both sides of each equation.

Proof. We give the proof for the $+$ choice on both sides of the top equation, the argument for the remaining three cases being identical in structure to this one. Fix $i \geq 1$, and, for each n, let $\{v_j^{(n)}\}_{j=1}^{i}$ be an $L^2(M^{\Xi}_{\mathrm{fb},n}, h_n^{\Xi})$ orthonormal set such that each $v_j^{(n)}$ is a j-th ($\{z=0\}$, $+$)-invariant eigenfunction of $Q_{\mathrm{N}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}}$. Fix $C > 0$, as afforded by Lemma 5.25, such that

$$\sup_n \sup_{1\leq j\leq i} (\|v_j^{(n)}\|_{C^0} + \lambda_j^{\{z=0\},+}(Q^{\widehat{M}^{\Xi}_{\mathrm{fb},n}})) \leq C.$$

Given any $\epsilon > 0$ (fixed from now on) and taking $s > 0$ and correspondingly $n_s > 0$ large enough, as afforded by Lemmas 5.21 and 5.23, we have

$$\mathscr{H}^2(h_n^{\Xi})(\Lambda_n(s)) < \epsilon, \quad \lambda_i^{\{z=0\},+}(Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}}) < \lambda_i^{\{z=0\},+}(Q_{\mathrm{N}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}(s)}) + \epsilon. \tag{5-29}$$

Now, for $n > n_s$ and any v in the span of $\{v_j^{(n)}\}_{j=1}^{i}$, we estimate

$$\|v\|^2_{L^2(M^{\Xi}_{\mathrm{fb},n}, h_n^{\Xi})} - \|v|_{M^{\Xi}_{\mathrm{fb},n}(s)}\|^2_{L^2(M^{\Xi}_{\mathrm{fb},n}(s), h_n^{\Xi})} \leq C^2 i\epsilon \|v\|^2_{L^2(M^{\Xi}_{\mathrm{fb},n}, h_n^{\Xi})},$$
$$\|\nabla_{h_n^{\Xi}} v|_{M^{\Xi}_{\mathrm{fb},n}(s)}\|_{L^2(M^{\Xi}_{\mathrm{fb},n}(s), h_n^{\Xi})} \leq \|\nabla_{h_n^{\Xi}} v\|_{L^2(M^{\Xi}_{\mathrm{fb},n}, h_n^{\Xi})},$$
$$\langle v|_{M^{\Xi}_{\mathrm{fb},n}(s)}, (\rho_n^{\Xi})^{-2}|A_n^{\Xi}|^2_{g_n^{\Xi}} v|_{M^{\Xi}_{\mathrm{fb},n}(s)}\rangle_{L^2(M^{\Xi}_{\mathrm{fb},n}(s), h_n^{\Xi})} \geq \langle v, (\rho_n^{\Xi})^{-2}|A_n^{\Xi}|^2_{g_n^{\Xi}} v\rangle_{L^2(M^{\Xi}_{\mathrm{fb},n}, h_n^{\Xi})} - 2C^2 i\epsilon \|v\|^2_{L^2(M^{\Xi}_{\mathrm{fb},n}, h_n^{\Xi})},$$

where for the last inequality we have used the fact that, on $\Lambda_n(s)$, the potential function appearing here is bounded above by 2, as is obvious from inspection of (5-8).

We conclude that, for all $n > n_s$, the set $\{v_j^{(n)}|_{M^{\Xi}_{\mathrm{fb},n}(s)}\}_{j=1}^{i}$ is linearly independent, and for all v as above we have

$$\frac{Q_{\mathrm{N}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}(s)}(v|_{M^{\Xi}_{\mathrm{fb},n}(s)}, v|_{M^{\Xi}_{\mathrm{fb},n}(s)})}{\|v|^2_{M^{\Xi}_{\mathrm{fb},n}(s)}\|^2_{L^2(M^{\Xi}_{\mathrm{fb},n}(s), h_n^{\Xi})}} \leq \frac{\lambda_j^{\{z=0\},+}(Q_{\mathrm{N}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}}) + 2C^2 i\epsilon}{1 - C^2 i\epsilon},$$

and so by virtue of the min-max characterization (2-13) for the eigenvalues, it follows that

$$\lambda_j^{\{z=0\},+}(Q_{\mathrm{N}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}(s)}) \leq \frac{\lambda_j^{\{z=0\},+}(Q_{\mathrm{N}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}}) + 2C^2 i\epsilon}{1 - C^2 i\epsilon}$$

for all $n > n_s$ and $1 \leq j \leq i$. Thus, using the second inequality in (5-29), we get in particular

$$\lambda_i^{\{z=0\},+}(Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}}) \leq \frac{\lambda_i^{\{z=0\},+}(Q_{\mathrm{N}}^{\widehat{M}^{\Xi}_{\mathrm{fb},n}}) + 2C^2 i\epsilon}{1 - C^2 i\epsilon} + \epsilon$$

for all $n > n_s$. The claim now follows since this inequality holds for all $\epsilon > 0$, with C independent of ϵ and n. □

By combining Lemmas 5.24 with 5.26, we immediately derive the following conclusion.

Corollary 5.27 (eigenvalues on $\widehat{M}^{\Xi}_{\mathrm{fb},n}$ and $\widehat{M}^{\Sigma}_{\mathrm{fb},m}$). *For each integer $i \geq 1$,*

$$\lim_{n\to\infty} \lambda_i^{\{z=0\},\pm}(Q^{\widehat{M}^{\Xi}_{\mathrm{fb},n}}_{\mathrm{N}}) = \lambda_i^{\{z=0\},\pm}(Q^{\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}}),$$
$$\lim_{m\to\infty} \lambda_i^{\{y=z=0\},\pm}(Q^{\widehat{M}^{\Sigma}_{\mathrm{fb},m}}_{\mathrm{N}}) = \lambda_i^{\{y=z=0\},\pm}(Q^{\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}})$$

for each common choice of sign $\pm$ on both sides of each equation.

Corollary 5.28 (equivariant index and nullity on M^{Ξ}_n and M^{Σ}_m). *There exist $n_0, m_0 > 0$ such that we have the following indices and nullities for all integers $n > n_0$ and $m > m_0$:*

S	G	$\mathrm{ind}_G(Q^S_{\mathrm{N}})$	$\mathrm{nul}_G(Q^S_{\mathrm{N}})$
M^{Ξ}_n	$\mathbb{P}_n$	1	0
M^{Σ}_m	$\mathbb{A}_{m+1}$	1	0

Additionally, still assuming $m > m_0$, we have the upper bound

$$\mathrm{ind}_{\mathbb{Y}_{m+1}}(Q^{M^{\Sigma}_m}_{\mathrm{N}}) + \mathrm{nul}_{\mathbb{Y}_{m+1}}(Q^{M^{\Sigma}_m}_{\mathrm{N}}) \leq 3.$$

Proof. All claims follow from the conjunction of Lemma 3.5 (to reduce to the appropriately even and odd indices and nullities on $n^{-1}M^{\Xi}_{\mathrm{fb},n}$ and $(m+1)^{-1}M^{\Sigma}_{\mathrm{fb},m}$ with Neumann boundary data), Proposition 3.11 (to dispense with the above scale factors n, $m+1$ and, more substantially, to pass from the natural metric to h^{Ξ}_n or h^{Σ}_m), Corollary 5.27 (to reduce to the appropriate indices and nullities of $\widehat{\mathbb{M}}^{\Xi}_{\mathrm{fb}}$ and $\widehat{\mathbb{M}}^{\Sigma}_{\mathrm{fb}}$), and finally Lemma 5.16 (which provides these last quantities). □

5.5. ***Proofs of Theorems 1.2 and 1.1.*** The following statement collects, from the broader analysis conducted in the previous section, those conclusions we shall need to prove the two main results stated in the introduction.

Corollary 5.29 (equivariant index and nullity upper bounds for $\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}$ and $\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}$). *There exists $m_0, n_0 > 0$ such that, for all integers $m > m_0$ and $n > n_0$, we have the bounds*

$$\mathrm{ind}_{\mathbb{A}_{m+1}}(\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}) + \mathrm{nul}_{\mathbb{A}_{m+1}}(\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}) \leq 2,$$
$$\mathrm{ind}_{\mathbb{Y}_{m+1}}(\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}) + \mathrm{nul}_{\mathbb{Y}_{m+1}}(\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}) \leq 6,$$
$$\mathrm{ind}_{\mathbb{P}_n}(\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}) + \mathrm{nul}_{\mathbb{P}_n}(\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}) \leq 2.$$

Proof. We apply item (ii) of Proposition 3.1 for the partition "into building blocks" defined in Section 5.3 (see Figure 5), in conjunction with Lemma 5.22 and Corollary 5.28 for the ancillary estimates for the index and nullity of the various blocks. We find that the three index-plus-nullity sums appearing in the statement are respectively bounded above by

$$\mathrm{ind}_{\mathbb{Y}_{m+1}}(Q^{K^{\Sigma}_m}_{\mathrm{N}}) + \mathrm{ind}_{\mathbb{A}_{m+1}}(Q^{M^{\Sigma}_m}_{\mathrm{N}}) + [\mathrm{ind}_{\mathbb{A}_{m+1}}(Q^{B^{\Sigma}_m}_{\mathrm{N}}) + \mathrm{nul}_{\mathbb{A}_{m+1}}(Q^{B^{\Sigma}_m}_{\mathrm{N}})] \leq 1+1+0 = 2,$$
$$2\,\mathrm{ind}_{\mathbb{Y}_{m+1}}(Q^{K^{\Sigma}_m}_{\mathrm{N}}) + \mathrm{ind}_{\mathbb{Y}_{m+1}}(Q^{M^{\Sigma}_m}_{\mathrm{N}}) + [\mathrm{ind}_{\mathbb{Y}_{m+1}}(Q^{B^{\Sigma}_m}_{\mathrm{N}}) + \mathrm{nul}_{\mathbb{Y}_{m+1}}(Q^{B^{\Sigma}_m}_{\mathrm{N}})] \leq 2+3+1 = 6,$$
$$\mathrm{ind}_{\mathbb{Y}_n}(Q^{K^{\Xi}_m}_{\mathrm{N}}) + [\mathrm{ind}_{\mathbb{P}_n}(Q^{M^{\Xi}_m}_{\mathrm{N}}) + \mathrm{nul}_{\mathbb{P}_n}(Q^{M^{\Xi}_m}_{\mathrm{N}})] \leq 1+1 = 2.$$

The first term in the first line arises as an upper bound for the $\mathbb{A}_{m+1}$-equivariant index of $K^{\Sigma}_m \cup \underline{\mathrm{R}}_{\{y=z=0\}} K^{\Sigma}_m$ subject to the natural (free boundary) Robin condition on the portion of its boundary in $\mathbb{S}^2$ and subject to

the homogeneous Neumann boundary condition on the remainder of the boundary. To obtain this upper bound, we have used the fact that a function on $K_m^\Sigma \cup \underline{\mathsf{R}}_{\{y=z=0\}} K_m^\Sigma$ (a disjoint union with each annulus disjoint from $\{z=0\}$) is $\mathbb{A}_{m+1}$-equivariant if and only if its restriction to K_m^Σ is $\mathbb{Y}_{m+1}$-equivariant and it is odd with respect to any one (so all) of the $m+1$ reflections through horizontal lines in $\mathbb{A}_{m+1}$. The first term of the final line is obtained in similar fashion. □

So, we are in position to fully determine the (maximally) equivariant index and nullity for the two families of free boundary minimal surfaces we constructed in [Carlotto et al. 2022b].

Proof of Theorem 1.2. We combine the upper bounds of the preceding corollary with the lower bounds from our earlier paper [Carlotto et al. 2022b], specifically with the content of Proposition 7.1 (see Remark 7.5) therein for what pertains to the index. At that stage, the fact that both nullities are zero then follows from the first and third inequality in Corollary 5.29. □

Finally, we can obtain the absolute estimates on the Morse index of the same families.

Proof of Theorem 1.1. The lower bounds have already been established: specifically, for $\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}$ this is just part of Proposition 5.4, while for $\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}$ it follows from just combining Proposition 5.4 with Proposition 5.5. For the upper bound we can apply the Montiel–Ros argument making use of the equivariant upper bounds above, as we are about to explain. In the case of $\Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0}$, the $\mathbb{P}_n$-equivariant upper bound on the Morse index (and nullity) is equivalent to an upper bound on the index and nullity on each domain $\Omega_i^n = \Xi_n^{-\mathbb{K}_0\cup\mathbb{K}_0} \cap W_i$, where $W_1, \dots, W_{4n}$ are the open domains defined, in $\mathbb{B}^3$, by the horizontal plane $\{z=0\}$ together with the n vertical planes passing through the origin and having equations $\theta = \pi/(2n) + i\pi/n$, $i = 0, 1, \dots, n-1$, (in the cylindrical coordinates defined at the beginning of Section 4), subject to Neumann conditions in the interior boundary as prescribed by Lemma 3.5. Thus the conclusion comes immediately by appealing to Corollary 3.2 given the third displayed equation of Corollary 5.29. Similarly, for $\Sigma_m^{-\mathbb{K}_0\cup\mathbb{B}^2\cup\mathbb{K}_0}$ we can interpret the second inequality in the statement of Corollary 5.29 as a statement on the index and nullity of the portions of surfaces that are contained in any of the $2(m+1)$ sets obtained by cutting with the $m+1$ vertical planes passing through the origin and having equations $\theta = \pi/(2(m+1)) + i\pi/(m+1)$, $i = 0, 1, \dots, m$, again subject to Neumann conditions. This completes the proof. □

Acknowledgements

The authors wish to express their sincere gratitude to Giada Franz for a number of conversations on themes related to the subject of the present manuscript. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 947923). The research of Schulz was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC-2044-390685587, Mathematics Münster: Dynamics–Geometry–Structure, and the Collaborative Research Centre CRC 1442, Geometry: Deformations and Rigidity. Part of this article was finalized while Carlotto was visiting the ETH-FIM, whose support and excellent working conditions are gratefully acknowledged.

References

[Ambrozio et al. 2018a] L. Ambrozio, A. Carlotto, and B. Sharp, "Comparing the Morse index and the first Betti number of minimal hypersurfaces", *J. Differential Geom.* **108**:3 (2018), 379–410. MR Zbl

[Ambrozio et al. 2018b] L. Ambrozio, A. Carlotto, and B. Sharp, "Index estimates for free boundary minimal hypersurfaces", *Math. Ann.* **370**:3-4 (2018), 1063–1078. MR Zbl

[Aronszajn 1957] N. Aronszajn, "A unique continuation theorem for solutions of elliptic partial differential equations or inequalities of second order", *J. Math. Pures Appl.* (9) **36** (1957), 235–249. MR Zbl

[Carlotto and Li 2024] A. Carlotto and C. Li, "Constrained deformations of positive scalar curvature metrics", *J. Differential Geom.* **126**:2 (2024), 475–554. MR Zbl

[Carlotto et al. 2022a] A. Carlotto, G. Franz, and M. B. Schulz, "Free boundary minimal surfaces with connected boundary and arbitrary genus", *Camb. J. Math.* **10**:4 (2022), 835–857. MR Zbl

[Carlotto et al. 2022b] A. Carlotto, M. B. Schulz, and D. Wiygul, "Infinitely many pairs of free boundary minimal surfaces with the same topology and symmetry group", 2022. To appear in Mem. Amer. Math. Soc. arXiv 2205.04861

[Choe 1990] J. Choe, "Index, vision number and stability of complete minimal surfaces", *Arch. Ration. Mech. Anal.* **109**:3 (1990), 195–212. MR Zbl

[Devyver 2019] B. Devyver, "Index of the critical catenoid", *Geom. Dedicata* **199** (2019), 355–371. MR Zbl

[Evans and Gariepy 2015] L. C. Evans and R. F. Gariepy, *Measure theory and fine properties of functions*, revised ed., CRC, Boca Raton, FL, 2015. MR Zbl

[Fischer-Colbrie 1985] D. Fischer-Colbrie, "On complete minimal surfaces with finite Morse index in three-manifolds", *Invent. Math.* **82**:1 (1985), 121–132. MR Zbl

[Folha et al. 2017] A. Folha, F. Pacard, and T. Zolotareva, "Free boundary minimal surfaces in the unit 3-ball", *Manuscripta Math.* **154**:3-4 (2017), 359–409. MR Zbl

[Franz 2022] G. Franz, *Contributions to the theory of free boundary minimal surfaces*, Ph.D. thesis, ETH Zürich, 2022. arXiv 2208.12188

[Fraser 2020] A. Fraser, "Extremal eigenvalue problems and free boundary minimal surfaces in the ball", pp. 1–40 in *Geometric analysis*, Lecture Notes in Math. **2263**, Springer, 2020. MR Zbl

[Fraser and Schoen 2011] A. Fraser and R. Schoen, "The first Steklov eigenvalue, conformal geometry, and minimal surfaces", *Adv. Math.* **226**:5 (2011), 4011–4030. MR Zbl

[Fraser and Schoen 2016] A. Fraser and R. Schoen, "Sharp eigenvalue bounds and minimal surfaces in the ball", *Invent. Math.* **203**:3 (2016), 823–890. MR Zbl

[Girouard and Lagacé 2021] A. Girouard and J. Lagacé, "Large Steklov eigenvalues via homogenisation on manifolds", *Invent. Math.* **226**:3 (2021), 1011–1056. MR Zbl

[Kapouleas 1997] N. Kapouleas, "Complete embedded minimal surfaces of finite total curvature", *J. Differential Geom.* **47**:1 (1997), 95–169. MR Zbl

[Kapouleas and Li 2021] N. Kapouleas and M. M.-c. Li, "Free boundary minimal surfaces in the unit three-ball via desingularization of the critical catenoid and the equatorial disc", *J. Reine Angew. Math.* **776** (2021), 201–254. MR Zbl

[Kapouleas and McGrath 2023] N. Kapouleas and P. McGrath, "Generalizing the linearized doubling approach, I: General theory and new minimal surfaces and self-shrinkers", *Camb. J. Math.* **11**:2 (2023), 299–439. MR Zbl

[Kapouleas and Wiygul 2020] N. Kapouleas and D. Wiygul, "The index and nullity of the Lawson surfaces $\xi_{g,1}$", *Camb. J. Math.* **8**:2 (2020), 363–405. MR Zbl

[Kapouleas and Wiygul 2023] N. Kapouleas and D. Wiygul, "Free boundary minimal surfaces with connected boundary in the 3-ball by tripling the equatorial disc", *J. Differential Geom.* **123**:2 (2023), 311–362. MR Zbl

[Kapouleas and Zou 2021] N. Kapouleas and J. Zou, "Free boundary minimal surfaces in the Euclidean three-ball close to the boundary", preprint, 2021. arXiv 2111.11308

[Karcher 1988] H. Karcher, "Embedded minimal surfaces derived from Scherk's examples", *Manuscripta Math.* **62**:1 (1988), 83–114. MR Zbl

[Karpukhin et al. 2014] M. Karpukhin, G. Kokarev, and I. Polterovich, "Multiplicity bounds for Steklov eigenvalues on Riemannian surfaces", *Ann. Inst. Fourier* (*Grenoble*) **64**:6 (2014), 2481–2502. MR Zbl

[Ketover 2016a] D. Ketover, "Equivariant min-max theory", preprint, 2016. arXiv 1612.08692

[Ketover 2016b] D. Ketover, "Free boundary minimal surfaces of unbounded genus", preprint, 2016. arXiv 1612.08691

[Li 2020] M. M.-c. Li, "Free boundary minimal surfaces in the unit ball: recent advances and open questions", pp. 401–435 in *Proceedings of the International Consortium of Chinese Mathematicians* (Guangzhou, China, 2017), edited by S. Y. Cheng et al., Int. Press, Boston, MA, 2020. MR Zbl

[Lima 2022] V. Lima, "Bounds for the Morse index of free boundary minimal surfaces", *Asian J. Math.* **26**:2 (2022), 227–252. MR Zbl

[Marques and Neves 2016] F. C. Marques and A. Neves, "Morse index and multiplicity of min-max minimal hypersurfaces", *Camb. J. Math.* **4**:4 (2016), 463–511. MR Zbl

[Marques and Neves 2018] F. C. Marques and A. Neves, "The space of cycles, a Weyl law for minimal hypersurfaces and Morse index estimates", pp. 319–329 in *Surveys in differential geometry* (Cambridge, MA, 2017), edited by H.-D. Cao et al., Surv. Differ. Geom. **22**, Int. Press, Somerville, MA, 2018. MR Zbl

[Marques and Neves 2020] F. C. Marques and A. Neves, "Applications of min-max methods to geometry", pp. 41–77 in *Geometric analysis*, Lecture Notes in Math. **2263**, Springer, 2020. MR Zbl

[Montiel and Ros 1991] S. Montiel and A. Ros, "Schrödinger operators associated to a holomorphic map", pp. 147–174 in *Global differential geometry and global analysis* (Berlin, 1990), Lecture Notes in Math. **1481**, Springer, 1991. MR Zbl

[Neves 2014] A. Neves, "New applications of min-max theory", pp. 939–957 in *Proceedings of the International Congress of Mathematicians*, *II*, edited by S. Y. Jang et al., Kyung Moon Sa, Seoul, 2014. MR Zbl

[Sargent 2017] P. Sargent, "Index bounds for free boundary minimal surfaces of convex bodies", *Proc. Amer. Math. Soc.* **145**:6 (2017), 2467–2480. MR Zbl

[Scherk 1835] H. F. Scherk, "Bemerkungen über die kleinste Fläche innerhalb gegebener Grenzen", *J. Reine Angew. Math.* **13** (1835), 185–208. MR Zbl

[Smith and Zhou 2019] G. Smith and D. Zhou, "The Morse index of the critical catenoid", *Geom. Dedicata* **201** (2019), 13–19. MR Zbl

[Taylor 1996] M. E. Taylor, *Partial differential equations, I*: *Basic theory*, Appl. Math. Sci. **115**, Springer, 1996. MR Zbl

[Tran 2020] H. Tran, "Index characterization for free boundary minimal surfaces", *Comm. Anal. Geom.* **28**:1 (2020), 189–222. MR Zbl

Received 15 Jan 2023. Revised 3 May 2024. Accepted 20 Jul 2024.

ALESSANDRO CARLOTTO: alessandro.carlotto@unitn.it
Dipartimento di Matematica, Università di Trento, Povo, Italy

MARIO B. SCHULZ: mario.schulz@unitn.it
Dipartimento di Matematica, Università di Trento, Povo, Italy

DAVID WIYGUL: davidjames.wiygul@unitn.it
Dipartimento di Matematica, Università di Trento, Povo, Italy

ANALYSIS AND PDE
Vol. 18 (2025), No. 7, pp. 1769–1804
DOI: 10.2140/apde.2025.18.1769

msp

THE FRACTAL UNCERTAINTY PRINCIPLE VIA DOLGOPYAT'S METHOD IN HIGHER DIMENSIONS

AIDAN BACKUS, JAMES LENG AND ZHONGKAI TAO

We prove a fractal uncertainty principle with exponent $\frac{1}{2}d - \delta + \varepsilon,\ \varepsilon > 0$, for Ahlfors–David regular subsets of $\mathbb{R}^d$ with dimension δ which satisfy a suitable "nonorthogonality condition". This generalizes the application of Dolgopyat's method by Dyatlov and Jin (2018) to higher dimensions. As a corollary, we get a quantitative essential spectral gap for the Laplacian on convex cocompact hyperbolic manifolds of arbitrary dimension with Zariski-dense fundamental groups.

1. Introduction

The *fractal uncertainty principle*, informally, is the assertion that a function cannot be microlocalized to a neighborhood of a fractal set in phase space. Such assertions have applications in spectral theory, where one can apply microlocal methods to show that fractal uncertainty principles imply the existence of essential spectral gaps [Dyatlov and Zahl 2016]. In particular, one can obtain $L^2 \to L^2$ bounds on the scattering resolvents of the Laplacian on convex cocompact hyperbolic manifolds, as well as improvements on the size of the maximal region in which certain zeta functions admit analytic continuation [Bourgain and Dyatlov 2018].

To make the fractal uncertainty principle more precise, we introduce the *semiclassical Fourier transform*

$$\mathscr{F}_h f(\xi) := (2\pi h)^{-d/2} \int_{\mathbb{R}^d} e^{-ix\cdot\xi/h} f(x)\,\mathrm{d}x,$$

where $h > 0$ is a small parameter. If we have sets $X,\ Y$ and we write $X_h,\ Y_h$ for the sumsets $X_h := X + B_h$, $Y_h := Y + B_h,\ B_h := B(0, h)$, then the fractal uncertainty principle for $X,\ Y$ asserts bounds of the form

$$\|1_{X_h}\mathscr{F}_h 1_{Y_h}\|_{L^2\to L^2} \lesssim h^\beta \tag{1-1}$$

for some $\beta > 0$ in the limit $h \to 0$. We are interested in the case that $X,\ Y$ are Ahlfors–David regular sets.

Definition 1.1. A compactly supported finite Borel measure μ on $\mathbb{R}^d$ is *Ahlfors–David regular* of dimension $\delta \in [0, d]$ on scales $[\alpha, \beta]$ with regularity constant $C_R \geq 1$ if, for every closed square box I with side length $r \in [\alpha, \beta]$ or closed ball I with radius $r \in [\alpha, \beta]$,

$$\mu(I) \leq C_R r^\delta$$

MSC2020: primary 28A80, 35B34; secondary 81Q50.
Keywords: fractal uncertainty principle, resonances.

and if, in addition, I is centered on a point in $X := \operatorname{supp}\mu$,

$$C_R^{-1} r^\delta \leq \mu(I).$$

In short we say that (X, μ) is δ-*regular*.

Applying Plancherel's theorem and Hölder's inequality, one can easily check that if X is δ-regular and Y is δ'-regular on scales $[h, 1]$ then

$$\|1_{X_h} \mathscr{F}_h 1_{Y_h}\|_{L^2 \to L^2} \lesssim h^{\max\left(0, \frac{1}{2}(d-\delta-\delta')\right)}; \tag{1-2}$$

this estimate is a straightforward modification of [Dyatlov 2019, (2.7)]. In fact, (1-2) is sharp if δ or δ' are either 0 or d, or if $X,\ Y$ are orthogonal line segments in $\mathbb{R}^2$.

Thus we say that $X,\ Y$ satisfy the *fractal uncertainty principle* if (1-1) holds for some

$$\beta > \max\left(0, \tfrac{1}{2}(d-\delta-\delta')\right).$$

We note that the two ranges $\delta + \delta' \geq d$ and $\delta + \delta' \leq d$ are very different and the corresponding fractal uncertainty principles usually hold for different reasons. There are several cases in which the fractal uncertainty principle is known:

(1) If $d = 1$ and $0 < \delta, \delta' < 1$, then the fractal uncertainty principle holds [Bourgain and Dyatlov 2018; Dyatlov and Jin 2018; Dyatlov and Zahl 2016].

(2) If $d < \delta + \delta' < 2d$, then the fractal uncertainty principle holds under the additional assumption that either Y can be decomposed as a product of Ahlfors–David fractals in $\mathbb{R}$ [Han and Schlag 2020] or Y is line-porous [Cohen 2023].

(3) If d is odd and $\delta,\ \delta'$ are very close to $\frac{1}{2}d$, then the fractal uncertainty principle holds [Cladek and Tao 2021].

(4) If $X,\ Y$ are arithmetic Cantor sets,[1] then the fractal uncertainty principle holds for $d = 1$ [Dyatlov and Jin 2017] and $d = 2,\ \delta + \delta' \geq 1$ under the condition that X does not contain any line [Cohen 2025].

1.1. ***The main theorem.*** In this paper we establish the fractal uncertainty principle for $0 < \delta + \delta' \leq d$ under the following additional hypothesis, which rules out the possibility that $X,\ Y$ are orthogonal line segments. For $\Phi(x, y) := -x \cdot y$, it is a quantitative form of the statement that "X and Y do not lie in submanifolds which have orthogonal tangent spaces".

Definition 1.2. Let $X, Y \subseteq \mathbb{R}^d$, and let $\Phi \in C^2(\mathbb{R}^d \times \mathbb{R}^d)$. We say that (X, Y) is Φ-*nonorthogonal* with constant $0 < c_N \leq 1$ from scales (α_0^X, α_0^Y) to (α_1^X, α_1^Y) if, for any $x_0 \in X,\ y_0 \in Y$ and $r_X \in (\alpha_0^X, \alpha_1^X)$ and $r_Y \in (\alpha_0^Y, \alpha_1^Y)$, there exists $x_1, x_2 \in X \cap B(x_0, r_X),\ y_1, y_2 \in Y \cap B(y_0, r_Y)$ such that

$$|\Phi(x_1, y_1) - \Phi(x_2, y_1) - \Phi(x_1, y_2) + \Phi(x_2, y_2)| \geq c_N r_X r_Y. \tag{1-3}$$

[1] We define these fundamental examples in Section 1.2.1, but for now the reader may view them as Cantor sets where the removed boxes have rational vertices.

The motivation for this definition is as follows: we want nonorthogonality to be visible on virtually all scales; after all, orthogonality of fractals is a local property, so we want nonorthogonal examples on most balls centered on a point in X and Y. The Ahlfors–David regularity condition guarantees that each such ball contributes roughly the same amount of fractal mass and is hence the reason why we upgrade "most" to "all". At the same time, we don't want nonorthogonal points to lie too close to each other. This is why we take the right-hand side to be $r_X r_Y$ instead of $|x_1 - x_0| \cdot |y_1 - y_0|$. One can verify that this definition of nonorthogonality generalizes the nonorthogonality hypothesis of [Dyatlov 2019, Proposition 6.5].

The nonorthogonality condition (1-3) is based on the *local nonintegrability condition* (LNI) of [Naud 2005; Stoyanov 2011], which itself can be traced back to the *uniform nonintegrability condition* of [Chernov 1998; Dolgopyat 1998]. In such papers one is concerned with the nonintegrability of the stable and unstable foliations of an Axiom A (or perhaps even Anosov) flow. Roughly speaking, given fractals $X,\ Y$, one may define two laminations (in the sense of [Thurston 1979, Chapter 8]) in $\mathbb{R}^d_x \times \mathbb{R}^d_\xi$, the *vertical lamination* $\{x \in X\}$ and *horizontal lamination* $\{\xi = \partial_x \Phi(x, y) : y \in Y\}$, and then (1-3) essentially asserts that the vertical and horizontal laminations satisfy LNI.

In order to state our result, we need one more condition which involves how the measure of a cube I varies when we double it.

Definition 1.3. A measure μ is *doubling* on scales $[h, 1]$ if there exists $C_D > 0$ such that, for every $r \in \left[h, \frac{1}{2}\right]$ and every cube I of side length r centered at $x \in \operatorname{supp} \mu$, we have $\mu(I \cdot 2) \leq C_D \mu(I)$, where $I \cdot 2$ is the cube with the same center as I and side length $2r$.

Clearly every regular measure is doubling; we highlight that our main theorem only needs to assume the measure is doubling rather than regular. It is essential that we only consider cubes centered at $x \in \operatorname{supp} \mu$ in the definition. One can compare this doubling property with the Federer property in [Dolgopyat 1998, §7], in which case the Gibbs measure is supported everywhere.

What follows is our main theorem.

Theorem 1.4. *Let $\mu_X,\ \mu_Y$ be doubling probability measures on scales $[h, 1]$ with compact supports $X \subset I_0,\ Y \subset J_0$, where $I_0, J_0 \subset \mathbb{R}^d$ are rectangular boxes with unit length. Let $\mathcal{B}_h$ be the semiclassical Fourier integral operator*

$$\mathcal{B}_h f(x) = \int_Y \exp\left(\frac{i\Phi(x, y)}{h}\right) p(x, y) f(y)\, \mathrm{d}\mu_Y(y), \tag{1-4}$$

where the phase Φ belongs to $C^3(I_0 \times J_0)$, $X,\ Y$ are Φ-nonorthogonal from scales h to 1, and the symbol p belongs to $C^1(I_0 \times J_0)$. Then there exists $\varepsilon_0 > 0$ such that

$$\|\mathcal{B}_h\|_{L^2(\mu_Y) \to L^2(\mu_X)} \lesssim h^{\varepsilon_0}.$$

We use the notation $\lesssim$ in the statements of our theorems to record the existence of a hidden constant. The constant could depend on the dimension d, the nonorthogonality constant c_N, the doubling constant C_D, the diameters of X, Y, $\|\Phi\|_{C^3}$, and $\|p\|_{C^1}$, but is *independent* of h.

If one additionally assumes $d = 1$ and that $\mu_X,\ \mu_Y$ are regular with dimension $\in (0, 1)$, then Theorem 1.4 was proven in [Dyatlov and Jin 2018], extending the method of [Dolgopyat 1998] which had already been

applied to construct spectral gaps. Using the construction of dyadic cubes in [Christ 1990], it might be possible that Theorem 1.4 can be generalized to doubling metric spaces. Since there is no immediate application for metric spaces, we have not attempted to write down the more general version.

Following the methods of [Dyatlov and Jin 2018], Theorem 1.4 implies the following fractal uncertainty principle which is interesting in the range $\frac{1}{2}(d-\delta-\delta')+\varepsilon_0>0$.

Theorem 1.5. *Let X and Y be Ahlfors–David regular sets in $\mathbb{R}^d$ which are nonorthogonal with respect to the dot product on $\mathbb{R}^d\times\mathbb{R}^d$. Assume that X is δ-regular, Y is δ'-regular, $0<\delta,\delta'<d$. Then there exists $\varepsilon_0>0$ such that*

$$\|1_{X_h}\mathscr{F}_h 1_{Y_h}\|_{L^2\to L^2}\lesssim h^{\frac{1}{2}(d-\delta-\delta')+\varepsilon_0}.$$

1.1.1. *Lower bounds on the uncertainty exponent.* If we let

$$L:=\frac{10^{14}d^3}{c_N^3}\max(1,\|\partial_{xy}^2\Phi\|_{C^1}^3), \tag{1-5}$$

then we can take in Theorem 1.4

$$\frac{1}{\varepsilon_0}\le 6\cdot 10^9 c_N^{-2}d^2(C_D(X)C_D(Y))^{4\lceil\log_2(20L^{5/3})\rceil}L^{2/3}\log L. \tag{1-6}$$

In the model case where $X=Y$ is regular, $d=1$, and $\Phi(x,y)=-xy$, we can always take $c_N=C_R^{-4/\delta}$ and $C_D=2^\delta C_R^2$, which gives a subexponential bound of the form $1/\varepsilon_0\lesssim e^{C(\delta)\log_2 C_R}$. This is because of the rather poor dependence of ε_0 on the doubling constant; if one modified our proof to use the Ahlfors–David regularity directly, they would obtain a bound of the form $1/\varepsilon_0\lesssim C_R^{O(1+1/\delta)}$, which is comparable with the bound $1/\varepsilon_0\lesssim C_R^{160/(\delta(1-\delta))}$ of [Dyatlov and Jin 2018].

In any case, it does not seem that one can use Dolgopyat's method to obtain sharp fractal uncertainty principles, which therefore remains an interesting and challenging open problem. To drive this point home, we recall that in the case $d=1$, $\delta=\frac{1}{2}$, an unpublished manuscript of Murphy claims that $1/\varepsilon_0\lesssim\log C_R\log\log C_R$ [Cladek and Tao 2021, §1].

1.1.2. *Applications to spectral gaps.* Suppose $M=\Gamma\backslash\mathbb{H}^{d+1}$ is a (noncompact) convex cocompact hyperbolic manifold and $\Lambda(\Gamma)$ is the limit set (see Section 5.2 for the definition). The Patterson–Sullivan measure μ on $\Lambda(\Gamma)$ is Ahlfors–David regular of dimension $\delta_\Gamma\in[0,d)$ [Sullivan 1979, Theorem 7]. Under the condition that Γ is Zariski dense in $G=\mathrm{SO}(d+1,1)_0$, we have that $(\Lambda(\Gamma),\mu)$ satisfies the nonorthogonality condition (1-3) for very general $\Phi(x,y)$ (see Corollary 5.4). So we have the fractal uncertainty principle for $\Lambda(\Gamma)$ with very general phase functions.

Dyatlov and Zahl [2016] showed that fractal uncertainty principles can be used to prove essential spectral gaps. Let Δ be the Laplace–Beltrami operator on M. Then the resolvent

$$R(\lambda):=\left(-\Delta-\tfrac{1}{4}d^2-\lambda^2\right)^{-1}:L^2_{\mathrm{comp}}(M)\to H^2_{\mathrm{loc}}(M)$$

is well defined for $\mathrm{Im}(\lambda)\gg 1$ with a meromorphic continuation to $\lambda\in\mathbb{C}$; see [Guillarmou 2005; Mazzeo and Melrose 1987] for (even) asymptotically hyperbolic manifolds and [Guillopé and Zworski 1995] for manifolds with constant negative curvature near infinity. Vasy [2013a; 2013b] gave a new construction of the meromorphic continuation, which is the one used in [Dyatlov and Zahl 2016].

The standard Patterson–Sullivan gap [Patterson 1976; Sullivan 1979] says

$$R(\lambda) \text{ has only finitely many poles in } \big\{\operatorname{Im}(\lambda) \geq -\max\big(0, \tfrac{1}{2}d - \delta_\Gamma\big)\big\}. \tag{1-7}$$

Moreover, there is no pole in $\big\{\operatorname{Im}(\lambda) > \delta_\Gamma - \frac{1}{2}d\big\}$, and there are conditions on δ_Γ such that $\lambda = i\big(\delta_\Gamma - \frac{1}{2}d\big)$ is the first pole (see [Sullivan 1979; Patterson 1988]). Using methods of [Dyatlov and Zahl 2016], we can improve the essential spectral gap when $\delta_\Gamma \leq \frac{1}{2}d$.

Theorem 1.6. *Let M be a noncompact convex cocompact hyperbolic $(d+1)$-dimensional manifold such that $\Gamma = \pi_1(M)$ is Zariski dense in $\mathrm{SO}(d+1,1)_0$. Let $\delta_\Gamma \in (0,d)$ be the Hausdorff dimension of the limit set $\Lambda(\Gamma)$. Then there exists $\varepsilon_0 > 0$ such that, for any $\varepsilon > 0$, $R(\lambda)$ has only finitely many poles λ with $\operatorname{Im}\lambda > \delta_\Gamma - \frac{1}{2}d - \varepsilon_0 + \varepsilon$. Moreover, for any $\chi \in C_0^\infty(M)$, there exists $C_0 = C_0(\varepsilon) > 0$ and $C = C(\varepsilon, \chi) > 0$ such that*

$$\|\chi R(\lambda)\chi\|_{L^2\to L^2} \leq C|\lambda|^{-1-2\min(0,\operatorname{Im}\lambda)+\varepsilon}, \quad |\lambda| > C_0, \quad \operatorname{Im}\lambda \geq \delta_\Gamma - \tfrac{1}{2}d - \varepsilon_0 + \varepsilon. \tag{1-8}$$

Dyatlov and Jin [2018, Theorem 2] showed Theorem 1.6 with $d = 1$ by proving Theorem 1.4 for $d = 1$ and X, Y δ-regular and applying [Dyatlov and Zahl 2016, Theorem 3]; our result is the natural higher-dimensional generalization of this theorem. Note the statement of the theorem holds for $\delta_\Gamma \in (0,d)$ in the whole range, but when $\delta_\Gamma > \frac{1}{2}d + \varepsilon_0$, our theorem says nothing more than the Lax–Phillips gap coming from unitarity. On the other hand, it improves the Lax–Phillips gap when $\delta_\Gamma < \frac{1}{2}d + \varepsilon_0$, which slightly passes the threshold $\frac{1}{2}d$.

The spectral gap in Theorem 1.6 was first proved [Naud 2005] for surfaces and generalized [Stoyanov 2011] to higher dimensions. The size of their gap is implicit but our method gives an explicit constant ε_0 as in (1-6) depending on the fractal dimension δ_Γ, the regularity constant and the nonorthogonality constant of the limit set $\Lambda(\Gamma)$. We give a method for computing nonorthogonality constants from the generators of a classical Schottky group $\Gamma \subset \mathrm{SL}(2,\mathbb{C})$ in the Appendix.

Another advantage of the method of [Dyatlov and Zahl 2016] is that we also get the resolvent estimate (1-8), which is hard to obtain using transfer operator techniques and in particular is not included in [Naud 2005; Stoyanov 2011]. The resolvent bound is useful in applications; see, e.g., [Vacossin 2023].

Corollary 1.7. *Let M be convex cocompact with Γ Zariski-dense. Let ζ_M be the* ***Selberg zeta function***

$$\zeta_M(s) = \prod_{l\in\mathcal{L}_M}\prod_{k=0}^{\infty}(1 - e^{-(s+k)l}), \quad s = \tfrac{1}{2}d - i\lambda,$$

where $\mathcal{L}_M$ consists of the lengths of all primitive closed geodesics on M (with multiplicity). Then $\zeta_M(s)$ has only finitely many singularities (i.e., zeroes or poles) in the half-plane $\{\operatorname{Re} s > \delta_\Gamma - \varepsilon_0 + \epsilon\}$ for any $\epsilon > 0$.

Proof. This follows from Theorem 1.6 and [Bunke and Olbrich 1999; Patterson and Perry 2001]. □

The spectral gap is closely related to asymptotics of closed geodesics and exponential decay of correlations, which are important and well-studied questions in dynamical systems. We list a few references.

- Chernov [1998] gave the first dynamical proof showing subexponential decay of correlations for 3-dimensional contact Anosov flows. The groundbreaking work of Dolgopyat [1998] showed exponential decay of correlations for transitive Anosov flows with jointly nonintegrable C^1 stable/unstable foliations.
- Naud [2005] applied Dolgopyat's method to establish a spectral gap for convex cocompact hyperbolic surfaces.
- Stoyanov [2008; 2011] showed exponential mixing for a general class of Axiom A flows satisfying his local nonintegrability condition.
- Sarkar and Winter [2021] used Dolgopyat's method to prove exponential mixing of the frame flow for convex cocompact hyperbolic manifolds. Chow and Sarkar [2022] extended it to locally symmetric spaces.
- It is interesting to ask what happens on hyperbolic manifolds with cusps. We direct the readers to [Li and Pan 2023; Li et al. 2023] for more details.

All the above works require certain *nonintegrability conditions* which should be thought of as the analogue of our nonorthogonality condition (1-3).

We would like to mention some other related works on the spectral gap for convex cocompact hyperbolic manifolds.

- Dyatlov and Zahl [2016], Dyatlov and Jin [2018], Bourgain and Dyatlov [2018], and Jin and Zhang [2020] proved the fractal uncertainty principle for $d = 1$ and hence gave explicit essential spectral gaps.
- Bourgain and Dyatlov [2017] used Fourier decay of the Patterson–Sullivan measure to get an essential spectral gap that only depends on δ_Γ when $d = 1$, $\delta_\Gamma \leq \frac{1}{2}$. This is generalized to Kleinian Schottky groups when $d = 2$ by Li, Naud and Pan [Li et al. 2021], but in this case the essential spectral gap will depend on δ_Γ and another quantity related to our nonorthogonality constant c_N (see [Li et al. 2021, Lemma 4.4]). See also [Khalil 2023; 2024] for a method using additive combinatorics.
- Oh and Winter [2016] showed a uniform spectral gap for a large family of congruence arithmetic surfaces, which was then generalized to arbitrary dimensions by Sarkar [2022].

1.2. *Idea of the proof.*

1.2.1. *Model problem*: *Arithmetic Cantor sets.* We first describe the problem in the model case that X, Y are arithmetic Cantor sets. Let $M \geq 3$ be an integer and $A, B \subseteq \{0, 1, \ldots, M-1\}^d$ be sets with

$$\delta_A := \frac{\log|A|}{\log(M)} \leq \tfrac{1}{2}d, \quad \delta_B := \frac{\log|B|}{\log(M)} \leq \tfrac{1}{2}d.$$

We let $N := M^k$ and define the *arithmetic Cantor sets*

$$C_{k,A} := \{a_0 + a_1 M + \cdots + a_{k-1}M^{k-1} : a_i \in A\} \subset (\mathbb{Z}/N\mathbb{Z})^d,$$

$$C_{k,B} := \{b_0 + b_1 M + \cdots + b_{k-1}M^{k-1} : b_i \in B\} \subset (\mathbb{Z}/N\mathbb{Z})^d.$$

We introduce the *discrete Fourier transform*

$$\mathcal{F}_N f(j) := N^{-d/2} \sum_{\ell \in \{0,1,\dots,N-1\}^d} \exp\Big(2\pi i j \cdot \frac{\ell}{N}\Big) f(\ell), \quad j \in (\mathbb{Z}/N\mathbb{Z})^d.$$

The fractal uncertainty principle states that there exists some $\varepsilon_0 > 0$ such that

$$\|1_{C_{k,A}} \mathcal{F}_N 1_{C_{k,B}}\|_{\ell^2 \to \ell^2} \lesssim N^{-\beta-\varepsilon_0}, \tag{1-9}$$

where $\beta := \frac{1}{2}(d - \delta_A - \delta_B)$ [Dyatlov and Jin 2017, §3]. Analyzing the Hilbert–Schmidt norm, we have

$$\|1_{C_{k,A}} \mathcal{F}_N 1_{C_{k,B}}\|_{\ell^2 \to \ell^2} \le \|1_{C_{k,A}} \mathcal{F}_N 1_{C_{k,B}}\|_{HS} = \sqrt{\frac{|A|^k |B|^k}{N^d}} = N^{-\beta}. \tag{1-10}$$

Thus, our goal is to obtain additional gain beyond β. To prove this, one can show as in [Dyatlov 2019, Lemma 6.4] that if we let

$$r_k := \|1_{C_{k,A}} \mathcal{F}_N 1_{C_{k,B}}\|_{\ell^2 \to \ell^2}$$

then $r_{k_1+k_2} \le r_{k_1} r_{k_2}$. This can be used to show that if we can get any gain at all at some scale k then we get a gain on all further levels, so we suppose for the sake of contradiction that we cannot obtain any gain at any scale, or that the inequality present in (1-10) is an equality. Then, since the Hilbert–Schmidt norm measures the square root of the sum of squares of the singular values and the operator norm measures the largest singular value, it follows that the operator $N^{d/2} 1_{C_{k,A}} \mathcal{F}_N 1_{C_{k,B}}$ must be rank 1. A simple computation then shows that the operator $N^{d/2} 1_{C_{k,A}} \mathcal{F}_N 1_{C_{k,B}}$ is the matrix $(\exp(2\pi i j \cdot \ell/N))_{j \in C_{k,A}, \ell \in C_{k,B}}$ (and is zero in the unspecified entries). Computing the determinant of 2×2 minors, we see that

$$\left| \det \begin{pmatrix} \exp(2\pi i j \cdot \ell/N) & \exp(2\pi i j' \cdot \ell/N) \\ \exp(2\pi i j \cdot \ell'/N) & \exp(2\pi i j' \cdot \ell'/N) \end{pmatrix} \right| = \left| \exp\Big(2\pi i \frac{\langle j - j', \ell - \ell' \rangle}{N}\Big) - 1 \right| = 0$$

for all $j, j' \in C_{k,A}$ and $\ell, \ell' \in C_{k,B}$. Thus, (1-9) holds as long as a *nonorthogonality* condition

$$\langle j - j', \ell - \ell' \rangle \neq 0$$

holds for some choice of $j, j' \in A,\ \ell, \ell' \in B$. If nonorthogonality is violated at all scales, then (1-9) cannot hold; see Example 1.9.

1.2.2. *Nonorthogonality and Dolgopyat's method.* Our proof and the proof of [Dyatlov and Jin 2018] lies in the continuous setting where the fractal is not necessarily self-similar. Thus, we must construct a tree of tiles that discretizes the doubling measure μ and which is regular enough so that each tile has two children which are spaced far enough apart. While very nice submultiplicativity does not hold as it does in the discrete case, we can still, via an induction on scales argument, propagate gain on one scale to gain on all scales. The key tool allowing us to obtain gain on all scales is nonorthogonality, which we formulated in (1-3); it asserts that we can find many points in the intersections of the vertical and horizontal laminations where the phase is "oscillating faster than the function $\mathcal{B}_h$ is being tested against" at every scale, and so we must obtain a gain at every scale. This technique, called *Dolgopyat's method*, has been used to obtain fractal uncertainty principles, spectral gaps, or exponential mixing in previous works, including [Dolgopyat 1998; Dyatlov and Jin 2018; Liverani 2004; Naud 2005; Stoyanov 2008; 2011; Tsujii and Zhang 2023].

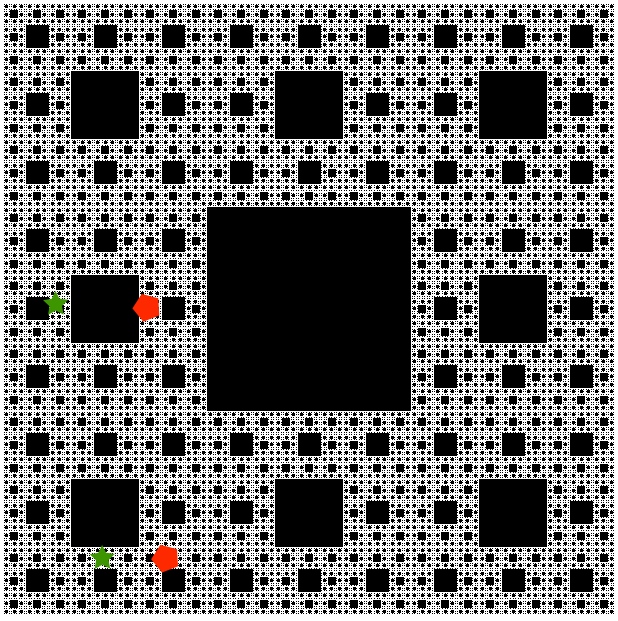

Figure 1. Nonorthogonality of the Sierpiński carpet X (the white region) to itself at scale $\frac{1}{3}$ (where $\operatorname{diam} X = \sqrt{2}$). Given any two points $x_1, y_1 \in X$ (green stars), we can find two points $x_2, y_2 \in X$ (red pentagons) such that $|x_1 - x_2|$ and $|y_1 - y_2|$ are both ≈ 0.15, and $|\sin \angle(x_2 - x_1, y_2 - y_1)| \ll 1$, so (X, X) is nonorthogonal with constant $(3 \cdot 0.14)^2 \approx 0.42$. Image adapted from [Rössel 2008].

The improvement on each child is measured in the spaces $C_\theta(I)$ that were introduced in [Naud 2005, Lemma 5.4]. Informally speaking, localizations of $\mathcal{B}_h$ to a tile I have roughly constant oscillation when normalized by $\theta \operatorname{diam}(I)$ for some appropriate choice of θ [Dyatlov and Jin 2018, §2.2]. The $C_\theta(I)$ norms are meant to capture this fact and to measure cancellation on scale I, similar to how algebraic manipulations on M^k-dimensional vectors can be used to measure cancellation in the arithmetic Cantor case.

1.2.3. *Improvements over Dyatlov–Jin.* The method of [Dyatlov and Jin 2018] does not immediately generalize to $d \geq 2$ for two reasons. First, in order to ensure that each interval has at least two children that are sufficiently far apart, Dyatlov–Jin allows intervals of varying length to appear in the tree by merging together consecutive intervals that intersect the fractal. However, in higher dimensions this leads to long, narrow, winding tiles appearing in the tree; these do not satisfy suitable doubling estimates, as exemplified by the following example.

Example 1.8. Let X be a Sierpiński carpet, and consider the merged discretization for X (see Section 3 or [Dyatlov and Jin 2018, §2.1]). Since X is path-connected, every scale consists of a single tile, the only child of the single tile at the previous scale! It is impossible to prove that every tile has two children which enjoy phase cancellation.

However, our method must be able to handle the Sierpiński carpet, since it meets the hypotheses of Theorem 1.5 if it is embedded in $\mathbb{R}^4$. Indeed, $2\delta_X \approx 3.8 < 4$. Moreover, X is nonorthogonal to itself at one scale (see Figure 1), so it is at every scale by self-similarity.

Secondly, as remarked above, one cannot obtain cancellation for arbitrary children $I_1,\ I_2$, but only those which are "not orthogonal to each other". Otherwise, even if we construct $I_1,\ I_2$ to be the appropriate distance apart to impose cancellation, it will not follow that the phases actually cancel each other.

Example 1.9. Let $X := [-5, 5] \times \{0\}$ and $Y := \{0\} \times [-5, 5]$. The Gaussian

$$f(x, y) := e^{-x^2/2 - y^2/(2h^2)}$$

is localized to X_{5h} and its Fourier transform is localized to Y_{5h}. So the fractal uncertainty principle is simply false for (X, Y), even though

$$\delta_X + \delta_Y = 2 \leq 2,$$

and we must use the nonorthogonality hypothesis somehow. One can also see that if $X' \subset X$ and $Y' \subset Y$ are fractals then the fractal uncertainty principle does not hold for (X', Y').

To overcome these difficulties, we improve on Dyatlov–Jin as follows:

(1) We carefully construct the tree, so that tiles in the tree are very close to cubes and therefore satisfy good doubling estimates, but also so that each tile contains two children with a suitable distance from each other.

(2) We prove that if $X,\ Y$ are nonorthogonal then we may choose tangent vectors to $X,\ Y$ so that the phases cannot decouple.

These goals are accomplished by Proposition 3.3, which asserts that we can construct the so-called *perturbed standard discretization* of μ, and Proposition 3.13, which asserts that many quadruples of tiles in the perturbed standard discretization have the properties above.

We found it convenient to use the language of probability theory to state Proposition 3.13, as we then could interpret the various quantities appearing in the induction on scale (Proposition 4.3) as expected values or variances of certain averages of $\mathcal{B}_h f$ taken over random tiles. The necessary estimates needed to obtain a contradiction then follow from the *second moment method* — namely, the observation that, if Proposition 4.3 is false, then the variance of such random variables is impossibly small given the large size of their tails. A similar approach was taken by [Dyatlov and Jin 2018], which used the strict convexity of balls in Hilbert spaces [Dyatlov and Jin 2018, Lemma 2.7] to accomplish the same goals.

1.3. ***Outline of the paper.*** In Section 2 we recall some preliminaries.

In Section 3 we construct our discretization and show that it has good statistical properties, as made precise by Proposition 3.13.

In Section 4 we carry out our inductive argument. The main proposition is the iterative step, Proposition 4.3; we then use this to prove Theorem 1.4.

We then turn to the applications in Section 5, where we reduce Theorems 1.5 and 1.6 to Theorem 1.4 by standard techniques.

In the Appendix, we demonstrate how one can compute the nonorthogonality constant in a typical application: classical Schottky groups.

2. Preliminaries

2.1. ***Probability theory.*** We shall have probability spaces A, B, and will denote by a, a', a'' and b, b', b'' outcomes in those spaces (or equivalently random variables with values in A, B). The expected value of a random variable X is denoted $\mathbb{E}\,X$, while $\mathbb{E}(X \mid E)$ refers to the conditional expectation of X assuming an event E. The probability of the event E is denoted $\Pr(E)$, and the variance of a random variable is

$$\operatorname{Var} X := \mathbb{E}(X^2) - (\mathbb{E}\,X)^2.$$

If $X,\ Y$ are i.i.d., then

$$\mathbb{E}\,|X-Y|^2 = \mathbb{E}\,|X|^2 + \mathbb{E}\,|Y|^2 - 2\,\mathbb{E}(XY) = 2(\mathbb{E}\,|X|^2 - (\mathbb{E}\,X)^2),$$

and so

$$\mathbb{E}\,|X-Y|^2 = 2\operatorname{Var} X. \tag{2-1}$$

We also record *Cantelli's inequality*, valid for any constant $\lambda > 0$ [Lugosi 2009, Theorem 1]:

$$\begin{aligned} \Pr(X \geq \mathbb{E}\,X + \lambda) &\leq \frac{\operatorname{Var} X}{\lambda^2 + \operatorname{Var} X}, \\ \Pr(X \leq \mathbb{E}\,X - \lambda) &\leq \frac{\operatorname{Var} X}{\lambda^2 + \operatorname{Var} X}. \end{aligned} \tag{2-2}$$

2.2. ***A geometric mean value theorem.*** We shall need an analogue of the mean value theorem for phase functions [Dyatlov and Jin 2018, Lemma 2.5]. To formulate it, we shall recall some differential geometry.

If R is a nondegenerate 2-dimensional rectangle in $\mathbb{R}^d_x \times \mathbb{R}^d_y$ and $v,\ w$ are unit vectors tangent to the edges of R, then we write $\gamma_R := v \otimes w$ for the *unit bitangent* to R and $\mathrm{d}A_R$ for the area element on R.[2] We will consider the case when $v \in \mathbb{R}^d_x$ and $w \in \mathbb{R}^d_y$. In that case, γ_R and the off-diagonal Hessian $\partial^2_{xy}\Phi$ both lie in $\mathbb{R}^d_x \otimes \mathbb{R}^d_y$, so we can consider their contraction

$$\langle \partial^2_{xy}\Phi,\, \gamma_R \rangle = \partial_v \partial_w \Phi.$$

Lemma 2.1. *Let $\Phi \in C^2(\mathbb{R}^d \times \mathbb{R}^d)$. Let $x_0, x_1, y_0, y_1 \in \mathbb{R}^d$, and let R be the rectangle with vertices $(x_i, y_j),\ i, j \in \{0, 1\}$. Then*

$$\int_R \langle \partial^2_{xy}\Phi,\, \gamma_R \rangle \,\mathrm{d}A_R = \Phi(x_0, y_0) - \Phi(x_0, y_1) - \Phi(x_1, y_0) + \Phi(x_1, y_1). \tag{2-3}$$

Proof. Both sides of (2-3) are preserved by orientation-preserving isometries which preserve the product structure on $\mathbb{R}^d \times \mathbb{R}^d$. In particular, we may take $x_0, y_0 = 0,\ x_1 = (\xi^*, 0, \ldots, 0)$, and $y_1 = (\eta^*, 0, \ldots, 0)$ for some $\xi^*, \eta^* \in \mathbb{R}$. We then set

$$\varphi(\xi, \eta) := \Phi((\xi, 0, \ldots, 0), (\eta, 0, \ldots, 0)).$$

[2] Strictly speaking, the unit bitangent should be defined using the exterior algebra, but since R is assumed nondegenerate this adds more complication for no gain.

Then by Fubini's theorem,

$$\begin{aligned}\int_R \langle \partial_{xy}^2 \Phi, \gamma_R\rangle \, \mathrm{d}A_R &= \int_0^{\xi^*}\int_0^{\eta^*} \partial_\xi \partial_\eta \varphi(\xi,\eta)\,\mathrm{d}\eta\,\mathrm{d}\xi = \int_0^{\xi^*} \partial_\xi \varphi(\xi,\eta^*) - \partial_\xi\varphi(\xi,0)\,\mathrm{d}\xi \\ &= \varphi(\xi^*,\eta^*) - \varphi(\xi^*,0) - (\varphi(0,\eta^*) - \varphi(0,0)) \\ &= \Phi(x_0,y_0) - \Phi(x_0,y_1) - \Phi(x_1,y_0) + \Phi(x_1,y_1). \qquad \square\end{aligned}$$

We now estimate the difference between (2-3) evaluated over two different rectangles R, R' by differentiating Φ along a homotopy between R, R'. This estimate will be useful when applying the nonorthogonality hypothesis.

Lemma 2.2. *Let $\Phi \in C^3(\mathbb{R}^d \times \mathbb{R}^d)$, and let $R_t = [x_0(t), x_1(t)] \times [y_0(t), y_1(t)]$, where $t = 0, 1$ and $x_i(t), y_i(t) \in \mathbb{R}^d$. Let $\gamma_t := \gamma_{R_t}$ be the unit bitangent to R_t. Assume that, for some $0 \le \varepsilon_x, \varepsilon_y, c_x, c_y \le 1$,*

(1) *for every $i \in \{0,1\}$, we have $|x_i(1) - x_i(0)| \le \varepsilon_x$ and $|y_i(1) - y_i(0)| \le \varepsilon_y$,*

(2) *for every $t \in \{0,1\}$, we have $|x_1(t) - x_0(t)| \le c_x$ and $|y_1(t) - y_0(t)| \le c_y$.*

Then

$$\left| \int_{R_1} \langle \partial_{xy}^2 \Phi, \gamma_1\rangle \, \mathrm{d}A_{R_1} - \int_{R_0} \langle \partial_{xy}^2 \Phi, \gamma_0\rangle \, \mathrm{d}A_{R_0} \right| \le 7 \|\partial_{xy}^2 \Phi\|_{C^1} (\varepsilon_x c_y + \varepsilon_y c_x). \tag{2-4}$$

Proof. By taking convex combinations, we define $x_i(t)$ and $y_i(t)$ for any $t \in [0,1]$ and hence also R_t and γ_t. Now introduce the parametrization

$$\Psi_t(\xi,\eta) := \begin{bmatrix} \xi x_1(t) + (1-\xi)x_0(t) \\ \eta y_1(t) + (1-\eta) y_0(t) \end{bmatrix} \in \mathbb{R}^d \times \mathbb{R}^d$$

which maps $[0,1]^2$ to R_t. Also let $v_t := x_1(t) - x_0(t)$ and $w_t := y_1(t) - y_0(t)$, so $|v_t||w_t|$ is the (unoriented) Jacobian of the map Ψ_t. We record for later that $|v_t| \le c_x$ and $|w_t| \le c_y$.

We estimate

$$\begin{aligned}\left| \int_{R_1} \langle \partial_{xy}^2 \Phi, \gamma_1\rangle \, \mathrm{d}A_{R_1} - \int_{R_0} \langle \partial_{xy}^2 \Phi, \gamma_0\rangle \, \mathrm{d}A_{R_0} \right| &= \left| \int_0^1 \partial_t \int_{R_t} \langle \partial_{xy}^2 \Phi, \gamma_t\rangle \, \mathrm{d}A_{R_t}\,\mathrm{d}t \right| \\ &\le \int_0^1\int_0^1\int_0^1 |\partial_t(\langle \partial_{xy}^2\Phi \circ \Psi_t(\xi,\eta), \gamma_t\rangle \cdot |v_t| \cdot |w_t|)|\,\mathrm{d}\xi\,\mathrm{d}\eta\,\mathrm{d}t.\end{aligned}$$

We next split up the above integrand:

$$\begin{aligned}&|\partial_t(\langle \partial_{xy}^2\Phi \circ \Psi_t(\xi,\eta), \gamma_t\rangle |v_t||w_t|)| \\ &\quad \le |\langle \partial_t(\partial_{xy}^2\Phi \circ \Psi_t(\xi,\eta)), \gamma_t\rangle| \cdot |v_t| \cdot |w_t| + |\langle \partial_{xy}^2\Phi \circ \Psi_t(\xi,\eta), \partial_t\gamma_t\rangle| \cdot |v_t| \cdot |w_t| \\ &\qquad\qquad + |\langle \partial_{xy}^2\Phi \circ \Psi_t(\xi,\eta), \gamma_t\rangle| \cdot |\partial_t|v_t|| \cdot |w_t| + |\langle \partial_{xy}^2\Phi \circ \Psi_t(\xi,\eta), \gamma_t\rangle| \cdot |v_t| \cdot |\partial_t|w_t|| \\ &\quad =: \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV}.\end{aligned}$$

To estimate I, we compute

$$\partial_t \Psi_t(\xi,\eta) = \begin{bmatrix} \xi(x_1(1) - x_1(0)) + (1-\xi)(x_0(1) - x_0(0)) \\ \eta(y_1(1) - y_1(0)) + (1-\eta)(y_0(1) - y_0(0)) \end{bmatrix}$$

and conclude that $\|\partial_t \Psi_t\|_{C^0} \le \varepsilon_x + \varepsilon_y$. Therefore, by the chain rule,

$$\mathrm{I} \le \|\nabla \partial_{xy}^2 \Phi\|_{C^0} \|\partial_t \Psi_t\|_{C^0} |v_t| \cdot |w_t| \le \|\partial_{xy}^2 \Phi\|_{C^1} c_x c_y (\varepsilon_x + \varepsilon_y) \le \|\partial_{xy}^2 \Phi\|_{C^1} (c_x \varepsilon_y + c_y \varepsilon_x).$$

We furthermore estimate

$$|\partial_t v_t| = |x_1(1) - x_1(0) - x_0(1) + x_0(0)| \le 2\varepsilon_x$$

and similarly for w_t. Now to estimate II, we recall

$$\gamma_t = \frac{v_t}{|v_t|} \otimes \frac{w_t}{|w_t|}.$$

By the product rule,

$$|\partial_t \gamma_t| \le \frac{2}{|v_t|} |\partial_t v_t| + \frac{2}{|w_t|} |\partial_t w_t| \le 4 \left[\frac{\varepsilon_x}{|v_t|} + \frac{\varepsilon_y}{|w_t|} \right].$$

So

$$\mathrm{II} \le 4 \|\partial_{xy}^2 \Phi\|_{C^0} (c_x \varepsilon_y + c_y \varepsilon_x) \le 4 \|\partial_{xy}^2 \Phi\|_{C^1} (c_x \varepsilon_y + c_y \varepsilon_x).$$

To estimate III, we use Kato's inequality $|\partial_t |v_t|| \le |\partial_t v_t|$ to bound

$$\mathrm{III} \le 2 \|\partial_{xy}^2 \Phi\|_{C^0} c_y \varepsilon_x \le 2 \|\partial_{xy}^2 \Phi\|_{C^1} c_y \varepsilon_x.$$

The estimate on IV is similar but with x and y swapped. Adding up these terms and integrating, we conclude the result. □

3. Discretization of sets and measures

3.1. ***A new discretization.*** As in previous works on the fractal uncertainty principle, such as [Bourgain and Dyatlov 2018; Dyatlov and Jin 2018], we will discretize fractals as trees.

Definition 3.1. Let $X \subseteq \mathbb{R}^d$ be a set. A *discretization* of X is a family $V(X) = (V_n(X))_{n \in \mathbb{Z}}$ of sets, where $V_n(X)$ is a set of nonempty subsets of $\mathbb{R}^d$ such that

- $X = \bigcup \{I \cap X : I \in V_n(X)\}$ for each n and the union is disjoint;
- for any $I \in V_n(x)$, there exist $I_k \in V_{n+1}(X)$ such that $I = \bigcup_k I_k$.

Given $I \in \bigcup_n V_n(X)$, the *height* of I is defined as $H(I) = \sup\{n : I \in V_n(X)\}$.

Definition 3.2. For a compact set $X \subset \mathbb{R}^d$ and base $L \ge 2$, its *standard L-adic discretization* $V^0 = (V_n^0)_{n \in \mathbb{Z}}$ is defined by $I \in V_n^0(X)$ if and only if

$$I = I_n(q) := [q_1, L^{-n} + q_1) \times [q_2, L^{-n} + q_2) \times \cdots \times [q_d, L^{-n} + q_d)$$

for some $q \in L^{-n}\mathbb{Z}^d$ and $I \cap X \ne \varnothing$.

The standard discretization was used in [Bourgain and Dyatlov 2018] to prove the fractal uncertainty principle in the case $d = 1$, $\delta > \frac{1}{2}$. The problem with the standard discretization is that a box in $V_n^0(X)$ may be too small for the fractal measure. Dyatlov and Jin [2018] addressed this issue in the case $d = 1$, $\delta \le \frac{1}{2}$, by considering a discretization that we call the *merged discretization*. Unfortunately, if $d \ge 2$ and

$\delta \geq 1$, then the merged discretization does not satisfy desirable estimates, as intimated by the fact that such estimates have a constant of the form $O(1)^{1/(\delta(1-\delta))}$ for $\delta < 1$ in [Dyatlov and Jin 2018].

We now construct a discretization which is more appropriate to our setting. Given a compact convex set I and a real number $\alpha > 0$, we denote by $I\alpha$ the dilation of I by α from its barycenter. For sets $A, B \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$, we use the ℓ^∞ Hausdorff distance,

$$\operatorname{dist}_\infty(x, A) := \inf_{a\in A} |a - x|_{\ell^\infty},$$
$$\operatorname{dist}_\infty(A, B) := \max\Big(\sup_{a\in A} \operatorname{dist}_\infty(a, B), \sup_{b\in B} \operatorname{dist}_\infty(b, A)\Big),$$

where, for points $x = (x_i)$ and $y = (y_i)$, we have $|x - y|_{\ell^\infty} := \max_{1\leq i\leq d} |x_i - y_i|$. Note that $\operatorname{dist}_\infty(x, A) \neq \operatorname{dist}_\infty(\{x\}, A)$ in general. We recall that for the Hausdorff distance we have the triangle inequalities: for every $x \in \mathbb{R}^d$ and every subset $X \subset \mathbb{R}^d$,

$$\operatorname{dist}_\infty(x, A) \leq \operatorname{dist}_\infty(x, B) + \operatorname{dist}_\infty(A, B), \quad \operatorname{dist}_\infty(X, A) \leq \operatorname{dist}_\infty(X, B) + \operatorname{dist}_\infty(A, B).$$

Proposition 3.3. *For every compact set $X \subset \mathbb{R}^d$, $N \in \mathbb{N}$, and $L \geq 10^3$, there is a discretization $V(X)$ of X such that, for all $I \in V_n(X)$ and $1 \leq n \leq N$,*

- *there exists $I^0 \in V_n^0(X)$ such that*

$$I^0(1 - L^{-2/3}) \subset I \subset I^0(1 + L^{-2/3}), \tag{3-1}$$

- *and there exists a point x_0 in $X \cap I$ such that*

$$\operatorname{dist}_\infty(x_0, \partial I) \geq \tfrac{1}{10} L^{-2/3-n}. \tag{3-2}$$

We call this discretization the *perturbed standard discretization*, and we call elements of the perturbed standard discretization *tiles* (to emphasize that they may not be cubes).

Remark 3.4. Christ [1990] constructed dyadic cubes with similar properties for metric spaces with a doubling measure μ as in Definition 1.3. It's possible that the construction there can also be applied to prove Theorem 1.4. Our construction is less general but does not rely on the existence of a doubling measure.

3.2. *Constructing the new discretization.*

We prove Proposition 3.3 in this section.

3.2.1. *Preliminaries.* We establish some terminology and notation that we will use in the construction of the new discretization. Let I be a cube such that $\bar{I} = [a_1, b_1] \times \cdots \times [a_d, b_d]$. For $1 \leq k \leq d$, define the k-boundary

$$\partial^k I := \bigcup_{j_1,\dots,j_k} [a_1, b_1] \times \cdots \times \{a_{j_i}, b_{j_i}\} \times \cdots \times [a_d, b_d].$$

In other words, $\partial^k I$ is the union of all $(d-k)$-dimensional faces of I, so that $\partial^1 I = \partial I$ and $\partial^d I$ is the set of all vertices of I. For a set $A \in \mathbb{R}^d$, $r > 0$, let the ℓ^∞ ball around A with radius r be

$$B_\infty(A, r) = \{x \in \mathbb{R}^d : \exists a \in A,\ |a - x|_{\ell^\infty} < r\}.$$

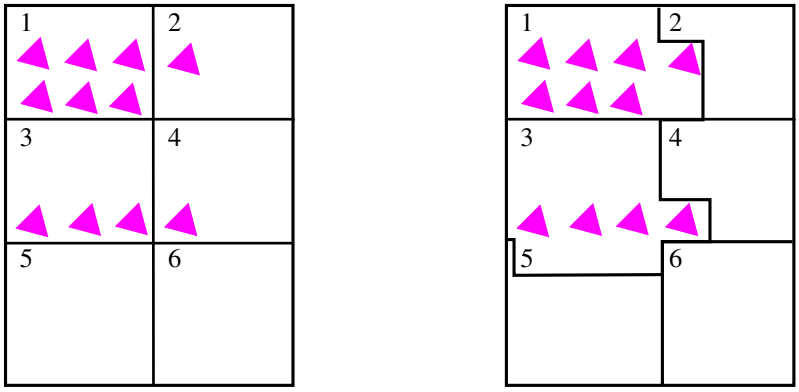

Figure 2. A standard (left) and perturbed standard (right) discretization. On the left, cube 1 is type 2, cubes 2 and 3 are type 1, cube 4 is type 0, and cubes 5 and 6 are type -1; on the right, tiles 1 and 3 are good and all other tiles are type -1.

We stress that a B without a subscript refers to the ℓ^2 ball, and in particular that the balls in the definition of nonorthogonality are ℓ^2 balls!

For a subset $P \subset \partial^k I$ of the k-boundary of a cube I, suppose without loss of generality that

$$P \subseteq \{a_1, b_1\} \times \cdots \times \{a_k, b_k\} \times [a_{k+1}, b_{k+1}] \times \cdots \times [a_d, b_d].$$

In that case, we define the *tubular neighborhoods*

$$B^t_\infty(P, r) := \big\{x \in \mathbb{R}^d : \text{ there exists } y = (y_i) \in P, \\ |x_1 - y_1| < r, \ldots, |x_k - y_k| < r, \, x_{k+1} = y_{k+1}, \ldots, x_d = y_d\big\} \tag{3-3}$$

and

$$B^t_\infty(P, r_1, r_2) := B^t_\infty(P, r_1) \cup B_\infty(P, r_2). \tag{3-4}$$

We will use these tubular neighborhoods to modify the standard cubes. Note that $B^t_\infty(P, r_1, r_2)$ has the advantage that $\operatorname{dist}_\infty(x, \partial B^t_\infty(P, r_1, r_2)) \geq \min(r_1, r_2)$ for any $x \in P$.

Let $V^0(X)$ be the standard discretization.

Definition 3.5. For every $n \leq N$ and $I \in V^0_n(X)$, the *type* of I is defined as follows:

- I is of type d if there exists a point $x \in X \cap I$ such that $\operatorname{dist}_\infty(x, \partial I) > \frac{1}{2}L^{-2/3-n}$.
- If $d - 1 > 0$, I is of type $d - 1$ if there exists a point $x \in X \cap I$ with $\operatorname{dist}_\infty(x, \partial^2 I) > \frac{1}{2}L^{-2/3-n}$ but I is not of type d.
- If $d - 2 > 0$, I is of type $d - 2$ if there exists a point $x \in X \cap I$ with $\operatorname{dist}_\infty(x, \partial^3 I) > \frac{1}{2}L^{-2/3-n}$ but I is not of type $\geq d - 1$.
- ....
- I is of type 0 if $X \cap I$ is nonempty and $\operatorname{dist}_\infty(X \cap I, \partial^d I) \leq \frac{1}{2}L^{-2/3-n}$.
- I is of type -1 if $X \cap I$ is empty.

See Figure 2 for an example of cube types.

We want to modify the cubes $I \in V_n^0(X)$ into tiles T, so that there exists $x_0 \in X \cap T$ satisfying

$$\operatorname{dist}_\infty(x_0, \partial T) \geq \tfrac{1}{5} L^{-2/3-n}. \tag{3-5}$$

We say that a tile T is *good* if (3-5) holds and otherwise that it is *bad*. For the remainder of the proof, we assume the following.

Invariant 3.6. *If a tile T constructed from a cube I is bad, then $T \subseteq I$.*

This invariant is true at the current stage of the proof; we necessarily have $T = I$ since we have not modified any tiles yet.

We want to do induction on the type of the tiles. In order to do so, we will need a notion of "type" for a bad tile. By Invariant 3.6, in order for type to be well defined, it suffices to define the type of a tile T which was modified from a cube I such that $T \subseteq I$.

Definition 3.7. Let T be a bad tile which was modified from a cube I such that $T \subseteq I$. Assume that I has type k with respect to $X \cap T$; that is, assume that I has type k in $V_n(X \cap T)$, where $V(X \cap T)$ consists of the restriction of elements of $V(X)$ that we are already defined to T. Then the *type* of T is k.

3.2.2. *Induction on type.* We now induct backwards on the largest type k of a bad tile. At every stage of the induction, we iterate over all bad tiles of type k. At each stage of this iteration, either we do nothing, or we replace a tile T with $T \cup \mathcal{T}$ for a tubular neighborhood $\mathcal{T}$ of some set. In the latter case, we replace all other tiles T' with $T' \setminus \mathcal{T}$. Here a tubular neighborhood $\mathcal{T}$ is always of the form (3-3) or (3-4). In short, we say that we *moved* the set $\mathcal{T}$. It will be important that we keep track of which sets we have already moved. As such, we make the following inductive assumptions, which are vacuous at the start of the inductive process when $k = d - 1$:

Invariant 3.8. *Every bad tile has type $\leq k$.*

Invariant 3.9. *If a tile T was constructed from a cube I, then* $\operatorname{dist}_\infty(\partial T, \partial I) \leq \frac{1}{2} L^{-n-2/3}$.

Invariant 3.10. *Let $\mathcal{T}$ be the tubular neighborhood of a set P. If $\mathcal{T}$ has been moved, then* $\dim P \geq k$.

Lemma 3.11. *Assume that $0 \leq k \leq d - 1$ and the above set of tiles satisfies Invariants 3.6, 3.8, 3.9, and 3.10. Then we may modify each tile to obtain a new set of tiles satisfying Invariants 3.6, 3.8, 3.9, and 3.10 but with k replaced by $k - 1$.*

Proof. Let T be a bad tile of type k modified from some cube I, and let P be a connected component of $\partial^{d-k} I \setminus B_\infty\big(\partial^{d-k+1} I, \frac{1}{2} L^{-2/3-n}\big)$ such that $B^t\big(P, \frac{1}{5} L^{-n-2/3}\big) \cap X \cap T \neq \varnothing$. We modify the adjacent tiles to P:

(1) If there is a good tile $T' \neq T$ adjacent to P, then we enlarge T' to contain the tubular neighborhood $\mathcal{T} := B_\infty^t\big(P, \frac{1}{2} L^{-2/3-n}\big) \cap (T' \cup T)$. Then:
   (a) T' is still good.
   (b) T no longer contains P.
   (c) Since $\mathcal{T}$ is contained in $T' \cup T$, no other tile is affected.

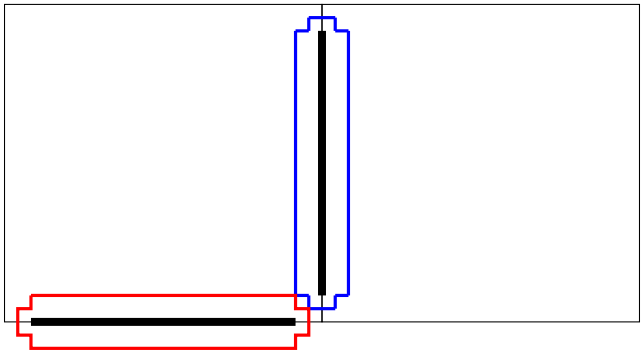

Figure 3. The proof of Lemma 3.11, Case (2). The boldest black lines represent components P of the boundary. The tubular neighborhoods around them do not intersect.

(2) Otherwise, by Invariant 3.8, every tile adjacent to P has type $\leq k$. In this case, we enlarge T by a tubular neighborhood $\mathcal{T} := B^t_\infty\left(P, \frac{1}{2}L^{-2/3-n}, \frac{1}{4}L^{-2/3-n}\right)$. Then:

(a) Let $\mathcal{T}'$ be a tubular neighborhood of a set P' which we already moved. Then:

(i) If $\dim P' > k$, then we claim that $\mathcal{T}$ is disjoint from $\mathcal{T}'$. If this is not true, then let T' be the tile containing $\mathcal{T}'$. Then T' is adjacent to P. By Invariant 3.6, T' is good, which contradicts the fact that we are in Case (2).

(ii) If $\dim P' = k$, then $\mathcal{T}$ is disjoint from $\mathcal{T}'$. See Figure 3.

(iii) We cannot have $\dim P' < k$, by Invariant 3.10.

Therefore $\mathcal{T}$ is disjoint from all tubular neighborhoods which were already moved.

(b) T becomes good.

(c) Every tile $T' \neq T$ adjacent to P no longer contains P.

We iterate the above procedure over all possible components P, stopping once there are no more components to consider. This happens after finitely many stages because of the following facts:

(1) If a tubular neighborhood of a component P is absorbed by a tile T of type k and its other neighboring tile is T', then T becomes good, and P can no longer witness that T' has type $\geq k$. Therefore we will not iterate over P again.

(2) At each stage, no new bad tiles are created, and no bad tiles are given more points and remain bad. Therefore Invariants 3.6 and 3.8 are preserved.

(3) Invariant 3.9 is preserved because if T was constructed from I then we only modify T in a neighborhood of distance $\frac{1}{2}L^{-n-2/3}$ of ∂I.

(4) Invariant 3.10 is preserved because we only moved tubular neighborhoods of sets of dimension k.

After iterating over all possible components P, Invariant 3.8 is improved, so that every bad tile has type $\leq k-1$. Indeed, if T is still bad and was type k, then every tubular neighborhood of a component P which could witness that T had type k was absorbed into a neighboring tile, so T must have type $\leq k-1$. □

After stage $k = 0$, every bad tile has type -1 by Invariant 3.8. However, if T is a tile of type -1, then by definition $X \cap T \cap I$ is empty. Then, by Invariant 3.6, $X \cap T$ is empty, and we may discard the tile T entirely.

Let $\widetilde{V}_n(X)$ be the set of good tiles that were constructed from $V_n(X)$ by the above procedure. Then every tile in $\widetilde{V}_n(X)$ satisfies (3-5) and

$$X = \bigsqcup_{T \in \widetilde{V}_n(X)} T \cap X.$$

However, $\widetilde{V}(X)$ may not have a tree structure, so it is not a discretization.

3.2.3. *Obtaining a tree structure.* We now modify $\widetilde{V}(X)$ to be a discretization $V(X)$. We again proceed by induction. For $n > N$, let $V_n(X) = V_n^0(X)$. Now suppose that $n \le N$ and we have constructed $(V_m(X))_{m \ge n+1}$ to be a discretization of X. For each element $T \in \widetilde{V}_n(X)$, we define subsets $\mathcal{C}(T)$ of $V_{n+1}(X)$ as follows:

- The subsets $\mathcal{C}(T)$ are all disjoint and their disjoint union is $V_{n+1}(X)$.
- If $S \in V_{n+1}(X)$ and $S \subseteq T$, then $S \in \mathcal{C}(T)$.
- If $S \in V_{n+1}(X)$ intersects multiple T, then we pick one T for which S lies in $\mathcal{C}(T)$.

We now define $V_n(X) = \{\bigcup_{S \in \mathcal{C}(T)} S : T \in \widetilde{V}_n(X)\}$. Thus, for each $I \in V_n(X)$, there exists an element $T \in \widetilde{V}_n(X)$ such that

$$\operatorname{dist}_\infty(\partial T, \partial I) \le 2L^{-n-1} \le \tfrac{1}{10} L^{-n-2/3}$$

(where the second inequality is because $L \ge 10^3$), and, for $x \in T$ satisfying (3-5), $x \in I$. Then, for every $x \in X$, there exists a unique $I' \in V_{n+1}(X)$ containing x by our inductive assumption, and a unique $I \in V_n(X)$ which is a superset of I' by the fact that $\{\mathcal{C}(T) : T \in \widetilde{V}_n(X)\}$ is a partition of $V_{n+1}(X)$. It follows that $(V_m(X))_{m \ge n}$ is a discretization of X.

By construction, there exists $x_0 \in X \cap T$ satisfying (3-5); hence

$$\operatorname{dist}_\infty(x_0, \partial I) \ge \operatorname{dist}_\infty(x_0, \partial T) - \operatorname{dist}_\infty(\partial I, \partial T) \ge \left(\tfrac{1}{5} - \tfrac{1}{10}\right) L^{-n-2/3} = \tfrac{1}{10} L^{-n-2/3},$$

and hence x_0 satisfies (3-2). If we denote by I^0 the cube that we modified to create T, then, by Invariant 3.9,

$$\operatorname{dist}_\infty(\partial I^0, \partial I) \le \operatorname{dist}_\infty(\partial I^0, \partial T) + \operatorname{dist}_\infty(\partial T, \partial I) \le \left(\tfrac{1}{2} + \tfrac{1}{10}\right) L^{-n-2/3},$$

which one can use to show (3-1). This completes the proof of Proposition 3.3.

3.3. ***Regularity of the discretization.*** We now show that if the compact set X is the support of a doubling measure then its perturbed standard discretization $V(X)$ satisfies regularity conditions similar to those established in [Dyatlov and Jin 2018, Lemma 2.1] for the merged discretization in the case $d = 1$.

We begin by showing that every pair of tiles $(I, J) \in V_n(X) \times V_m(Y)$ has children which contain points for which the estimate

$$|\Phi(x_0, y_0) - \Phi(x_0, y_1) - \Phi(x_1, y_0) + \Phi(x_1, y_1)| \gtrsim |x_0 - x_1| \cdot |y_0 - y_1|$$

holds. This is the key new estimate needed in the higher-dimensional case.

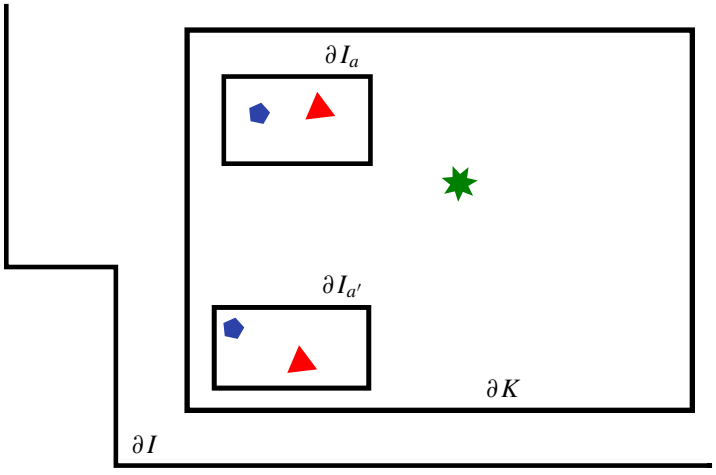

Figure 4. A typical situation in the proof of Lemma 3.12. The child tiles $I_a,\ I_{a'}$ are contained in the cube $K \subset I$ and are much smaller than I. The green 7-pointed star denotes $\underline{x}$, the red triangles denote $\tilde{x}_a,\ \tilde{x}_{a'}$, and the blue pentagons denote $x_a,\ x_{a'}$.

Lemma 3.12. *Let $\Phi \in C^3(\mathbb{R}^d \times \mathbb{R}^d)$, and let $X, Y \subseteq \mathbb{R}^d$ be Φ-nonorthogonal with constant c_N from scales (L^{-K_X}, L^{-K_Y}) to 1. Let $V(X),\ V(Y)$ be the perturbed standard discretizations of $X,\ Y$. Then, for*

$$L \geq \max(180^3, 10^{10} c_N^{-3} \|\partial_{xy}^2 \Phi\|_{C^1}^3) d^{3/2} \tag{3-6}$$

and every $n < K_X,\ m < K_Y,\ I \in V_n(X),\ J \in V_m(Y)$, there exist children $I_a,\ I_{a'}$ of I and $J_b,\ J_{b'}$ of J such that, for every $x_\alpha \in I_\alpha,\ y_\beta \in J_\beta$, and $\omega_{\alpha\beta} := \Phi(x_\alpha, y_\beta)$, we have

$$\frac{c_N}{1000} \leq L^{m+n+4/3} |\omega_{ab} - \omega_{a'b} - \omega_{ab'} + \omega_{a'b'}| \leq \frac{\|\partial_{xy}^2 \Phi\|_{C^0}}{20} \tag{3-7}$$

and

$$L^{n+2/3} |x_a - x_{a'}|,\ L^{m+2/3} |y_b - y_{b'}| \leq \tfrac{1}{2}. \tag{3-8}$$

Moreover, we may assume:

$$\text{for any } x_a \in I_a \text{ and } x_{a'} \in I_{a'},\ \text{the line segment } \overline{x_a x_{a'}} \text{ always lies in } I. \tag{3-9}$$

Proof. By Proposition 3.3, we may choose $\underline{x} \in X \cap I$ and $\underline{y} \in Y \cap J$ such that

$$\min(L^{n+2/3} \text{dist}_\infty(\underline{x}, \partial I),\ L^{m+2/3} \text{dist}_\infty(\underline{y}, \partial J)) \geq \tfrac{1}{10}.$$

Let $r_X = \frac{1}{20} L^{-n-2/3}$ and $r_Y = \frac{1}{20} L^{-m-2/3}$. One can show that, if (3-6) holds, then $L \geq 20^3$ and

$$(1 + 2L^{-2/3})^4 \leq \tfrac{9}{8}. \tag{3-10}$$

Since $n \leq K_X - 1$ and $L \geq 20^3$,

$$r_X = \tfrac{1}{20} L^{-n-2/3} \geq \tfrac{1}{20} L^{-K_X} L^{1/3} \geq L^{-K_X},$$

and similarly $r_Y \geq L^{-K_Y}$. By nonorthogonality, there exist $\tilde{x}_a, \tilde{x}_{a'} \in X \cap B(\underline{x}, r_X)$ and $\tilde{y}_b, \tilde{y}_{b'} \in Y \cap B(\underline{y}, r_Y)$ such that for $\tilde{\omega}_{\alpha\beta} := \Phi(\tilde{x}_\alpha, \tilde{y}_\beta)$,

$$|\tilde{\omega}_{ab} - \tilde{\omega}_{a'b} - \tilde{\omega}_{ab'} + \tilde{\omega}_{a'b'}| \geq c_N r_X r_Y. \tag{3-11}$$

In the other direction, (2-3) and the triangle inequality gives

$$|\tilde{\omega}_{ab} - \tilde{\omega}_{a'b} - \tilde{\omega}_{ab'} + \tilde{\omega}_{a'b'}| \leq \|\partial_{xy}^2 \Phi\|_{C^0} \cdot |\tilde{x}_a - \tilde{x}_{a'}| \cdot |\tilde{y}_b - \tilde{y}_{b'}|. \tag{3-12}$$

Let I_α be the child of I containing $\tilde{x}_\alpha$ and J_β be the child of J containing $\tilde{y}_\beta$. Pick arbitrary points $x_\alpha \in I_\alpha$ and $y_\beta \in J_\beta$. We first use (3-1), (3-10), and (3-6) to bound

$$\begin{aligned} |x_a - x_{a'}| &\leq 2r_X + \operatorname{diam} I_a + \operatorname{diam} I_{a'} \\ &\leq \tfrac{1}{10} L^{-n-2/3} + 2d^{1/2} L^{-n-1}(1 + 2L^{-2/3})^2 \\ &\leq \tfrac{1}{10} L^{-n-2/3} + 5d^{1/2} L^{-n-1} \\ &\leq \tfrac{1}{2} L^{-n-2/3}. \end{aligned}$$

A similar estimate holds on $|y_b - y_{b'}|$, which proves the upper bound in (3-8).

To prove (3-7), let $c_x := 2r_X,\ c_y := 2r_Y,\ \varepsilon_x := \max(\operatorname{diam} I_a, \operatorname{diam} I_{a'}),\ \varepsilon_y := \max(\operatorname{diam} J_b, \operatorname{diam} J_{b'})$. Then, by (2-3), Lemma 2.2, (3-1), (3-10), and (3-6),

$$\begin{aligned} |\omega_{ab} - \omega_{ab'} - \omega_{a'b} + \omega_{a'b'} - \tilde{\omega}_{ab} + \tilde{\omega}_{ab'} + \tilde{\omega}_{a'b} - \tilde{\omega}_{a'b'}| &\leq 7\|\partial_{xy}^2 \Phi\|_{C^1}(c_x \varepsilon_y + c_y \varepsilon_x) \\ &\leq \tfrac{7}{5}\|\partial_{xy}^2 \Phi\|_{C^1}(d^{1/2} L^{-n-m-5/3}(1 + 2L^{-2/3})^2) \\ &\leq 2\|\partial_{xy}^2 \Phi\|_{C^1} d^{1/2} L^{-n-m-5/3}. \end{aligned}$$

Combining this estimate with (3-11) and (3-6), we obtain

$$\begin{aligned} &|\omega_{ab} - \omega_{ab'} - \omega_{a'b} + \omega_{a'b'}| \\ &\qquad \geq |\tilde{\omega}_{ab} - \tilde{\omega}_{a'b} - \tilde{\omega}_{ab'} + \tilde{\omega}_{a'b'}| - |\omega_{ab} - \omega_{ab'} - \omega_{a'b} + \omega_{a'b'} - \tilde{\omega}_{ab} + \tilde{\omega}_{ab'} + \tilde{\omega}_{a'b} - \tilde{\omega}_{a'b'}| \\ &\qquad \geq \tfrac{1}{400} c_N L^{-n-m-4/3} - 2\|\partial_{xy}^2 \Phi\|_{C^1} d^{1/2} L^{-n-m-5/3} \\ &\qquad \geq \tfrac{1}{1000} c_N L^{-n-m-4/3}, \end{aligned}$$

which is the desired lower bound in (3-7). For the upper bound, since $\tilde{x}_a, \tilde{x}_{a'} \in B(\underline{x}, r_X)$ and $\tilde{y}_b, \tilde{y}_{b'} \in B(\underline{y}, r_Y)$ we use (2-3):

$$\begin{aligned} |\omega_{ab} - \omega_{ab'} - \omega_{a'b} + \omega_{a'b'}| &\leq 4\|\partial_{xy}^2 \Phi\|_{C^0}(r_X + \sqrt{d} L^{-n-1})(r_Y + \sqrt{d} L^{-m-1}) \\ &\leq \tfrac{1}{20}\|\partial_{xy}^2 \Phi\|_{C^0} L^{-n-m-4/3}. \end{aligned}$$

Finally we prove (3-9). We use (3-1), (3-10), (3-6), and the fact that $\operatorname{dist}_\infty(a, b) \leq |a - b|$ to estimate

$$\operatorname{dist}_\infty(x_a, \underline{x}) \leq \operatorname{dist}_\infty(x_a, \tilde{x}_a) + \operatorname{dist}_\infty(\tilde{x}_a, \underline{x}) \leq 2L^{-n-1} + r_X \leq \tfrac{1}{15} L^{-n-2/3}.$$

The same bound holds for $x_{a'}$, and it follows that $x_a,\ x_{a'}$ are contained in the convex set

$$K := B_\infty\big(\underline{x}, \tfrac{1}{15} L^{-n-2/3}\big).$$

In particular, $\ell := \overline{x_a x_{a'}}$ satisfies $\ell \subset K$. This implies $\ell \subset I$ since

$$\operatorname{dist}_\infty(\partial K, \partial I) \geq \operatorname{dist}_\infty(\underline{x}, \partial I) - \tfrac{1}{15} L^{-n-2/3} \geq \tfrac{1}{30} L^{-n-2/3},$$

so that $K \subseteq I$. □

We now give a probabilistic interpretation of the above lemmas. To establish notation, suppose that $I \in V_n(X)$ for some compact set X and some n. We write $\{I_a : a \in A\}$ for the set of children of I. This induces the structure of a probability space on A; namely,

$$\Pr(a) := \frac{\mu_X(I_a)}{\mu_X(I)}.$$

Proposition 3.13. *Let $\Phi \in C^3(\mathbb{R}^d \times \mathbb{R}^d)$, and suppose that L satisfies* (3-6)*. Let (X, μ_X) be doubling with constant $C_D(X)$ on scales $[L^{-K_X}, 1]$, let (Y, μ_Y) be doubling with constant $C_D(Y)$ on scales $[L^{-K_Y}, 1]$, let $V(X),\ V(Y)$ be their perturbed standard discretizations, and assume that (X, Y) is Φ-nonorthogonal with constant c_N from scales (L^{-K_X}, L^{-K_Y}) to 1, $n < K_X,\ m < K_Y,\ I \in V_n(X)$, and $J \in V_m(Y)$, and $\{I_a : a \in A\}$ and $\{J_b : b \in B\}$ are the sets of children of $I,\ J$. Furthermore, choose, for each $a \in A$ and $b \in B,\ x_a \in I_a$ and $y_b \in J_b$, and set $\omega_{ab} := \Phi(x_a, y_b)$.*

Draw independent random outcomes $a, a' \in A$ and $b, b' \in B$. Then, with probability

$$\rho \geq C_D(X)^{-2\lceil \log_2(20L^{5/3})\rceil} C_D(Y)^{-2\lceil \log_2(20L^{5/3})\rceil}, \tag{3-13}$$

we have

$$\tfrac{1}{1000} c_N L^{-1/3} \leq L^{n+m+1} |\omega_{ab} - \omega_{a'b} - \omega_{ab'} + \omega_{a'b'}| \leq \pi \tag{3-14}$$

and

$$L^{n+2/3}|x_a - x_{a'}|,\ L^{m+2/3}|y_b - y_{b'}| \leq \tfrac{1}{2}. \tag{3-15}$$

Moreover, we may assume,

$$\text{for any } x_a \in I_a \text{ and } x_{a'} \in I_{a'},\ \text{the line segment } \overline{x_a x_{a'}} \text{ always lies in } I. \tag{3-16}$$

Proof. By Lemma 3.12, there exist $a,\ b,\ a',\ b'$ satisfying (3-7) and (3-8). By definition of the perturbed standard discretization, there exists $x_* \in I_a \cap X$ with $I_* := \frac{1}{10} B_\infty(x_*, L^{-n-5/3}) \subset I_a$. Moreover,

$$I \subset B_\infty(x_0, 2L^{-n}) = I_*(20L^{5/3}).$$

Therefore,

$$\Pr(a) = \frac{\mu_X(I_a)}{\mu_X(I)} \geq \frac{\mu_X(I_*)}{\mu_X(I_*(20L^{5/3}))} \geq C_D(X)^{-\lceil \log_2(20L^{5/3})\rceil}.$$

We have analogous lower bounds on $\Pr(b)$, $\Pr(a')$, and $\Pr(b')$. Then, by independence,

$$\rho \geq \Pr(a) \Pr(a') \Pr(b) \Pr(b'),$$

which gives (3-13), and (3-7) and (3-8) clearly imply (3-15) and the lower bound on (3-14). The condition (3-16) comes from (3-9). For the upper bound we apply (3-7) and (3-6). □

4. The induction on scales

We now begin the proof of Theorem 1.4. Let $\Phi \in C^3(\mathbb{R}^d \times \mathbb{R}^d)$ and $p \in C^1(\mathbb{R}^d \times \mathbb{R}^d)$ be the phase and symbol of $\mathcal{B}_h$, and let $K := \lfloor -\log_L h \rfloor$.

Let (X, μ_X), (Y, μ_Y) be doubling with constants $C_D(X)$, $C_D(Y)$ on scales $\geq h$, let $V(X)$, $V(Y)$ be their perturbed standard discretizations, and assume that (X, Y) is Φ-nonorthogonal with constant c_N from scales (h, h) to 1.

For $I \in V_n(X)$ and $J \in V_m(Y)$, where $n + m + 1 = K$, we set

$$F_J(x) = \frac{1}{\mu_Y(J)} \int_J \exp\left(i \frac{\Phi(x, y) - \Phi(x, y_J)}{h}\right) p(x, y) f(y) \, \mathrm{d}\mu_Y(y).$$

Here y_J is the center of J^0, the box in the standard discretization associated to J. Let $\{I_a : a \in A\}$ and $\{J_b : b \in B\}$ be sets of children with their usual probability measures. Let $x_a := \arg\max_{I_a} |F_J|$ and $y_b := y_{J_b}$.

4.1. ***Mean value space.*** We need to generalize the space $C_\theta(I)$ where $d = 1$ (see [Dyatlov and Jin 2018, §2.2] and also [Naud 2005, Lemma 5.4]), which is supposed to locally measure oscillation on I whilst also being "scale-invariant".[3] This will allow us to get some gain out of the cancellation obtained from nonorthogonality while performing induction on scales.

Definition 4.1. Given $I \in V_n(X)$ and $\theta \in (0, 1)$, we define the $C_\theta(I)$ norm for functions $f \in C^1(I)$ by

$$\|f\|_{C_\theta(I)} := \max(\|f\|_{C^0(I)}, \theta \operatorname{diam}(I) \|\nabla f\|_{C^0(I)}).$$

Given $J \in V_m(Y)$, we define $\Psi_b : I \to \mathbb{R}$ as

$$\Psi_b(x) := \frac{\Phi(x, y_{J_b}) - \Phi(x, y_J)}{h}.$$

Lemma 4.2. *Let*

$$\theta \leq \frac{1}{8 \max(1, \|\partial_{xy}^2 \Phi\|_{C^0(I_{\mathrm{conv}})})}$$

(*where* I_{conv} *is the convex hull of* I) *and* $L \geq 10$. *Then, for* $f \in C_\theta(I)$,

$$\|e^{i\Psi_b} f\|_{C_\theta(I_a)} \leq \|f\|_{C_\theta(I)}. \tag{4-1}$$

Proof. Observe that if ψ is a smooth function on $I_{a,\mathrm{conv}}$ then any $f \in C_\theta(I_a)$ satisfies

$$|\nabla(e^{i\psi} f)| = |i e^{i\psi} f \nabla\psi + e^{i\psi} \nabla f| \leq |f \nabla\psi| + |\nabla f|.$$

Hence

$$\theta \operatorname{diam}(I_a) |\nabla(e^{i\Psi_b} f)(x)| \leq \theta \operatorname{diam}(I_a) \|\nabla\Psi_b\|_{C^0(I_a)} \|f\|_{C^0(I_a)} + \theta \operatorname{diam}(I_a) \|\nabla f\|_{C^0(I_a)}.$$

[3] We cannot use the space $C^1(I)$ with its norm $\|f\|_{C^1(I)} := \|f\|_{C^0(I)} + \|\nabla f\|_{C^0(I)}$, because the first and second terms in the norm will scale differently if we rescale I.

We estimate that

$$|\nabla\Psi_b(x)| = \frac{1}{h}|\partial_x(\Phi(x, y_{J_b}) - \Phi(x, y_J))| \le \frac{1}{h}|y_{J_b} - y_J|\|\partial_{xy}^2\Phi\|_{C^0(I_{\text{conv}})} \le \frac{\operatorname{diam}(J)}{h}\|\partial_{xy}^2\Phi\|_{C^0(I_{\text{conv}})}.$$

So, by hypothesis on θ and L,

$$\begin{aligned}\theta\operatorname{diam}(I_a)\|\nabla\Psi_b\|_{C^0(I_a)}\|f\|_{C^0(I_a)} &\le \theta\frac{\operatorname{diam}(I_a)\operatorname{diam}(J)}{h}\|\partial_{xy}^2\Phi\|_{C^0(I_{\text{conv}})}\|f\|_{C^0(I)}\\ &\le \theta(1+L^{-2/3})^2\|\partial_{xy}^2\Phi\|_{C^0(I_{\text{conv}})}\|f\|_{C_\theta(I)}\\ &\le \frac{1}{4}\|f\|_{C_\theta(I)}.\end{aligned}$$

In addition, by hypothesis on L,

$$\theta\operatorname{diam}(I_a)\|\nabla f\|_{C^0(I_a)} \le \frac{2}{L}\theta\operatorname{diam}(I)\|\nabla f\|_{C^0(I)} \le \tfrac{1}{4}\|f\|_{C_\theta(I)}.$$

Summing up,

$$\theta\operatorname{diam}(I_a)\|\nabla(e^{i\Psi_b}f)\|_{C^0(I_a)} \le \|f\|_{C_\theta(I)}.$$

We also trivially have

$$\|f\|_{C^0(I_a)} \le \|f\|_{C^0(I)} \le \|f\|_{C_\theta(I)},$$

which proves (4-1). □

4.2. ***Inductive step.*** Our next task is to prove the following analogue of [Dyatlov and Jin 2018, Lemma 3.2].

Proposition 4.3. *Let $I \in V_n(X)$, $J \in V_m(Y)$, where $n+m+1 = K$. Draw a random $b \in B$, and assume that* (3-14) *and* (3-15) *hold with probability ρ. Assume that*

$$L \ge \max\left(\frac{10^{12}d^3}{c_N^3\theta^{3/2}}, \frac{10^{10}\|\partial_{xy}^2\Phi\|_{C^1}^3 d^{3/2}}{c_N^3}\right), \tag{4-2}$$

$$\varepsilon_1 \le \frac{\rho^2 c_N^2}{10^9 d^2 L^{2/3}}. \tag{4-3}$$

Then we have the improvement

$$\mathop{\mathbb{E}}_{a\in A}\|F_J\|_{C_\theta(I_a)}^2 \le (1-\varepsilon_1)\mathop{\mathbb{E}}_{b\in B}\|F_{J_b}\|_{C_\theta(I)}^2. \tag{4-4}$$

By Proposition 3.13, we can always choose L and $\varepsilon_1 > 0$ such that the hypotheses of this proposition are met.

4.2.1. *The contradiction assumption.* We set up the proof of Proposition 4.3 by first recording

$$F_J = \mathop{\mathbb{E}}_{b\in B} e^{i\Psi_b}F_{J_b}. \tag{4-5}$$

We have the following lemma which is nearly identical to [Dyatlov and Jin 2018, Lemma 3.3].

Lemma 4.4. *For each $a \in A$,*

$$\|F_J\|_{C_\theta(I_a)}^2 \le \Big(\mathop{\mathbb{E}}_{b\in B}\|F_{J_b}\|_{C_\theta(I)}\Big)^2 \le \mathop{\mathbb{E}}_{b\in B}\|F_{J_b}\|_{C_\theta(I)}^2. \tag{4-6}$$

Proof. By (4-1),

$$\|e^{i\Psi_b} F_{J_b}\|_{C_\theta(I_a)} \le \|F_{J_b}\|_{C_\theta(I)}.$$

The assertions of (4-6) now follow from (4-5) and the Cauchy–Schwarz inequality. □

We set $R := \mathbb{E}_{b\in B} \|F_{J_b}\|^2_{C_\theta(I)}$. Draw $a \in A$ independently of b. Taking expectations in (4-6), we obtain

$$\sigma^2 := \mathop{\mathbb{E}}_{b\in B} \|F_{J_b}\|^2_{C_\theta(I)} - \mathop{\mathbb{E}}_{a\in A} \|F_J\|^2_{C_\theta(I_a)} \ge \mathop{\mathrm{Var}}_{b\in B} \|F_{J_b}\|_{C_\theta(I)}. \tag{4-7}$$

In particular, (4-7) can be written as

$$\sigma^2 = R - \mathop{\mathbb{E}}_{a\in A} \|F_J\|^2_{C_\theta(I_a)}.$$

If we knew that $\sigma^2 \ge \varepsilon_1 R$, then the improvement (4-4) would follow. So, we assume towards a contradiction that

$$\sigma^2 < \varepsilon_1 R. \tag{4-8}$$

Let

$$F_{ab} := F_{J_b}(x_a), \quad \omega_{ab} := \Psi_b(x_a), \quad f_{ab} := e^{i\omega_{ab}} F_{ab}. \tag{4-9}$$

Note carefully that ω_{ab} disagrees with the phase in Proposition 3.13 by a factor of h. We compute

$$F_J(x_a) = \mathop{\mathbb{E}}_{b\in B} f_{ab} \tag{4-10}$$

and, for each $a \in A$,

$$\mathop{\mathbb{E}}_{b\in B} |F_{ab}|^2 \le \mathop{\mathbb{E}}_{b\in B} \|F_{J_b}\|^2_{C_\theta(I)} = R. \tag{4-11}$$

4.2.2. *Outline of the proof.* By our contradiction assumption (4-8) and variance bound (4-7), the $C_\theta(I)$ norms of the functions F_{J_b} are all almost independent of b. One can show that f_{ab} is almost independent of b (see (4-12)). By the mean value theorem, F_{ab} does not vary too much in a (see (4-23)). However, the events (3-14) and (3-15) have positive probability, so we may condition on them without losing too much, and, after conditioning, the phases of f_{ab} and $f_{a'b'}$ cannot be too correlated by (3-14) and (3-15). So we expect cancellation between f_{ab} and $f_{a'b'}$ whenever a, a', b, b' are drawn at random by the square-root cancellation heuristic. This cancellation implies that the conditional expectation of $|F_{ab}|^2$ is both very small and comparable to R, a contradiction.

4.2.3. *Two unconditional moment estimates.* We now make two unconditional moment estimates; we shall later use Cantelli's inequality to show that weaker versions of the same moment estimates hold even when we condition on the events (3-14) and (3-15).

Lemma 4.5. *One has*

$$\mathop{\mathbb{E}}_{a\in A} \mathop{\mathrm{Var}}_{b\in B} f_{ab} \le \mathop{\mathbb{E}}_{\substack{a\in A\\ b\in B}} |F_{ab}|^2 - R + 2\sigma^2 \le 2\sigma^2, \tag{4-12}$$

$$\mathop{\mathbb{E}}_{\substack{a\in A\\ b\in B}} |F_{ab}| \ge (1-2\varepsilon_1)\sqrt{R}. \tag{4-13}$$

Proof. We follow [Dyatlov and Jin 2018, Lemma 3.5]. By Lemma 4.2, for each $a,\ b$,

$$\theta \|\nabla(e^{i\Psi_b} F_{J_b})\|_{C^0(I_a)} \operatorname{diam} I_a \le \tfrac{1}{2}\|F_{J_b}\|_{C_\theta(I)}.$$

From the definition of $C_\theta(I_a)$, (4-5), and the triangle inequality, for each $a \in A$,

$$\begin{aligned}\|F_J\|_{C_\theta(I_a)} &= \max(\|F_J\|_{C^0(I_a)}, \theta\|\nabla F_J\|_{C^0(I_a)} \operatorname{diam} I_a)\\ &\le \max\Big(\|F_J\|_{C^0(I_a)}, \theta \operatorname{diam} I_a \mathop{\mathbb{E}}_{b\in B} \|\nabla(e^{i\Psi_b} F_{J_b})\|_{C^0(I_a)}\Big)\\ &\le \max\Big(\|F_J\|_{C^0(I_a)}, \tfrac{1}{2}\mathop{\mathbb{E}}_{b\in B} \|F_{J_b}\|_{C_\theta(I)}\Big).\end{aligned}$$

We estimate the squares of the two terms in the maximum using (4-6):

$$\|F_J\|^2_{C^0(I_a)} \le \tfrac{1}{2}(\|F_J\|^2_{C^0(I_a)} + \|F_J\|^2_{C_\theta(I_a)}) \le \tfrac{1}{2}(\|F_J\|^2_{C^0(I_a)} + R)$$

and

$$\Big(\tfrac{1}{2}\mathop{\mathbb{E}}_{b\in B} \|F_{J_b}\|_{C_\theta(I)}\Big)^2 \le \tfrac{1}{4}\mathop{\mathbb{E}}_{b\in B} \|F_{J_b}\|^2_{C_\theta(I)} \le \tfrac{1}{4}R \le \tfrac{1}{2}(\|F_J\|^2_{C^0(I_a)} + R).$$

In summary, we have

$$\|F_J\|^2_{C_\theta(I_a)} \le \tfrac{1}{2}(\|F_J\|^2_{C^0(I_a)} + R). \tag{4-14}$$

After taking expectations and applying (4-7), we get

$$\mathop{\mathbb{E}}_{a\in A} \|F_J\|^2_{C^0(I_a)} \ge 2 \mathop{\mathbb{E}}_{a\in A} \|F_J\|^2_{C_\theta(I_a)} - R = R - 2\sigma^2. \tag{4-15}$$

We also record that, by (4-5), (4-10), and the fact that x_a maximizes $|F_J|$,

$$\mathop{\mathbb{E}}_{a\in A}\Big|\mathop{\mathbb{E}}_{b\in B} f_{ab}\Big| = \mathop{\mathbb{E}}_{a\in A}\Big|\mathop{\mathbb{E}}_{b\in B} e^{i\Psi_b(x_a)} F_{J_b}(x_a)\Big| = \mathop{\mathbb{E}}_{a\in A} |F_J(x_a)| = \mathop{\mathbb{E}}_{a\in A} \|F_J\|_{C^0(I_a)}. \tag{4-16}$$

Combining this fact with (4-15),

$$\mathop{\mathbb{E}}_{a\in A}\Big|\mathop{\mathbb{E}}_{b\in B} f_{ab}\Big|^2 \ge R - 2\sigma^2.$$

Therefore,

$$\mathop{\mathbb{E}}_{\substack{a\in A\\ b\in B}} |F_{ab}|^2 = \mathop{\mathbb{E}}_{\substack{a\in A\\ b\in B}} |f_{ab}|^2 = \mathop{\mathbb{E}}_{a\in A}\Big(\Big|\mathop{\mathbb{E}}_{b\in B} f_{ab}\Big|^2 + \mathop{\operatorname{Var}}_{b\in B} f_{ab}\Big) \ge R - 2\sigma^2 + \mathop{\mathbb{E}}_{a\in A} \mathop{\operatorname{Var}}_{b\in B} f_{ab}. \tag{4-17}$$

Rearranging, we obtain

$$\mathop{\mathbb{E}}_{a\in A} \mathop{\operatorname{Var}}_{b\in B} f_{ab} \le \mathop{\mathbb{E}}_{\substack{a\in A\\ b\in B}} |F_{ab}|^2 - R + 2\sigma^2. \tag{4-18}$$

Then (4-12) follows from (4-11).

To obtain (4-13), we first estimate

$$\mathop{\mathbb{E}}_{\substack{a\in A\\ b\in B}} |F_{ab}| = \mathop{\mathbb{E}}_{\substack{a\in A\\ b\in B}} |f_{ab}| \ge \mathop{\mathbb{E}}_{a\in A}\Big|\mathop{\mathbb{E}}_{b\in B} f_{ab}\Big| = \mathop{\mathbb{E}}_{a\in A} \sqrt{\mathop{\mathbb{E}}_{b\in B} |f_{ab}|^2 - \mathop{\operatorname{Var}}_{b\in B} f_{ab}}. \tag{4-19}$$

From (4-17) and (4-11), and the contradiction assumption (4-8),

$$\begin{aligned}\mathop{\mathbb{E}}_{a\in A}\sqrt{\mathop{\mathbb{E}}_{b\in B}|f_{ab}|^2-\mathop{\mathrm{Var}}_{b\in B}f_{ab}} &\geq \frac{1}{\sqrt{\max_{a\in A}\mathbb{E}_{b\in B}|f_{ab}|^2}}\mathop{\mathbb{E}}_{a\in A}\Big(\mathop{\mathbb{E}}_{b\in B}|f_{ab}|^2-\mathop{\mathrm{Var}}_{b\in B}f_{ab}\Big)\\ &\geq \mathop{\mathbb{E}}_{a\in A}\frac{\mathbb{E}_{b\in B}|f_{ab}|^2-\mathrm{Var}_{b\in B}f_{ab}}{\sqrt{R}}\geq\frac{R-2\sigma^2}{\sqrt{R}}\\ &\geq (1-2\varepsilon_1)\sqrt{R}.\end{aligned}$$

□

4.2.4. *Drawing random nonorthogonal tiles.* By (4-6) and the Cauchy–Schwarz inequality,

$$\mathop{\mathbb{E}}_{b\in B}\|F_{J_b}\|_{C_\theta(I)}\leq\sqrt{R}. \tag{4-20}$$

Let T be the event that $\|F_{J_b}\|_{C_\theta(I)}\leq 2\sqrt{R}$. By the moment bounds (4-20) and (4-7), the contradiction assumption (4-8), and Cantelli's inequality (2-2),

$$\Pr(T)>1-\varepsilon_1. \tag{4-21}$$

We let T' be the respective event for b', where $a',\ b'$ are drawn independently from $a,\ b$. From (4-12), (4-21), and (2-1), we obtain

$$\begin{aligned}\mathop{\mathbb{E}}_{\substack{a\in A\\ b,b'\in B}}(|f_{ab}-f_{ab'}|^2\mid T\cap T') &\leq \frac{1}{\Pr(T\cap T')}\mathop{\mathbb{E}}_{\substack{a\in A\\ b,b'\in B}}|f_{ab}-f_{ab'}|^2\\ &\leq\frac{2}{\Pr(T)^2}\mathop{\mathbb{E}}_{a\in A}\mathop{\mathrm{Var}}_{b\in B}f_{ab}\leq 2.5\cdot 2\sigma^2=5\sigma^2.\end{aligned} \tag{4-22}$$

If T and (3-16) hold, then, by Lemma 4.2,

$$|F_{ab}-F_{a'b}|\leq\frac{2\sqrt{R}}{\theta}L^{H(I)}|x_a-x_{a'}|. \tag{4-23}$$

Let S be the intersection of $T,\ T'$, and the events (3-14), (3-15) and (3-16). By (4-3), $\varepsilon_1\leq\frac{1}{10}\rho$, so by (4-21),

$$\frac{\Pr(S)}{\Pr(T)^2}\geq\frac{\rho-2(1-\Pr(T))}{\Pr(T)^2}\geq\frac{\rho-2\varepsilon_1}{(1-\varepsilon_1)^2}\geq\tfrac{1}{2}\rho. \tag{4-24}$$

If S holds, then by (4-23) and (3-15),

$$|F_{ab}-F_{a'b}|\leq\frac{\sqrt{R}}{L^{2/3}\theta}. \tag{4-25}$$

4.2.5. *Conditional second moment bounds.* We now use (4-24) and (4-25) to obtain lower and upper bounds on $\mathbb{E}(|F_{ab}|^2\mid S)$ which are not both tenable.

Lemma 4.6. *For* $M:=8000000$,

$$\mathop{\mathbb{E}}_{\substack{a\in A\\ b\in B}}(|F_{ab}|^2\mid S)\leq Md^2\bigg(\frac{R}{c_N^2L^{2/3}\theta}+2\frac{L^{2/3}\sigma^2}{c_N^2\rho}\bigg). \tag{4-26}$$

Proof. We take all expectations and probabilities over $a,\ a',\ b,\ b'$. Write

$$\tau := \omega_{ab} - \omega_{ab'} - \omega_{a'b} + \omega_{a'b'};$$

so if S holds then

$$|e^{i\tau} - 1|^2 \geq |\tau|^2 \geq 10^{-6} c_N^2 L^{-2/3}$$

by (3-14) and [Dyatlov and Jin 2018, Lemma 2.6]. Following [Dyatlov and Jin 2018, 19], we rewrite (recalling the notation set in (4-9))

$$\begin{aligned}|(e^{i\tau} - 1)F_{ab}| &= |e^{i(\omega_{ab} - \omega_{ab'})} F_{ab} - e^{i(\omega_{a'b} - \omega_{a'b'})} F_{ab}| \\ &= |e^{-i\omega_{ab'}}(f_{ab} - f_{ab'}) + F_{ab'} - F_{a'b'} - e^{-i\omega_{a'b'}}(f_{a'b} - f_{a'b'}) + e^{i(\omega_{a'b} - \omega_{a'b'})}(F_{a'b} - F_{ab})|.\end{aligned}$$

So by the triangle inequality in L^2,

$$\mathbb{E}(|(e^{i\tau} - 1)F_{ab}|^2 \mid S) \leq 4\,\mathbb{E}(|F_{ab} - F_{a'b}|^2 + |F_{a'b'} - F_{ab'}|^2 + |f_{ab} - f_{ab'}|^2 + |f_{a'b'} - f_{a'b}|^2 \mid S).$$

So

$$\begin{aligned}&\mathbb{E}(|F_{ab}|^2 \mid S) \\ &\quad \leq 10^6 \cdot \frac{d^2 L^{2/3}}{c_N^2}\,\mathbb{E}(|(e^{i\tau} - 1)F_{ab}|^2 \mid S) \\ &\quad \leq \frac{M d^2 L^{2/3}}{2c_N^2}\,\mathbb{E}(|F_{ab} - F_{a'b}|^2 + |F_{a'b'} - F_{ab'}|^2 \mid S) + \frac{M d^2 L^{2/3}}{2c_N^2}\,\mathbb{E}(|f_{ab} - f_{ab'}|^2 + |f_{a'b'} - f_{a'b}|^2 \mid S).\end{aligned}$$

Applying (4-25),

$$|F_{ab} - F_{a'b}|^2 + |F_{a'b'} - F_{ab'}|^2 \leq \frac{2R}{L^{4/3}\theta}.$$

Since S implies $T \cap T'$ and $a,\ a'$ are independent,

$$\mathbb{E}(|f_{ab} - f_{ab'}|^2 + |f_{a'b'} - f_{a'b}|^2 \mid S) \leq 2\frac{\Pr(T)^2}{\Pr(S)}\,\mathbb{E}(|f_{ab} - f_{ab'}|^2 \mid T \cap T').$$

By (4-24), $\Pr(T)^2/\Pr(S) \leq 2/\rho$. Summing all this up and applying (4-22), we conclude (4-26). □

Lemma 4.7. *One has*

$$\mathop{\mathbb{E}}_{\substack{a\in A\\ b\in B}}(|F_{ab}|^2 \mid S) \geq \tfrac{1}{6}R. \tag{4-27}$$

Proof. By (4-13), we conclude that

$$\Pr_{a\in A}\Big(\mathop{\mathbb{E}}_{b\in B}|F_{ab}| < (1 - 2\sqrt{\epsilon_1})\sqrt{R}\Big) \leq \sqrt{\epsilon_1}.$$

By Cantelli's inequality (2-2),

$$\Pr_{b\in B}\Big(|F_{ab}| \leq \mathop{\mathbb{E}}_{b\in B}|F_{ab}| - \tfrac{1}{2}\sqrt{R}\Big) \leq \frac{\mathrm{Var}_{b\in B}\,|F_{ab}|}{\mathrm{Var}_{b\in B}\,|F_{ab}| + \frac{1}{4}R}.$$

Since $|F_{ab}| = |f_{ab}|$, it follows from (4-12) and (4-8) that

$$\begin{aligned}\Pr\big(|F_{ab}|^2 \le \tfrac{1}{5}R\big) &\le \Pr_{a\in A}\Big(\mathop{\mathbb{E}}_{b\in B}|F_{ab}| < (1-2\sqrt{\epsilon_1})\sqrt{R}\Big) + \Pr\Big(\mathop{\mathbb{E}}_{b\in B}|F_{ab}| \ge (1-2\sqrt{\epsilon_1})\sqrt{R},\ |F_{ab}|^2 \le \tfrac{1}{5}R\Big)\\ &\le \sqrt{\epsilon_1} + \mathop{\mathbb{E}}_{a\in A}\Pr_{b\in B}\Big(|F_{ab}| \le \mathop{\mathbb{E}}_{b\in B}|F_{ab}| - \tfrac{1}{2}\sqrt{R}\Big)\\ &\le \sqrt{\epsilon_1} + \frac{4\,\mathbb{E}_{a\in A}\operatorname{Var}_{b\in B} f_{ab}}{R}\\ &\le \sqrt{\epsilon_1} + \frac{8\sigma^2}{R} < 2\sqrt{\epsilon_1}.\end{aligned}$$

But by (4-24),

$$\Pr\big(|F_{ab}|^2 \le \tfrac{1}{5}R \;\big|\; S\big) = \frac{\Pr\big(\big(|F_{ab}|^2 \le \frac{1}{5}R\big)\cap S\big)}{\Pr(S)} \le \frac{2\Pr\big(|F_{ab}|^2 \le \frac{1}{5}R\big)}{\rho}.$$

The definition (4-3) of ε_1 then implies

$$\Pr\big(|F_{ab}|^2 \le \tfrac{1}{5}R \;\big|\; S\big) \le \frac{4\sqrt{\varepsilon_1}}{\rho} < L^{-1/3}.$$

Therefore

$$\Pr\big(|F_{ab}|^2 \ge \tfrac{1}{5}R \;\big|\; S\big) \ge 1 - L^{-1/3},$$

so by Markov's inequality and the assumption (4-2),

$$\mathop{\mathbb{E}}_{\substack{a\in A\\ b\in B}}(|F_{ab}|^2 \mid S) \ge \tfrac{1}{5}R\Pr\big(|F_{ab}|^2 \ge \tfrac{1}{5}R \;\big|\; S\big) \ge \tfrac{1}{6}R. \qquad \square$$

4.2.6. *Deriving a contradiction.* The two above conditional second moment bounds contradict (4-2) and (4-3) and the contradiction assumption (4-8). To be more precise, combining (4-26) with (4-27) and (4-8), we obtain

$$\tfrac{1}{6}R \le \mathop{\mathbb{E}}_{\substack{a\in A\\ b\in B}}(|F_{ab}|^2 \mid S) \le Md^2\bigg(\frac{R}{c_N^2 L^{2/3}\theta} + \frac{2L^{2/3}\sigma^2}{c_N^2\rho}\bigg) < Md^2\bigg(\frac{R}{c_N^2 L^{2/3}\theta} + \frac{2L^{2/3}\varepsilon_1 R}{c_N^2\rho}\bigg).$$

Dividing both sides by RM and applying (4-2) and (4-3), we obtain

$$2\cdot 10^{-8} < \frac{1}{48\cdot 10^6} = \frac{1}{6M} \le \frac{d^2}{c_N^2 L^{2/3}\theta} + \frac{2d^2 L^{2/3}\varepsilon_1}{c_N^2\rho} \le \frac{1}{10^8} + \frac{2}{10^9} = 1.2\cdot 10^{-8}.$$

This is a contradiction that proves that $\sigma^2 \ge \varepsilon_1 R$ and so completes the proof of Proposition 4.3.

4.3. ***Proof of main theorem.*** To prove Theorem 1.4, we iterate Proposition 4.3. For each J, we define

$$E_J : V_{K-H(J)}(X) \to \mathbb{R}, \quad I \mapsto \|F_J\|_{C_\theta(I)}.$$

We endow $V_n(X)$ with the discrete measure induced by μ_X, namely $\mu_X(\{I\}) = \mu_X(I)$, and J with the restricted fractal measure μ_Y.

First suppose that $J \in V_K(Y)$. Then, by the Cauchy–Schwarz inequality, it follows that

$$\begin{aligned}|\nabla F_J(x)| &= \frac{1}{\mu_Y(J)} \int_J i\,\partial_x \Psi_J(x,y) \exp(i(\Psi_J(x,y)))p(x,y)f(x,y) \\ &\qquad + \exp(i(\Psi_J(x,y)))\partial_x p(x,y) f(y)\,\mathrm{d}\mu_Y(y) \\ &\leq \frac{1}{\sqrt{\mu_Y(J)}}\left(\frac{\operatorname{diam} J}{h}\|\partial_{xy}^2\Phi\|_{C^0}\|f\|_{L^2(J)}\|p\|_{C^0} + \|\partial_x p\|_{C^0}\|f\|_{L^2(J)}\right)\end{aligned}$$

and

$$\|F_J\|_{C^0} \leq \frac{\|p\|_{C^0}\|f\|_{L^2(J)}}{\sqrt{\mu_Y(J)}}.$$

Thus,

$$E_J(I) = \|F_J\|_{C_\theta(I)} \leq \frac{\|p\|_{C^1}\|f\|_{L^2(J)}}{\sqrt{\mu_Y(J)}}. \tag{4-28}$$

Taking L^2 norms of both sides of (4-28), we get

$$\|E_J\|_{L^2}^2 \leq \frac{\|p\|_{C^1}^2 \mu_X(X)}{\mu_Y(J)}\|f\|_{L^2(J)}^2. \tag{4-29}$$

If we take L^2 norms of both sides of (4-4), we get

$$\|E_J\|_{L^2}^2 \leq (1-\varepsilon_1) \mathop{\mathbb{E}}_{b\in B} \|E_{J_b}\|_{L^2}^2. \tag{4-30}$$

Inducting backwards on $H(J)$ with (4-29) as base case and (4-30) as inductive case, we conclude that, if J is a tile in Y such that $H(J) = 0$,

$$\|E_J\|_{L^2}^2 \leq \frac{\|p\|_{C^1}^2 \mu_X(X)}{\mu_Y(J)}(1-\varepsilon_1)^K\|f\|_{L^2(J)}^2.$$

Summing both sides in J, we obtain (note $\|\mathcal{B}_h(1_J f)\|_{L^2} \leq \mu_Y(J)\|E_J\|_{L^2}$ and $\mu_X(X) = \mu_Y(Y) = 1$)

$$\|\mathcal{B}_h f\|_{L^2}^2 \lesssim \|p\|_{C^1}^2 \mu_X(X)\mu_Y(Y)(1-\varepsilon_1)^K\|f\|_{L^2}^2 \lesssim (1-\varepsilon_1)^K\|f\|_{L^2}^2.$$

We now can set

$$\varepsilon_0 := \frac{\varepsilon_1}{6\log L} \leq \frac{\log(1-\varepsilon_1)^{-1}}{2\log L}$$

and plug in θ in (4-2) to obtain (1-5) and (1-6). Then $(1-\varepsilon_1)^{K/2} \leq h^{\varepsilon_0}$, so

$$\|\mathcal{B}_h\|_{L^2(\mu_Y)\to L^2(\mu_X)} \lesssim h^{\varepsilon_0},$$

which completes the proof of Theorem 1.4.

5. Applications

5.1. ***Classical fractal uncertainty principle.*** We now prove Theorem 1.5 following [Dyatlov and Jin 2018, Theorem 1, Remarks 1]. The classical version of the fractal uncertainty principle, Theorem 1.5, uses Lebesgue measure. In order to use our main Theorem 1.4, we need to define a rescaling of the Lebesgue measure. This is the content of the following lemma.

Lemma 5.1. *Let (X, μ) be δ-regular on scales $[h, 1]$, $h > 0$, where $\delta \in [0, d]$ and μ is the δ-dimensional Hausdorff measure. Let $X_h := X + B_h$ and*

$$\mu_h(A) := h^{\delta-d}|X_h \cap A|.$$

Then (X_h, μ_h) is δ-regular on scales $[2h, 1]$ with constant

$$C_R(X_h) := 6^d \max(|\mathbb{B}^d|, |\mathbb{B}^d|^{-1})C_R(X)^2.$$

Here $|X_h \cap A|$ is the Lebesgue measure of $X_h \cap A$ and $\mathbb{B}^d$ is the unit ball in $\mathbb{R}^d$.

Proof. Let $N = N_X(x, r, h)$ be the cardinality of a maximal h-separated subset of $X \cap B(x, r)$ for $x \in X$ and $r \geq 2h$. By [Dyatlov and Zahl 2016, Lemma 7.4], we have

$$C_R(X)^{-2}\frac{r^\delta}{h^\delta} \leq N_X(x, r, h) \leq C_R(X)^2\left(1 + \frac{2r}{h}\right)^\delta.$$

If $\{x_1, \ldots, x_N\}$ is such a maximal set and $I_n := B(x_n, 2h)$, then $X_h \cap B(x, r) \subseteq \bigcup_{n=1}^N I_n$, so we have

$$\mu_h(B(x, r)) \leq h^{\delta-d}\sum_{n=1}^N |I_n| \leq 2^d h^\delta |\mathbb{B}^d| N \leq 2^d |\mathbb{B}^d| C_R(X)^2 (h + 2r)^\delta \leq C_R(X_h) r^\delta.$$

Conversely, if $J_n := B\big(x_n, \frac{1}{2}h\big)$, then J_n and J_m are disjoint and $\bigcup_{n=1}^N J_n \subseteq X_h \cap B(x, r)$, so we have

$$\mu_h(B(x, r)) \geq \sum_{n=1}^N h^{\delta-d}|J_n| \geq N\frac{h^\delta}{2^d}|\mathbb{B}^d| \geq C_R(X)^{-2}2^{-d}|\mathbb{B}^d| r^\delta \geq C_R(X_h)^{-1} r^\delta. \qquad \square$$

We now show that the nonorthogonality condition also holds for (X_h, Y_h).

Lemma 5.2. *Let (X, Y) be Φ-nonorthogonal on scales $[h, 1]$, $h > 0$. Then (X_h, Y_h) is Φ-nonorthogonal on scales $[2h, 1]$ with constant $c_N(X_h, Y_h) := \frac{1}{4}c_N(X, Y)$.*

Proof. Let $x_0 \in X_h$, $y_0 \in Y_h$, and $r_X, r_Y \geq 2h$; then there exist $\tilde{x}_0 \in X$ and $\tilde{y}_0 \in Y$ with

$$\max(|x_0 - \tilde{x}_0|, |y_0 - \tilde{y}_0|) \leq h.$$

Putting $\tilde{r}_X := r_X - h$ and $\tilde{r}_Y := r_Y - h$, we can find by Φ-nonorthogonality of (X, Y) points

$$x_1, x_2 \in X \cap B(\tilde{x}_0, \tilde{r}_X) \subseteq X \cap B(x_0, r_X)$$

and

$$y_1, y_2 \in Y \cap B(\tilde{y}_0, \tilde{r}_Y) \subseteq Y \cap B(y_0, r_Y)$$

such that

$$|\Phi(x_1, y_1) - \Phi(x_1, y_2) - \Phi(x_2, y_1) + \Phi(x_2, y_2)| \geq c_N(X)\tilde{r}_X\tilde{r}_Y \geq c_N(X_h) r_X r_Y. \qquad \square$$

Proof of Theorem 1.5. We define $(X_h, \mu_{X,h})$, $(Y_h, \mu_{Y,h})$ as in Lemma 5.1 and introduce the Fourier integral operator

$$\mathcal{B}_h f(\xi) := \int_{Y_h} e^{ix\cdot\xi/h} f(x)\, \mathrm{d}\mu_{Y,h}(x).$$

By the above lemmas, $(X_h, \mu_{X,h})$ is δ-regular, $(Y_h, \mu_{Y,h})$ is δ'-regular, and (X_h, Y_h) is Φ-nonorthogonal. Thus, by Theorem 1.4,[4] there exists $\varepsilon_0 > 0$ such that

$$\|1_{X_h}\mathscr{F}_h 1_{Y_h}\|_{L^2\to L^2} = \frac{h^{(d-\delta-\delta')/2}}{(2\pi)^{d/2}}\|\mathcal{B}_h\|_{L^2(\mu_{Y,h})\to L^2(\mu_{X,h})} \lesssim h^{(d-\delta-\delta')/2+\varepsilon_0}. \qquad \square$$

5.2. ***Convex cocompact hyperbolic manifolds.*** In this section we prove Theorem 1.6. First we recall some preliminaries for convex cocompact hyperbolic manifolds.

Let $\mathbb{H}^{d+1}$ be the $(d+1)$-dimensional hyperbolic space (with constant curvature -1). The orientation preserving isometry group is given by $G = \mathrm{SO}(d+1,1)_0$. Let $K = \mathrm{SO}(d+1)$ be a maximal compact subgroup, so that $\mathbb{H}^{d+1} = G/K$. We are interested in infinite volume hyperbolic manifolds given by $M = \Gamma\backslash G/K$, where $\Gamma \subset G$ is a convex cocompact Zariski-dense torsion-free discrete subgroup.

The *limit set* is defined as $\Lambda(\Gamma) := \overline{\Gamma x} \cap \partial_\infty(\mathbb{H}^{d+1}) \subset \overline{\mathbb{H}^{d+1}}$ for any $x \in \mathbb{H}^{d+1}$ (one can show that the definition is independent of the choice of x). Let $\mathrm{Hull}(\Lambda(\Gamma))$ be the convex hull of $\Lambda(\Gamma)$ in $\mathbb{H}^{d+1}$. Then Γ is called *convex cocompact* if the convex core $\mathrm{Core}(M) := \Gamma\backslash\mathrm{Hull}(\Lambda(\Gamma)) \subset M$ is compact, and Γ is *Zariski dense* if Γ is not contained in the zero set of some nontrivial polynomial on $\mathrm{SO}(d+1,1)_0$. We identify the sphere $\mathbb{S}^d$ with $\mathbb{R}^d \cup \{\infty\}$. In the Poincaré upper half-space model, the limit set $\Lambda(\Gamma) \subset \mathbb{S}^d$ is a compact set of dimension $\delta_\Gamma \in (0, d)$ (see [Sarkar and Winter 2021, §2]), and we may assume that $\Lambda(\Gamma)$ is a compact subset of $\mathbb{R}^d$.

We recall the following nonconcentration property from [Sarkar and Winter 2021, Proposition 6.6].

Proposition 5.3. *Let $\Gamma \subset G$ be a convex cocompact subgroup such that Γ is Zariski dense in G. Then there exists $c_0 > 0$ such that, for any $x \in \Lambda(\Gamma) \cap \mathbb{R}^d$, $\varepsilon \in (0,1)$, and $w \in \mathbb{R}^d$ with $|w| = 1$, there exists $y \in \Lambda(\Gamma) \cap B(x,\varepsilon)$ such that*

$$|\langle y - x, w\rangle| > c_0\varepsilon. \tag{5-1}$$

As a corollary we have the following.

Corollary 5.4. *Let M be a convex cocompact hyperbolic $(d+1)$-dimensional manifold such that Γ is Zariski dense in G. Then, for any $\Phi \in C^3(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R})$ such that $\partial^2_{xy}\Phi(x,y)$ is nonvanishing, the pair $(\Lambda(\Gamma), \Lambda(\Gamma))$ is Φ-nonorthogonal with some constant $c_N > 0$ from scales 0 to 1.*

Proof. By the mean value theorem, for $x_1, x_2 \in B(x_0, r_X)$ and $y_1, y_2 \in B(y_0, r_Y)$,

$$|\Phi(x_0,y_0) - \Phi(x_1,y_0) - \Phi(x_0,y_1) + \Phi(x_1,y_1) - \langle \partial_{xy}\Phi(x_0,y_0)(x_1-x_0), y_1-y_0\rangle| \le \|\Phi\|_{C^3} r_X r_Y (r_X + r_Y).$$

Let $H = \ker(\partial^2_{xy}\Phi(x_0,y_0))$ and v be a unit normal vector to H (if $H = \{0\}$, then we choose v arbitrarily). By Proposition 5.3, there exists $x_1 \in \Lambda(\Gamma) \cap B(x_0, r_X)$ such that $|\langle x_1 - x_0, v\rangle| > c_0 r_X$. This would imply, for some $c_1 \in (0,1)$, that

$$|\partial^2_{xy}\Phi(x_0,y_0)(x_1 - x_0)| > c_1 c_0 r_X.$$

By Proposition 5.3 again, there exists $y_1 \in \Lambda(\Gamma) \cap B(y_0, r_Y)$ such that

$$|\langle \partial^2_{xy}\Phi(x_0,y_0)(x_1 - x_0), y_1 - y_0\rangle| > c_1 c_0^2 r_X r_Y.$$

[4]The fact that regularity and nonorthogonality only hold up to scale $2h$ causes us to incur a loss of a power of 2, but this is irrelevant.

Thus we may choose $r_X, r_Y \le \frac{1}{10}c_1c_0^2\|\Phi\|_{C^3}^{-1}$ such that

$$|\Phi(x_0, y_0) - \Phi(x_1, y_0) - \Phi(x_0, y_1) + \Phi(x_1, y_1)| > \tfrac{1}{2}c_1c_0^2 r_X r_Y,$$

i.e., nonorthogonality holds with

$$c_N = \frac{c_1^3 c_0^6}{200(1 + \|\Phi\|_{C^3})^2} > 0. \qquad \square$$

Proof of Theorem 1.6. As before, we may assume that $\Lambda(\Gamma)$ is contained in a compact region of $\mathbb{R}^d \subset \mathbb{S}^d$. On such a region, the stereographic projection $\varphi : \mathbb{S}^d \setminus \{\infty\} \to \mathbb{R}^d$ is bounded in C^3 with C^3-bounded inverse. Consider the measure

$$\mu_h(A) := h^{\delta_\Gamma - d}|\Lambda(\Gamma)_h \cap A|,$$

which is defined using the Euclidean metric but is comparable to the measure defined using the spherical metric due to the bounds on the stereographic projection. By Lemma 5.1, μ_h is δ_Γ-regular on scales $[2h, 1]$.

Let $\chi \in C_0^\infty(\mathbb{S}^d \times \mathbb{S}^d \setminus \{(x, x) : x \in \mathbb{S}^d\})$, and define $B_\chi(h) : L^2(\mathbb{S}^d) \to L^2(\mathbb{S}^d)$ by

$$B_\chi(h)u(x) = (2\pi h)^{-d/2}\int_{\mathbb{S}^d} |x - y|^{2i/h}\chi(x, y)u(y)\,\mathrm{d}y.$$

Then $1_{\Lambda(\Gamma)_h}B_\chi(h)1_{\Lambda(\Gamma)_h}$ can be rewritten as $(2\pi h)^{-d/2}h^{d-\delta_\Gamma}\mathcal{B}_h$, where $\mathcal{B}_h$ is the operator studied in Theorem 1.4 with $p(y')\,\mathrm{d}y' = \chi(\varphi^{-1}(y))\varphi_*(\mathrm{d}y)$, $\mu_X = \mu_Y = \mu_h$, and $\Phi(x, y) = 2\log|\varphi^{-1}(x) - \varphi^{-1}(y)|$. Combining Theorem 1.4 with the bounds on the stereographic projection and Corollary 5.4, we conclude the fractal uncertainty bound

$$\|1_{\Lambda(\Gamma)_h}B_\chi(h)1_{\Lambda(\Gamma)_h}\|_{L^2(\mathbb{S}^d)\to L^2(\mathbb{S}^d)} \le Ch^{d/2-\delta_\Gamma+\varepsilon_0}.$$

By a covering argument as in [Bourgain and Dyatlov 2018, Proposition 4.2], we have, for $\rho \in (0, 1)$,

$$\|1_{\Lambda(\Gamma)_{h^\rho}}B_\chi(h)1_{\Lambda(\Gamma)_{h^\rho}}\|_{L^2(\mathbb{S}^d)\to L^2(\mathbb{S}^d)} \le Ch^{d/2-\delta_\Gamma+\varepsilon_0-2(1-\rho)}.$$

Thus, $\Lambda(\Gamma)$ satisfies the fractal uncertainty principle with exponent $\beta = \frac{1}{2}d - \delta_\Gamma + \varepsilon_0$ in the sense of [Dyatlov and Zahl 2016, Definition 1.1]. Applying [Dyatlov and Zahl 2016, Theorem 3], we conclude that the Laplacian on M has only finitely many resonances in $\left\{\operatorname{Im}\lambda > \delta_\Gamma - \frac{1}{2}d - \varepsilon_0 + \varepsilon\right\}$ for any $\varepsilon > 0$, proving Theorem 1.6. $\square$

Appendix: The nonorthogonality constant of a classical Schottky group

In this appendix we demonstrate a simple way to estimate the nonorthogonality constant for classical Schottky groups Γ in $\mathrm{SO}(3, 1)_0 = \mathrm{PSL}(2, \mathbb{C})$, pointed out to us by Qiuyu Ren. The key idea is to use the fact that Möbius transformations are conformal maps and preserve circles in order to derive (5-1).

We illustrate this by considering Schottky groups of genus 2. Let $D_1,\ D_2,\ D_3,\ D_4$ be four disjoint closed disks in $\mathbb{CP}^1 = \partial_\infty\mathbb{H}^3$, and let $\gamma_1, \gamma_2 \in \mathrm{PSL}(2, \mathbb{C})$ such that

$$\gamma_1(\overline{D_3^c}) = D_1, \quad \gamma_2(\overline{D_4^c}) = D_2, \quad \gamma_3 = \gamma_1^{-1}, \quad \gamma_4 = \gamma_2^{-1}.$$

Let $\Gamma = \langle\gamma_1, \gamma_2\rangle$ be the free group generated by γ_1 and γ_2. Thus, Γ is a Schottky group of genus 2.

We identify $\mathbb{CP}^1$, $\mathbb{S}^2$, and $\mathbb{R}^2 \cup \{\infty\}$. We may assume that the D_i do not contain ∞ and hence are contained in $\mathbb{R}^2$. Given vectors $v, w \in \mathbb{R}^2$, let $\angle(v, w)$ denote the angle between v, w. The notion of a circle is not completely invariant under conformal transformations of $\mathbb{CP}^1$. We recall that a *generalized circle* is either a circle or a line; conformal transformations map generalized circles to generalized circles. We will choose the disks $D_1,\ D_2,\ D_3,\ D_4$ such that

$$\text{no generalized circle passes though all four disks.} \tag{A-1}$$

The circle taken here is not necessarily a great circle.

Let $\bar{a} \equiv a + 2 \bmod 4$ for $a \in \mathcal{A} = \{1, 2, 3, 4\}$, so that $\bar{1} = 3,\ \bar{2} = 4$. The limit set $\Lambda(\Gamma)$ is given by the Cantor-like procedure

$$\Lambda(\Gamma) = \bigcap_{n=1}^{\infty} \bigsqcup_{\boldsymbol{a} \in \mathcal{W}^n} D_{\boldsymbol{a}}, \quad \mathcal{W}^n = \{a_1 a_2 \cdots a_n \in \mathcal{A}^n : \bar{a}_i \neq a_{i+1}\},$$

where $D_{\boldsymbol{a}} = \gamma_{a_1}(\gamma_{a_2}(\cdots(\gamma_{a_{n-1}}(D_{a_n}))))$.

The nonorthogonality condition (1-3) follows from the nonconcentration property (5-1). Thus it suffices to find absolute constants $0 < c_1 < 1$ and $\kappa = \kappa(\Gamma) > 0$ such that, for each $x \in \Lambda(\Gamma),\ \epsilon > 0$, and unit vector $w \in \mathbb{R}^2$, there exists an element $y \in \Lambda(\Gamma) \cap B(x, \epsilon) \setminus B(x, c_1\epsilon)$ such that

$$|\cos \angle(x - y, w)| \geq \kappa.$$

Suppose $x \in D_{\boldsymbol{a}} = D_{\boldsymbol{a}_0 b}$ and $B(x, \epsilon)$ is roughly of the size of $D_{\boldsymbol{a}_0}$. Then there are two other disks in $D_{\boldsymbol{a}_0}$, which we call $D_{\boldsymbol{a}_0 c}$ and $D_{\boldsymbol{a}_0 d}$. By condition (A-1) and conformal invariance of the action of Γ, we know that, for any $y_c \in D_{\boldsymbol{a}_0 c} \cap \Lambda(\Gamma)$ and $y_d \in D_{\boldsymbol{a}_0 d} \cap \Lambda(\Gamma)$,

$$\text{the circle passing through } x,\ y_c,\ y_d \text{ lies inside } D_{\boldsymbol{a}_0}. \tag{A-2}$$

A Möbius transformation preserving the unit disk is a composition of a rotation and the map

$$z \mapsto \frac{a - z}{1 - \bar{a}z}.$$

A simple computation shows the angles of the triangle $\Delta(x, y_c, y_d)$ are uniformly lower bounded under conformal maps preserving $D_{\boldsymbol{a}_0}$ if we assume (A-2). This implies that

$$\theta < \angle(y_c - x, y_d - x) < \pi - \theta$$

for some constant θ depending on the initial angles between $\gamma_a(D_b),\ a \neq \bar{b}$. Thus, by the pigeonhole principle,

$$\max(|\cos \angle(y_c - x, w)|, |\cos \angle(y_d - x, w)|) \geq \cos\Big(\frac{\pi - \theta}{2}\Big).$$

If we assume, moreover,

$$\text{for any } b \neq \bar{a} \neq c, \text{ there exist } a' \neq a,\ b' \neq \bar{a}' \text{ such that}$$
$$\text{no circle passes through } \gamma_a(D_b),\ \gamma_a(D_c),\ \gamma_{a'}(D_{b'}), \text{ and } D_{\bar{a}} \tag{A-3}$$

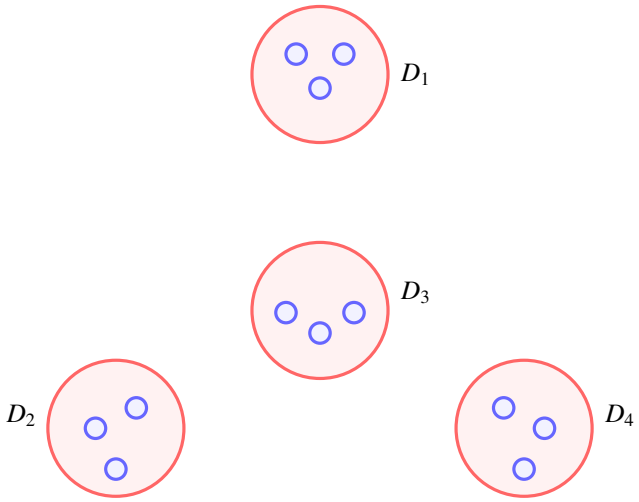

Figure 5. Iteration of disks under a Schottky group.

(which can be achieved if we choose the disks D_a to be small and with generic centers), then we can derive a lower bound on c_1 in a similar way. To be more precise, let $x \in D_{\boldsymbol{a}} = D_{\boldsymbol{a}_0 b} = S_{\boldsymbol{a}_1 ab}$ as before; then by assumption (A-3), there exists $a' \neq a$ and $b' \neq \bar{a}'$ such that

$$\text{the circle passing through } D_{\boldsymbol{a}_0 b},\ D_{\boldsymbol{a}_0 c},\ \text{and } D_{\boldsymbol{a}_1 a' b'} \text{ lies inside } D_{\boldsymbol{a}_1}. \tag{A-4}$$

In particular, for any $y_{a'b'} \in D_{\boldsymbol{a}_1 a' b'}$, the angles of the triangle $\Delta(x, y_c, y_{a'b'})$ are lower bounded. This in particular implies that the length of $\overline{xy_c}$ is comparable to the length of $\overline{y_c y_{a'b'}}$, which by the previous step is comparable with the size of $D_{\boldsymbol{a}_0}$. This allows us to compute a lower bound of c_1.

If one runs this procedure carefully, then it would be possible to compute an explicit nonorthogonality constant in terms of the angles between the disks $\gamma_a(D_b)$ in the initial step and the uniform constants in doing conformal transformations.

We do not bother to do the computation here, but we include Figure 5 to indicate how the procedure works. Conformal invariance ensures that the small blue disks always have an angle that lies in $[\theta, \pi - \theta]$.

While one needs to compute the above parameters κ and θ for any given Zariski-dense classical Schottky group Γ, we claim that this is always possible in principle, at least after passing to a finer scale. We say that a pair of words $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{W}^n,\ n \in \mathbb{N} \cup \{+\infty\}$, is *$\varepsilon$-separated* if their weighted Hamming distance satisfies

$$\sum_{i=1}^{n} \frac{1_{a_i \neq b_i}}{2^i} \geq \varepsilon.$$

Lemma A.1. *Let Γ be a classical Schottky group which is Zariski dense in* $\mathrm{PSL}(2, \mathbb{C})$. *For every $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that, for every $n \geq N$ and every triple of words $\boldsymbol{a}^n, \boldsymbol{b}^n, \boldsymbol{c}^n \in \mathcal{W}^n$ which are pairwise ε-separated, there exists $\boldsymbol{d}^n \in \mathcal{W}^n$ such that, for every circle X which meets all three disks $D_{\boldsymbol{a}^n}$, $D_{\boldsymbol{b}^n}$, $D_{\boldsymbol{c}^n}$, we have that X does not meet $D_{\boldsymbol{d}^n}$.*

Proof. We first prove an analogous result for the set of infinite words $\mathcal{W}^\infty$ and then reduce the finite case to the infinite case. To formulate it, let $x_{\boldsymbol{a}}$ be the unique point in $\lim_n D_{a_1\cdots a_n}$ (so $\boldsymbol{a} \mapsto x_{\boldsymbol{a}}$ is a homeomorphism $\mathcal{W}^\infty \to \Lambda(\Gamma)$, where $\mathcal{W}^\infty$ is given the product topology).

Let $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathcal{W}^\infty$ be distinct. Then there is a unique circle $X_{\boldsymbol{abc}} \subset \mathbb{CP}^1$ passing through $x_{\boldsymbol{a}},\ x_{\boldsymbol{b}},\ x_{\boldsymbol{c}}$. We claim that there exists $\boldsymbol{d} \in \mathcal{W}^\infty$ such that $x_{\boldsymbol{d}} \notin X_{\boldsymbol{abc}}$. Otherwise $\Lambda(\Gamma)$ is contained in a circle, which contradicts Proposition 5.3.

We now address the finite case. Suppose that the lemma fails on some $\boldsymbol{a}^n, \boldsymbol{b}^n, \boldsymbol{c}^n \in \mathcal{W}^n$ for each $n \in \mathbb{N}$ which are ε-separated, so, for every $\boldsymbol{d}^n \in \mathcal{W}^n$, there exists a circle $X(\boldsymbol{d}^n)$ which meets all disks $D_{\boldsymbol{a}^n},\ D_{\boldsymbol{b}^n},\ D_{\boldsymbol{c}^n},\ D_{\boldsymbol{d}^n}$. Let $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathcal{W}^\infty$ be the limits of $\boldsymbol{a}^n$, etc., and let $\boldsymbol{d} \in \mathcal{W}^\infty$ be given. Then $\boldsymbol{d} = \lim_n \boldsymbol{d}^n$ for some sequence $\boldsymbol{d}^n \in \mathcal{W}^n$, and we can define

$$X := \lim_n X(\boldsymbol{d}^n)$$

in Hausdorff distance. Then, $x_{\boldsymbol{a}}, x_{\boldsymbol{b}}, x_{\boldsymbol{c}}, x_{\boldsymbol{d}} \in X$, and $\boldsymbol{a},\ \boldsymbol{b},\ \boldsymbol{c}$ are ε-separated and hence distinct. Moreover, X is the limit of circles in $\mathbb{CP}^1$ whose radii are bounded from below (by ε-separation), so X is a circle, and hence $X = X_{\boldsymbol{abc}}$. This contradicts the infinite case. □

Assuming Lemma A.1, for $D_{\boldsymbol{a}} = D_{a_1\cdots a_{2n}}$, we can find $\boldsymbol{b}, \boldsymbol{c} \in \mathcal{W}^{2n}$ such that any circle passing through $D_{\boldsymbol{a}},\ D_{\boldsymbol{b}}$, and $D_{\boldsymbol{c}}$ lies in the disk D_{a_1}. This is because, given $D_{a_1\cdots a_{2n}}$ and $\overline{D^c_{a_1}}$, we have

$$\gamma_{\bar{a}_n}\cdots\gamma_{\bar{a}_2}\gamma_{\bar{a}_1}(D_{a_1\cdots a_{2n}}) = D_{a_{n+1}\cdots a_{2n}}, \quad \gamma_{\bar{a}_n}\cdots\gamma_{\bar{a}_2}\gamma_{\bar{a}_1}(\overline{D^c_{a_1}}) = D_{\bar{a}_n\cdots\bar{a}_2\bar{a}_1}.$$

By Lemma A.1, there exists $\boldsymbol{b}_0, \boldsymbol{c}_0 \in \mathcal{W}^n$ such that no circle passes through $D_{a_{n+1}\cdots a_{2n}},\ D_{\bar{a}_n\cdots\bar{a}_2\bar{a}_1},\ D_{\boldsymbol{b}_0}$, and $D_{\boldsymbol{c}_0}$. Applying $\gamma_{a_1}\cdots\gamma_{a_n}$, we conclude any circle passing through

$$D_{a_1\cdots a_{2n}}, \quad D_{a_1\cdots a_n\boldsymbol{b}_0}, \quad D_{a_1\cdots a_n\boldsymbol{c}_0}$$

lies inside D_{a_1} (there might be cancellations for the words $a_1\cdots a_n\boldsymbol{b}_0$ and $a_1\cdots a_n\boldsymbol{c}_0$ but one can always pass to a smaller disk). This allows us to compute the angle θ as before for general Zariski-dense classical Schottky groups.

Acknowledgments

The authors would like to thank Semyon Dyatlov for suggesting this problem and for helpful comments on earlier drafts. We would also like to thank Frédéric Naud for suggesting the references [Sarkar and Winter 2021; Stoyanov 2008; 2011], Pratyush Sarkar for suggesting the references [Guillopé and Zworski 1995; Mazzeo and Melrose 1987; Sarkar 2022], Terence Tao for helpful discussions and for suggesting the reference [Christ 1990], Qiuyu Ren for proposing the method we use in the Appendix, and Long Jin and Ruixiang Zhang for helpful discussions. We thank the anonymous referees for helpful suggestions to improve the paper.

Backus was supported by the National Science Foundation's Graduate Research Fellowship Program under Grant No. DGE-2040433, Leng was supported by the NSF's GRFP under Grant No. DGE-2034835, and Tao was partially supported by the NSF grant DMS-1952939 and Simons Targeted Grant award no. 896630.

References

[Bourgain and Dyatlov 2017] J. Bourgain and S. Dyatlov, "Fourier dimension and spectral gaps for hyperbolic surfaces", *Geom. Funct. Anal.* **27**:4 (2017), 744–771. MR Zbl

[Bourgain and Dyatlov 2018] J. Bourgain and S. Dyatlov, "Spectral gaps without the pressure condition", *Ann. of Math.* (2) **187**:3 (2018), 825–867. MR Zbl

[Bunke and Olbrich 1999] U. Bunke and M. Olbrich, "Group cohomology and the singularities of the Selberg zeta function associated to a Kleinian group", *Ann. of Math.* (2) **149**:2 (1999), 627–689. MR Zbl

[Chernov 1998] N. I. Chernov, "Markov approximations and decay of correlations for Anosov flows", *Ann. of Math.* (2) **147**:2 (1998), 269–324. MR Zbl

[Chow and Sarkar 2022] M. Chow and P. Sarkar, "Exponential mixing of frame flows for convex cocompact locally symmetric spaces", preprint, 2022. arXiv 2211.14737

[Christ 1990] M. Christ, "A $T(b)$ theorem with remarks on analytic capacity and the Cauchy integral", *Colloq. Math.* **60/61**:2 (1990), 601–628. MR Zbl

[Cladek and Tao 2021] L. Cladek and T. Tao, "Additive energy of regular measures in one and higher dimensions, and the fractal uncertainty principle", *Ars Inven. Anal.* (2021), art. id. 1. MR Zbl

[Cohen 2023] A. Cohen, "Fractal uncertainty in higher dimensions", preprint, 2023. arXiv 2305.05022

[Cohen 2025] A. Cohen, "Fractal uncertainty for discrete two-dimensional Cantor sets", *Anal. PDE* **18**:3 (2025), 743–772. MR Zbl

[Dolgopyat 1998] D. Dolgopyat, "On decay of correlations in Anosov flows", *Ann. of Math.* (2) **147**:2 (1998), 357–390. MR Zbl

[Dyatlov 2019] S. Dyatlov, "An introduction to fractal uncertainty principle", *J. Math. Phys.* **60**:8 (2019), art. id. 081505. MR Zbl

[Dyatlov and Jin 2017] S. Dyatlov and L. Jin, "Resonances for open quantum maps and a fractal uncertainty principle", *Comm. Math. Phys.* **354**:1 (2017), 269–316. MR Zbl

[Dyatlov and Jin 2018] S. Dyatlov and L. Jin, "Dolgopyat's method and the fractal uncertainty principle", *Anal. PDE* **11**:6 (2018), 1457–1485. MR Zbl

[Dyatlov and Zahl 2016] S. Dyatlov and J. Zahl, "Spectral gaps, additive energy, and a fractal uncertainty principle", *Geom. Funct. Anal.* **26**:4 (2016), 1011–1094. MR Zbl

[Guillarmou 2005] C. Guillarmou, "Meromorphic properties of the resolvent on asymptotically hyperbolic manifolds", *Duke Math. J.* **129**:1 (2005), 1–37. MR Zbl

[Guillopé and Zworski 1995] L. Guillopé and M. Zworski, "Polynomial bounds on the number of resonances for some complete spaces of constant negative curvature near infinity", *Asymptotic Anal.* **11**:1 (1995), 1–22. MR Zbl

[Han and Schlag 2020] R. Han and W. Schlag, "A higher-dimensional Bourgain–Dyatlov fractal uncertainty principle", *Anal. PDE* **13**:3 (2020), 813–863. MR Zbl

[Jin and Zhang 2020] L. Jin and R. Zhang, "Fractal uncertainty principle with explicit exponent", *Math. Ann.* **376**:3-4 (2020), 1031–1057. MR Zbl

[Khalil 2023] O. Khalil, "Exponential mixing via additive combinatorics", preprint, 2023. arXiv 2305.00527

[Khalil 2024] O. Khalil, "Polynomial fourier decay for Patterson–Sullivan measures", preprint, 2024. arXiv 2404.09424

[Li and Pan 2023] J. Li and W. Pan, "Exponential mixing of geodesic flows for geometrically finite hyperbolic manifolds with cusps", *Invent. Math.* **231**:3 (2023), 931–1021. MR Zbl

[Li et al. 2021] J. Li, F. Naud, and W. Pan, "Kleinian Schottky groups, Patterson–Sullivan measures, and Fourier decay", *Duke Math. J.* **170**:4 (2021), 775–825. MR Zbl

[Li et al. 2023] J. Li, W. Pan, and P. Sarkar, "Exponential mixing of frame flows for geometrically finite hyperbolic manifolds", preprint, 2023. arXiv 2302.03798

[Liverani 2004] C. Liverani, "On contact Anosov flows", *Ann. of Math.* (2) **159**:3 (2004), 1275–1312. MR Zbl

[Lugosi 2009] G. Lugosi, "Desigualtats de concentració", *Butl. Soc. Catalana Mat.* **24**:2 (2009), 97–136, 209. MR

[Mazzeo and Melrose 1987] R. R. Mazzeo and R. B. Melrose, "Meromorphic extension of the resolvent on complete spaces with asymptotically constant negative curvature", *J. Funct. Anal.* **75**:2 (1987), 260–310. MR Zbl

[Naud 2005] F. Naud, "Expanding maps on Cantor sets and analytic continuation of zeta functions", *Ann. Sci. École Norm. Sup.* (4) **38**:1 (2005), 116–153. MR Zbl

[Oh and Winter 2016] H. Oh and D. Winter, "Uniform exponential mixing and resonance free regions for convex cocompact congruence subgroups of $\mathrm{SL}_2(\mathbb{Z})$", *J. Amer. Math. Soc.* **29**:4 (2016), 1069–1115. MR Zbl

[Patterson 1976] S. J. Patterson, "The limit set of a Fuchsian group", *Acta Math.* **136**:3-4 (1976), 241–273. MR Zbl

[Patterson 1988] S. J. Patterson, "On a lattice-point problem in hyperbolic space and related questions in spectral theory", *Ark. Mat.* **26**:1 (1988), 167–172. MR Zbl

[Patterson and Perry 2001] S. J. Patterson and P. A. Perry, "The divisor of Selberg's zeta function for Kleinian groups", *Duke Math. J.* **106**:2 (2001), 321–390. MR Zbl

[Rössel 2008] J. Rössel, Image of Sierpinski's carpet, 2008, available at https://w.wiki/EGk$.

[Sarkar 2022] P. Sarkar, "Generalization of Selberg's 3/16 theorem for convex cocompact thin subgroups of $\mathrm{SO}(n, 1)$", *Adv. Math.* **409** (2022), art. id. 108610. MR Zbl

[Sarkar and Winter 2021] P. Sarkar and D. Winter, "Exponential mixing of frame flows for convex cocompact hyperbolic manifolds", *Compos. Math.* **157**:12 (2021), 2585–2634. MR Zbl

[Stoyanov 2008] L. Stoyanov, "Spectra of Ruelle transfer operators for contact flows", *Serdica Math. J.* **34**:1 (2008), 219–252. MR Zbl

[Stoyanov 2011] L. Stoyanov, "Spectra of Ruelle transfer operators for axiom A flows", *Nonlinearity* **24**:4 (2011), 1089–1120. MR Zbl

[Sullivan 1979] D. Sullivan, "The density at infinity of a discrete group of hyperbolic motions", *Inst. Hautes Études Sci. Publ. Math.* 50 (1979), 171–202. MR Zbl

[Thurston 1979] W. P. Thurston, "The geometry and topology of three-manifolds", lecture notes, Princeton University, 1979, available at https://url.msp.org/gt3m.

[Tsujii and Zhang 2023] M. Tsujii and Z. Zhang, "Smooth mixing Anosov flows in dimension three are exponentially mixing", *Ann. of Math.* (2) **197**:1 (2023), 65–158. MR Zbl

[Vacossin 2023] L. Vacossin, "Resolvent estimates in strips for obstacle scattering in 2D and local energy decay for the wave equation", *Pure Appl. Anal.* **5**:4 (2023), 1009–1039. MR Zbl

[Vasy 2013a] A. Vasy, "Microlocal analysis of asymptotically hyperbolic and Kerr-de Sitter spaces", *Invent. Math.* **194**:2 (2013), 381–513. MR Zbl

[Vasy 2013b] A. Vasy, "Microlocal analysis of asymptotically hyperbolic spaces and high-energy resolvent estimates", pp. 487–528 in *Inverse problems and applications*: *inside out*, *II*, edited by G. Uhlmann, Math. Sci. Res. Inst. Publ. **60**, Cambridge Univ. Press, 2013. MR Zbl

Received 14 Apr 2023. Revised 27 May 2024. Accepted 20 Jul 2024.

AIDAN BACKUS: aidan_backus@brown.edu
Department of Mathematics, Brown University, Providence, RI, United States

JAMES LENG: jamesleng@math.ucla.edu
Department of Mathematics, University of California, Los Angeles, Los Angeles, CA, United States

ZHONGKAI TAO: ztao@math.berkeley.edu
Department of Mathematics, University of California, Berkeley, Berkeley, CA, United States

ANALYSIS AND PDE
Vol. 18 (2025), No. 7, pp. 1805–1834
DOI: 10.2140/apde.2025.18.1805

msp

CONSEQUENCES OF THE RANDOM MATRIX SOLUTION TO THE PETERSON–THOM CONJECTURE

BEN HAYES, DAVID JEKEL AND SRIVATSAV KUNNAWALKAM ELAYAVALLI

We show various new structural properties of free group factors using the recent resolution (due independently to Belinschi and Capitaine, and Bordenave and Collins) of the Peterson–Thom conjecture. These results include the resolution to the coarseness conjecture due independently to the first author and Popa, a generalization of Ozawa and Popa's celebrated strong solidity result using vastly more general versions of the normalizer (and in an ultraproduct setting), a dichotomy result for intertwining of maximal amenable subalgebras of interpolated free group factors, as well as applications to ultraproduct embeddings of nonamenable subalgebras of interpolated free group factors.

1. Introduction

The structure of the group of von Neumann algebras associated to the countable free groups (also known as free group factors) has been a constant source of new results and new mysteries. Murray and von Neumann [1936] showed that the free group factors are full, i.e., they have no central sequences, and they used this structural property to distinguish the free group factors from the separable hyperfinite II_1-factor, thus giving the first example of two provably nonisomorphic separable II_1-factors. Their work left behind the now notorious open question of whether the free group factors themselves are isomorphic for different numbers of generators. Almost a century has passed in the development of II_1-factors, in which the quest to understand the structure of free group factors has been a recurring theme with several remarkable achievements.

One avenue of this research is the structure of subalgebras of free group factors. A foundational discovery of Popa [1983a] showed that every subalgebra that strictly contains the generator MASA (maximally abelian subalgebra) in a free group factor must be full and in particular nonamenable (amenability is equivalent to hyperfiniteness by fundamental work of Connes [1976]). This answered in the negative a question of Kadison at the 1967 Baton Rouge conference who asked if every self-adjoint operator in a II_1-factor is contained in a hyperfinite subfactor. The technique of asymptotic orthogonality developed by Popa to achieve the above result has been used successfully to establish this maximal amenability property for various natural subalgebras of the free group factors, such as the radial MASA [Cameron et al. 2010] (see also [Brothier and Wen 2016; Parekh et al. 2018]). Recently Boutonnet and Popa [2023] also constructed a continuum size family $(M_\alpha)_\alpha$ of interesting maximally amenable subalgebras in any free product of diffuse tracial von Neumann algebras (in particular for free group factors) with the property that M_α is not unitarily conjugate to M_β if $\alpha \neq \beta$.

MSC2020: 46L09, 46L53, 46L54.
Keywords: free group factors, 1-bounded entropy, anti-coarse space, coarseness conjecture.

Maximal amenability can be enhanced to an absorption phenomenon as follows. We say that a diffuse $P \leq M$ has the *absorbing amenability property* if, whenever $Q \leq M$ is amenable and $P \cap Q$ is diffuse, we have $Q \leq P$. By modifying Popa's asymptotic orthogonality property, it was shown in [Houdayer 2015; Wen 2016] respectively that the generator MASA and the radial MASA admit the absorbing amenability property. This work inspired many papers establishing the absorbing amenability property (and other absorption properties such as Gamma stability) in many examples; see [Brothier and Wen 2016; Hayes et al. 2021b; Parekh et al. 2018].

Given a finite von Neumann algebra M,[1] we say that M has *unique maximal amenable extensions* if, for every diffuse, amenable subalgebra $Q \leq M$, there is a *unique* maximal amenable $P \leq M$ with $Q \subseteq P$. Peterson and Thom [2011] conjectured that any diffuse, amenable subalgebra of a free group factor has unique maximal amenable extensions, which came to be known as the *Peterson–Thom conjecture*. This conjecture was motivated by both Peterson and Thom's analogous insights on groups with positive first L^2-Betti numbers and previous work of Ozawa and Popa [2010a], Peterson [2009], and Jung [2007]. One can apply Zorn's lemma to show that, for any von Neumann algebra M and any amenable $Q \leq M$, there is *some* maximal amenable $P \leq M$ with $Q \subseteq P$. The novelty of the Peterson–Thom conjecture is that such a P should be *unique*. The Peterson–Thom conjecture is then equivalent to the statement that every maximal amenable subalgebra of a free group factor has the absorbing amenability property.

In [Hayes 2022], the first author formulated a conjecture on random matrices, which he showed implies the Peterson–Thom conjecture. Several works in random matrices made progress towards this random matrix conjecture [Bandeira et al. 2023; Collins et al. 2022; Parraud 2023]. Recent breakthroughs of Belinschi and Capitaine [2022] and of Bordenave and Collins [2023] independently prove this random matrix conjecture, thus resolving the Peterson–Thom conjecture in the positive.

The reduction of the Peterson–Thom conjecture to a random matrix problem uses Voiculescu's microstates free entropy dimension theory (see [Voiculescu 1991; 1995; 1994; 1996]), namely the 1-bounded entropy defined implicitly in [Jung 2007, Theorem 3.2] and explicitly by the first author in [Hayes 2018]. We denote the 1-bounded entropy of an algebra M by $h(M)$ (see the Appendix for the precise definition, which we will not need in the main body of the paper). The 1-bounded entropy has several permanence properties which show that the collection of algebras Q with $h(Q : M) \leq 0$ is invariant under various operations such as taking subalgebras, taking the von Neumann algebra generated by its normalizer (or other weakenings of the normalizer) of an algebra, or taking the join of two algebras with diffuse intersection; see Section 2.2 for a list of such properties. Because of the permanence properties that the 1-bounded entropy enjoys, solving the Peterson–Thom conjecture via 1-bounded entropy proves several results beyond showing that free group factors have the absorbing amenability property. This paper will explain in detail several corollaries of the recent resolution of the Peterson–Thom conjecture. As shown in [Popa 2021], solving the Peterson–Thom property via 1-bounded entropy also resolves it for the interpolated free group factors $L(\mathbb{F}_t)$, independently defined by Dykema [1994] and Rădulescu [1994],

[1] It is not necessary for M to be finite. However, it could be argued that in the general setting one should require all subalgebras to be images of normal conditional expectations. We leave it to those more versed in Tomita–Takesaki theory to work out the appropriate definition here.

for $t > 1$. We give a separate proof of this in Section 3. In fact, because of our work in Section 3, all the main results in this paper apply to interpolated free group factors and not just free group factors.

Our first main result is the positive resolution of the *coarseness conjecture* independently due to the first author [Hayes 2018, Conjecture 1.12] and Popa [2021, Conjecture 5.2]. If M is a von Neumann algebra, an M-M bimodule $\mathcal{H}$ is a Hilbert space with normal left and right actions of M which commute. We use ${}_M\mathcal{H}_M$ to mean that $\mathcal{H}$ is an M-M bimodule. If $\mathcal{H}$ and $\mathcal{K}$ are M-M bimodules, we use ${}_M\mathcal{H}_M \preceq_M {}_M\mathcal{K}_M$ to mean that there is an M-bimodular isometry $\mathcal{H} \to \mathcal{K}$. If $\mathcal{H}_1 \subseteq \mathcal{H}_2$ are Hilbert spaces, we use $\mathcal{H}_2 \ominus \mathcal{H}_1$ for $\mathcal{H}_1^{\perp} \cap \mathcal{H}_2$.

Theorem 1.1. *Let $t > 1$. For any maximal amenable subalgebra $P \leq L(\mathbb{F}_t)$, we have*

$$ {}_P[L^2(L(\mathbb{F}_t)) \ominus L^2(P)]_P \preceq_P (L^2(P) \otimes L^2(P))^{\oplus\infty}{}_P. $$

In [Popa 2021], this property is referred to as *coarseness* of the inclusion $P \leq L(\mathbb{F}_t)$. As explained in the introduction of that paper, we may think of coarseness as the "most random" position a subalgebra can be in.

It is of interest to specialize Theorem 1.1 to the case where P is abelian. Suppose (M, τ) is a tracial von Neumann algebra and $A \leq M$ is a maximal abelian $*$-subalgebra. Write $A = L^\infty(X, \mu)$ for some compact Hausdorff space X and some Borel probability measure μ on X. The representation

$$ \pi : C(X) \otimes C(X) \to B(L^2(M) \ominus L^2(A)), $$

given by

$$ \pi(f \otimes g)\xi = f\xi g, $$

gives rise to a spectral measure E on $X \times X$ whose marginals are Radon–Nikodym equivalent to μ. We say that $\nu \in \mathrm{Prob}(X \times X)$ is a *left/right measure* of $A \leq M$ if it is Radon–Nikodym equivalent to E. One often abuses terminology and refers to *the* left/right measure to refer to any element of this equivalence class of measures. The measure class of this measure ν together with the multiplicity function $m : X \times X \to \mathbb{N} \cup \{0\} \cup \{\infty\}$ is usually called the measure-multiplicity invariant (see [Feldman and Moore 1977; Neshveyev and Størmer 2002]), which has also been investigated in [Dykema and Mukherjee 2013; Mukherjee 2013; Popa 2021]. The essential range of the multiplicity function is called the Pukánszky invariant; this was defined in [Pukánszky 1960] and further studied in [Dykema et al. 2006; Popa 2019; Rădulescu 1991; Robertson and Steger 2010; Sinclair and Smith 2005; White 2008]. By [Ge and Popa 1998, Theorem 4.1] and [Popa 2019, Corollary 3.8 (1)], we know that every MASA in an interpolated free group factor has unbounded multiplicity function.

Theorem 1.2. *Let $M = L(\mathbb{F}_t)$ for $t > 1$. Suppose that $A \leq M$ is abelian and a maximal amenable subalgebra of M. Write $A = L^\infty(X, \mu)$ for some compact metrizable space X and some Borel probability measure on X. Then the left/right measure of $A \leq M$ is absolutely continuous with respect to $\mu \otimes \mu$.*

Our work shows that, for MASAs in interpolated free group factors which are also maximal amenable, the measure given in the measure-multiplicity invariant has to be absolutely continuous with respect to the product measure $\mu \otimes \mu$. More concretely, if we use the fact that all standard probability spaces are isomorphic to reduce to the case where (X, μ) is $[0, 1]$ with Lebesgue measure, then the measure given

in the measure-multiplicity invariant has to be absolutely continuous with respect to Lebesgue measure on the unit square.

One of the landmark structural results about free group factors is the solidity property that the commutant of any diffuse subalgebra is amenable. Ozawa [2009] achieved this result first by introducing C*-algebraic boundary techniques. Popa [2007] then gave a different proof using his influential s-malleable deformations and spectral gap rigidity ideas. Peterson [2009] verified solidity for more examples using a conceptually new approach based on the theory of closable derivations. All of these results apply to algebras much more general than free group factors, for instance, von Neumann algebras of hyperbolic groups; see [Chifan and Sinclair 2013; Ding et al. 2023; Ozawa and Popa 2010b; Popa and Vaes 2014a; 2014b; Sinclair 2011].

Despite the early success of free entropy theory in establishing global structural properties of free group factors (for instance, absence of Cartan subalgebras [Voiculescu 1996] and primeness [Ge 1998]), solidity was still out of reach by free entropy methods. In this paper our first result is a proof of Ozawa's solidity theorem based on free entropy dimension techniques, which is completely different from previous arguments. In fact, for interpolated free group factors, we strengthen the celebrated strong solidity theorem of Ozawa and Popa [2010a, Corollary 1] using a vastly more general version of the normalizer. A first example is the 1-sided quasinormalizer, defined in [Izumi et al. 1998; Pimsner and Popa 1986; Popa 1999] (building off of ideas in [Popa 1983b]),

$$q^1\mathcal{N}_M(N) = \left\{x \in M : \text{there exists } x_1, \ldots, x_n \in M \text{ such that } xN \subseteq \sum_{j=1}^{n} Nx_j\right\}.$$

We also consider the wq-normalizer, defined in [Galatan and Popa 2017; Ioana et al. 2008; Popa 2006c; 2006b],

$$\mathcal{N}_M^{wq}(N) = \{u \in \mathcal{U}(M) : uNu^* \cap N \text{ is diffuse}\}$$

and its cousin the very weak quasinormalizer

$$\mathcal{N}_M^{vwq}(N) = \{u \in \mathcal{U}(M) : \text{there exists } v \in \mathcal{U}(M) \text{ such that } uNv \cap N \text{ is diffuse}\}.$$

Since $uNv \cap N$ is not an algebra, the phrase "$uNv \cap N$ is diffuse" should be interpreted as saying that there is a sequence of unitaries $v_n \in uNv \cap N$ which tend to zero weakly. We also consider the *weak intertwining space* $wI_M(Q, Q)$ due to Popa [2005; 2021] (we restate the definition in Definition 2.1 of this paper). As shown in [Hayes 2018, Proposition 3.2] and Section 2.1, all of these are contained in the anti-coarse space

$$\mathcal{H}_{\text{anti-c}}(N \leq M) = \bigcap_{T \in \text{Hom}_{N-N}(L^2(M), L^2(N) \otimes L^2(N))} \ker(T).$$

Here $\text{Hom}_{N-N}(L^2(M), L^2(N) \otimes L^2(N)$ is the space of bounded, linear, N-N bimodular maps

$$T : L^2(M) \to L^2(N) \otimes L^2(N).$$

Our next main result is a statement about the most general setting of the anti-coarse space; however, for the reader's sake we state it in the context of these four examples.

Theorem 1.3. *Let $t > 1$ and $Q \leq L(\mathbb{F}_t)$ be a diffuse, amenable subalgebra. Then $\mathrm{W}^*(\mathcal{H}_{\text{anti-c}}(Q \leq L(\mathbb{F}_t)))$ remains amenable. In particular, for any*

$$X \subseteq q^1\mathcal{N}_{L(\mathbb{F}_t)}(Q) \cup \mathcal{N}^{wq}_{L(\mathbb{F}_t)}(Q) \cup wI_{L(\mathbb{F}_t)}(Q, Q) \cup \mathcal{N}^{vwq}_{L(\mathbb{F}_t)}(Q),$$

we have that $\mathrm{W}^*(X)$ is amenable.

We refer the reader to Section 4.2 for the precise definition of $\mathrm{W}^*(Y)$ for $Y \subseteq L^2(M)$. The case $X = \mathcal{N}_{L(\mathbb{F}_t)}(Q)$ in the above theorem recovers the strong solidity theorem of Ozawa and Popa [2010a, Corollary 1] for the special case of interpolated free group factors. Theorem 1.3 is already new for $t \in \mathbb{N}$ for and when

$$X \in \{q^1\mathcal{N}_{L(\mathbb{F}_t)}(Q),\ \mathcal{N}^{wq}_{L(\mathbb{F}_t)}(Q),\ wI_{L(\mathbb{F}_t)}(Q, Q),\ \mathcal{N}^{vwq}_{L(\mathbb{F}_t)}(Q)\}.$$

We may, in fact, deduce a further generalization of strong solidity for interpolated free group factors in an ultraproduct framework.

Theorem 1.4. *Let $t \in (1, +\infty)$, and let ω be a free ultrafilter on $\mathbb{N}$. Suppose that $Q \leq L(\mathbb{F}_t)^\omega$ is a diffuse, amenable subalgebra. Suppose we are given Neumann subalgebras Q_α defined for ordinals α satisfying*

- $Q_0 = Q$,
- *if α is a successor ordinal, then $Q_\alpha = \mathrm{W}^*(X_\alpha, Q_{\alpha-1})$, where*

$$X_\alpha \subseteq \mathcal{H}_{\text{anti-c}}(Q_{\alpha-1} \leq L(\mathbb{F}_t))$$

  (*for example if*

$$X_\alpha \subseteq q^1\mathcal{N}_{L(\mathbb{F}_t)^\omega}(Q_{\alpha-1}) \cup \mathcal{N}^{wq}_{L(\mathbb{F}_t)^\omega}(Q_{\alpha-1}) \cup wI_{L(\mathbb{F}_t)^\omega}(Q_{\alpha-1}, Q_{\alpha-1}) \cup \mathcal{N}^{vwq}_{M}(Q_{\alpha-1})),$$

- *if α is a limit ordinal, then $Q_\alpha = \overline{\bigcup_{\beta<\alpha} Q_\beta}^{\mathrm{SOT}}$.*

Then, for any ordinal α we have that $Q_\alpha \cap L(\mathbb{F}_t)$ is amenable. In particular, $L(\mathbb{F}_t)$ has the following Gamma stability property: if $Q \leq L(\mathbb{F}_t)^\omega$ and if $Q' \cap L(\mathbb{F}_t)^\omega$ is diffuse, then $Q \cap L(\mathbb{F}_t)$ is amenable.

This recovers the previous Gamma stability results from [Houdayer 2015] (recovering also instances of the results of [Ding and Kunnawalkam Elayavalli 2024; Ding et al. 2023]).

The case of the weak intertwining space itself leads to a dichotomy in terms of Popa's deformation/rigidity theory for maximal amenable subalgebras of free group factors. We recall the fundamental notion of intertwining introduced in [Popa 2006a; 2006c]. If M is a finite von Neumann algebra and $P, Q \leq M$, we say that *a corner of Q intertwines into P inside of M* and write $Q \prec P$ if there are nonzero projections $f \in Q$ and $e \in P$, a unital $*$-homomorphism $\Theta : fQf \to ePe$, and a nonzero partial isometry $v \in M$ such that

- $xv = v\Theta(x)$ for all $x \in fQf$,
- $vv^* \in (fQf)' \cap fMf$,
- $v^*v \in \Theta(fqf)' \cap eMe$.

This can be thought of intuitively as "Q can be unitarily embedded into P after cutting by a projection".

Theorem 1.5. *Fix $t > 1$, and let Q and P be maximal amenable subalgebras of $L(\mathbb{F}_t)$. Then exactly one of the following occurs*:

(1) *either there are nonzero projections $e \in Q$, $f \in P$ and a unitary $u \in L(\mathbb{F}_t)$ such that $u^*(ePe)u = fQf$,*

(2) *or, for any diffuse $Q_0 \leq Q$, we have that $Q_0 \not\prec P$.*

In particular, if Q, P are hyperfinite subfactors of $L(\mathbb{F}_t)$ that are maximal amenable subalgebras in $L(\mathbb{F}_t)$, then either they are unitarily conjugate or no corner of any diffuse subalgebra of one can be intertwined into the other inside of $L(\mathbb{F}_t)$.

So, given any pair P, Q of maximal amenable subalgebras of $L(\mathbb{F}_t)$, they either have unitarily conjugate corners or no diffuse subalgebra of Q can be "essentially conjugated" into P. The reader should compare this with [Popa 2006a, Theorem A.1], where a similar result is shown for MASAs in *any* tracial von Neumann algebra.

We close with an application to embeddings into matrix ultraproducts.

Corollary 1.6. *Let $t > 1$, and let $N \leq L(\mathbb{F}_t)$ be a nonamenable subfactor. Then there is a free ultrafilter ω and an embedding $\iota : N \to \prod_{k\to\omega} M_k(\mathbb{C})$, with $\iota(N)' \cap \prod_{k\to\omega} M_k(\mathbb{C}) = \mathbb{C}1$.*

The corollary is proved as follows. The results of [Hayes 2022] and [Belinschi and Capitaine 2022; Bordenave and Collins 2023] imply that every nonamenable $N \leq L(\mathbb{F}_t)$ satisfies $h(N : L(\mathbb{F}_t)) > 0$, and hence $h(N) > 0$; see Corollary 3.1 for details. Work of the second author shows that if $h(N) > 0$, then there exists some embedding of N into a matrix ultraproduct which has trivial relative commutant [Jekel 2023, Corollary 1.3, see the statement in corrigendum].

A comment on proofs. We give a few remarks on how 1-bounded entropy is used in the paper. The first is that, for any tracial von Neumann algebra, the 1-bounded entropy leads to a natural class of subalgebras called *Pinsker algebras*. To say that $P \leq M$ is Pinsker means that P is a maximal subalgebra such that $h(P : M) \leq 0$, where $h(Q : N)$ for $Q \leq N$ is the 1-bounded entropy of Q in the presence of N defined in [Hayes 2018]. Intuitively this means two things: firstly that P has "very few" embeddings into ultraproducts of matrices which extend to M, and that P is maximal with respect to inclusion among algebras which have "very few" embeddings into ultraproducts of matrices which extend to M. We refer to Section 2.3 for the precise definition of a Pinsker algebra. Given a diffuse subalgebra $Q \leq M$ with $h(Q : M) \leq 0$, general properties of 1-bounded entropy show that there is a unique Pinsker algebra $P \leq M$ with $Q \subseteq P$. In the context of the Peterson–Thom conjecture, the unique maximal amenable extension of a diffuse subalgebra can be identified exactly as the Pinsker algebra containing this subalgebra. Thus the 1-bounded entropy cannot only be used to solve the Peterson–Thom conjecture but also naturally identify the maximal amenable extensions of a given amenable subalgebra.

The second remark is that, as was previously mentioned, the 1-bounded entropy enjoys several permanence properties, which we will list in Section 2.2. For the proofs of all the theorems mentioned so far, we will only use these permanence properties as well as the fact that the results of [Belinschi and Capitaine 2022; Bordenave and Collins 2023; Hayes 2022] show that the Pinsker algebras in interpolated free group factors are precisely the maximal amenable subalgebras. In particular, we never have to work

with the precise definition of 1-bounded entropy, just these permanence properties. This tells us that the axiomatic treatment of 1-bounded entropy via these general properties can be used to deduce many interesting and new results on the structure of free group factors.

Organization of the paper. In Section 2.1, we recall the anti-coarse space and expand the results in [Hayes 2018, Proposition 3.2], showing that it contains several other weakenings of the normalizer. In Section 2.2, we list the permanence properties of 1-bounded entropy we will use in this paper. For the proofs of all of our applications of [Hayes 2022] and [Belinschi and Capitaine 2022; Bordenave and Collins 2023], we will only use these permanence properties and not the precise definition of 1-bounded entropy, so these properties give an axiomatic approach to most of the proofs in this paper. In Section 2.3, we recall the notion of Pinsker algebras for 1-bounded entropy introduced in [Hayes et al. 2021b] and recast the results of [Belinschi and Capitaine 2022; Bordenave and Collins 2023] in these terms. In Section 3, we will explain how these results apply not just to maximal amenable subalgebras of free group factors, but also to those of interpolated free group factors. Section 4.1 contains the proof of the coarseness conjecture, as well as applications to maximal abelian subalgebras of free group factors. In Section 4.2, we give several generalizations of Ozawa and Popa's celebrated strong solidity theorem. In Section 4.3, we give a dichotomy for intertwining between maximal amenable subalgebras of interpolated free group factors (and more generally for Pinsker algebras in any tracial von Neumann algebra). In the Appendix, we give the definition of 1-bounded entropy and prove that it is independent of the choice of generators (a fact proved first implicitly in [Jung 2007, Theorem 3.2] and later explicitly in [Hayes 2018, Theorem A.9]). Here we give a significant conceptual and technical simplification of the proof using the noncommutative functional calculus due to the second author.

2. Preliminaries

2.1. ***The anti-coarse space.*** For an inclusion $N \leq M$ of tracial von Neumann algebras, we let

$$\mathcal{H}_{\text{anti-c}}(N \leq M) = \bigcap_{T \in \mathrm{Hom}_{N-N}(L^2(M), L^2(N) \otimes L^2(N))} \ker(T),$$

where $\mathrm{Hom}_{N-N}(L^2(M), L^2(N) \otimes L^2(N)$ is the space of bounded, linear, N-N bimodular maps

$$T : L^2(M) \to L^2(N) \otimes L^2(N).$$

It is shown in [Hayes 2018, Proposition 3.2] that this contains the following generalizations of the normalizer of $N \leq M$:

$$q^1\mathcal{N}_M(N) = \left\{ x \in M : \text{there exists } x_1, \ldots, x_n \in M \text{ such that } xN \subseteq \sum_{j=1}^{n} Nx_j \right\},$$

$$\mathcal{N}_M^{wq}(N) = \{u \in \mathcal{U}(M) : uNu^* \cap N \text{ is diffuse}\}.$$

We recall the following definition, due to Popa [2005; 2021] (see also [Galatan and Popa 2017; Ioana et al. 2008; Popa 2006b] for related concepts).

Definition 2.1 [Popa 2021, Definition 2.6.1]. Let (M, τ) be a tracial von Neumann algebra. For $Q, P \leq M$ diffuse, we define the *intertwining space from Q to P inside M*, denoted by $I_M(Q, P)$, to be the set of $\xi \in L^2(M)$ such that

$$\overline{\operatorname{span}\{a\xi b : a \in Q, b \in P\}}^{\|\cdot\|_2}$$

has finite dimension as a right P-module. We define the *weak intertwining space from Q to P inside M* by

$$wI_M(Q, P) = \bigcup_{Q_0 \leq Q \text{ diffuse}} I_M(Q_0, P).$$

The following is a well-known result due to [Popa 2021, Proposition 2.6.3], but we include the proof for completeness.

Proposition 2.2. *Let (M, τ) be a tracial von Neumann algebra. For $Q \leq M$ diffuse, we have*

$$wI_M(Q, Q) \subseteq \mathcal{H}_{\text{anti-c}}(Q \leq M).$$

Proof. Fix $Q_0 \leq Q$ diffuse. It suffices to show that

$$(I_M(Q_0, Q))^\perp \supseteq \mathcal{H}_{\text{anti-c}}(Q \leq M)^\perp.$$

It is a folklore result that $\mathcal{H}_{\text{anti-c}}(Q \leq M)^\perp$ can be embedded into an infinite direct sum of $L^2(Q) \otimes L^2(Q)$ as a Q-Q bimodule (see, e.g., [Hayes 2018, Proposition 3.3] for a complete proof). Since Q_0 is diffuse, we can find a sequence $u_n \in \mathcal{U}(Q_0)$ which tend to zero weakly. We leave it as an exercise to check that, for all $\xi, \eta \in L^2(Q) \otimes L^2(Q)$, we have

$$\lim_{n\to\infty} \sup_{y \in Q : \|y\| \leq 1} |\langle u_n \xi y, \eta\rangle| = 0 = \lim_{n\to\infty} \sup_{y \in Q : \|y\| \leq 1} |\langle y \xi u_n, \eta\rangle|$$

(to prove this one can, for instance, check it on simple tensors and then conclude using linearity and density). Since $\mathcal{H}_{\text{anti-c}}(Q \leq M)^\perp$ embeds into an infinite direct sum of $L^2(Q) \otimes L^2(Q)$ as a Q-Q bimodule, it follows that, for all $\xi, \eta \in \mathcal{H}_{\text{anti-c}}(Q \leq M)^\perp$, we have

$$\lim_{n\to\infty} \sup_{y \in Q : \|y\| \leq 1} |\langle u_n \xi y, \eta\rangle| = 0 = \lim_{n\to\infty} \sup_{y \in Q : \|y\| \leq 1} |\langle y \xi u_n, \eta\rangle|.$$

Since

$$\sup_{y \in Q : \|y\| \leq 1} |\langle u_n \xi y, \eta\rangle| = \|\mathbb{E}_Q(\eta^* u_n \xi)\|_1,$$

we have

$$0 = \lim_{n\to\infty} \|\mathbb{E}_Q(\eta^* u_n \xi)\|_1.$$

So [Popa 2019, Theorem 1.3.2] implies that $\mathcal{H}_{\text{anti-c}}(Q \leq M)^\perp \subseteq I_M(Q_0, Q)^\perp$, as desired. □

To further illustrate the generality of the anti-coarse space, we show that it contains the following very weak normalizer of $(N \subseteq M)$:

$$\mathcal{N}_M^{vwq}(N) = \{u \in \mathcal{U}(M) : \text{there exists } v \in \mathcal{U}(M) \text{ such that } uNv \cap N \text{ is diffuse}\}.$$

Here by "diffuse" we mean that $uNv \cap N$ contains a sequence $(u_n)_{n\in\mathbb{N}}$ of unitaries with $u_n \to 0$ in WOT as $n \to \infty$.

Proposition 2.3. *Let (M, τ) be a tracial von Neumann algebra. For $N \leq M$ diffuse, we have*

$$\mathcal{N}_M^{vwq}(N) \subseteq \mathcal{H}_{\text{anti-c}}(N \leq M).$$

Proof. The argument proceeds exactly as in [Hayes 2018, proof of Proposition 3.2], but we include it for the reader's convenience. Let $u \in \mathcal{N}_M^{wqv}(N)$. By definition, it suffices to show that, for every

$$T \in \text{Hom}_{N-N}(L^2(M), L^2(N) \otimes L^2(N)),$$

we have $T(u) = 0$. Let $v \in \mathcal{U}(M)$ and $u_n \in \mathcal{U}(N) \cap uNv$ be such that $u_n \to_{n\to\infty}^{\text{WOT}} 0$. Write

$$w_n = u^* u_n v^* \in \mathcal{U}(N)$$

and observe that $w_n \to 0$ in WOT. Then, using that T is N-N bimodular and that $u_n, w_n \in \mathcal{U}(N)$,

$$\begin{aligned}\|T(u)\|_2^2 = \|T(u)w_n\|_2^2 &= \langle T(u)w_n, T(u)w_n\rangle = \langle T(u)w_n, T(uw_n)\rangle \\ &= \langle T(u)w_n, T(u_n v^*)\rangle = \langle u_n^* T(u)w_n, T(v^*)\rangle.\end{aligned}$$

Since $u_n^* \to_{n\to\infty}^{\text{WOT}} 0$ and $w_n \to_{n\to\infty}^{\text{WOT}} 0$, it follows as in Proposition 2.2 that $\langle u_n^* \xi w_n, \eta\rangle \to_{n\to\infty} 0$ for all $\xi, \eta \in L^2(N) \otimes L^2(N)$. Taking limits as $n \to \infty$ above thus shows that $T(u) = 0$. □

2.2. 1-*bounded entropy*. One of the main ideas going into the proof of the Peterson–Thom conjecture is the 1-bounded entropy h of a tracial von Neumann algebra, a numerical invariant which appeared implicitly in [Jung 2007] and was made explicit by the first author in [Hayes 2018]. We will need the more general notion of 1-bounded entropy in the presence, which is defined for inclusions $N \leq M$ of tracial von Neumann algebras and is denoted by $h(N : M)$. For detailed expositions and recent work on this topic, see [Charlesworth et al. 2023; Chifan et al. 2023; Hayes et al. 2021a; 2021b; 2024; Jekel 2023; Kunnawalkam Elayavalli 2023].

For the applications in this paper, we will not need to use the definition of h directly, only the properties listed below. We include the precise definition of h in the Appendix, along with a streamlined proof that the definition is independent of the choice of the generating sets for the von Neumann algebras. The name "1-bounded entropy" derives from the following result, connecting the 1-bounded entropy to the strong 1-boundedness of Jung [2007] .

Theorem 2.4 [Hayes 2018, Proposition A.16]. *A tracial von Neumann algebra M is strongly* 1*-bounded* (in the sense of [Jung 2007]) *if and only if M is finitely generated and diffuse and satisfies $h(M) < \infty$.*

For this reason, we say that any tracial von Neumann algebra (M, τ) (even if M is not finitely generated or diffuse) is *strongly* 1*-bounded* if $h(M) < +\infty$.

Let (M, τ) be a tracial von Neumann algebra. The 1-bounded entropy in the presence enjoys the following properties:

(P1) $h(M) = h(M : M)$ for every tracial von Neumann algebra (M, τ).

(P2) Suppose $N \leq M$. Then $h(N : M) \geq 0$ if M embeds into an ultrapower of $\mathcal{R}$, and $h(N : M) = -\infty$ if M does not embed into an ultrapower of $\mathcal{R}$. (We leave this as an exercise.)

(P3) $h(N_1 : M_1) \leq h(N_2 : M_2)$ if $N_1 \leq N_2 \leq M_2 \leq M_1$. (We leave this as an exercise.)

(P4) $h(N : M) \leq 0$ if $N \leq M$ and N is hyperfinite. (We leave this as an exercise.)

(P5) $h(M) = \infty$ if M is diffuse, and $M = \mathrm{W}^*(x_1, \ldots, x_n)$, where $x_j \in M_{\mathrm{sa}}$ for all $1 \leq j \leq n$ and $\delta_0(x_1, \ldots, x_n) > 1$. For example, this applies to $M = L(\mathbb{F}_n)$ for $n > 1$ because of [Voiculescu 1994; 1996] together with Theorem 2.4 and [Jung 2007, Corollary 3.5].

(P6) $h(N_1 \vee N_2 : M) \leq h(N_1 : M) + h(N_2 : M)$ if $N_1, N_2 \leq M$ and $N_1 \cap N_2$ is diffuse. (See [Hayes 2018, Lemma A.12].)

(P7) Suppose that $(N_\alpha)_\alpha$ is an increasing chain of diffuse von Neumann subalgebras of a von Neumann algebra M. Then

$$h\left(\bigvee_\alpha N_\alpha : M\right) = \sup_\alpha h(N_\alpha : M).$$

(See [Hayes 2018, Lemma A.10].)

(P8) $h(N : M) = h(N : M^\omega)$ if ω is a free ultrafilter on an infinite set. (See [Hayes 2018, Proposition 4.5].)

(P9) $h(\mathrm{W}^*(\mathcal{H}_{\text{anti-c}}(N \leq M)) : M) = h(N : M)$ if $N \leq M$ is diffuse. (See [Hayes 2018, Theorem 3.8]. See also Section 4.2 for the definition of $\mathrm{W}^*(Y)$ for $Y \subseteq L^2(M)$.)

(P10) Let I be a countable set and $M = \bigoplus_{i \in I} M_i$ with M_i diffuse for all i. Suppose that τ is a faithful trace on M and that λ_i is the trace of the identity on M_i. Endow M_i with the trace $\tau_i = \tau|_{M_i}/\lambda_i$. Fix $N_i \leq M_i$ for all $i \in I$. Then

$$h(N : M, \tau) \leq \sum_i \lambda_i^2 h(N_i : M_i, \tau_i).$$

(See the proof of [Hayes 2018, Proposition A.13 (i)].)

(P11) If $z \in \mathrm{Proj}(Z(M))$, $N \leq M$, and $h(N : M) \leq 0$, then $h(Nz : Mz) \leq 0$. (See [Hayes et al. 2021a, Lemma 4.2].)

(P12) $h(pNp : pMp) = (1/\tau(p)^2)h(N : M)$ if $N \leq M$ is diffuse, p is a nonzero projection in N, and M is a factor. (This follows from modifying the proofs of [Hayes 2018, Proposition A.13 (ii)] and [Hayes et al. 2021a, Proposition 4.6].)

2.3. *Pinsker algebras.*

Definition 2.5. Let (M, τ) be a tracial von Neumann algebra. We say that $P \leq M$ is *Pinsker* if $h(P : M) \leq 0$ and, for any $P \leq Q \leq M$ with $P \neq Q$, we have $h(Q : M) > 0$.

By properties (P6) and (P7), if $Q \leq M$ is diffuse and $h(Q : M) \leq 0$, then there is a unique Pinsker algebra $P \leq M$ with $Q \subseteq P$. E.g.,

$$P = \bigvee_{N \leq M, N \supseteq Q, h(N:M) \leq 0} N.$$

We call P the *Pinsker algebra of* $Q \subseteq M$. By [Hayes 2022] and the recent breakthrough result in [Belinschi and Capitaine 2022] and [Bordenave and Collins 2023], we have the following classification of Pinsker subalgebras of free group factors.

Theorem 2.6. *Fix $r \in \mathbb{N} \cup \{\infty\}$. Then*

(i) *$Q \leq L(\mathbb{F}_r)$ is amenable if and only if $h(Q : L(\mathbb{F}_r)) = 0$,*

(ii) *$P \leq L(\mathbb{F}_r)$ is Pinsker if and only if it is maximal amenable.*

As remarked in [Hayes et al. 2021b], we may think of 1-bounded entropy as analogous to the Kolmogorov–Sinaĭ entropy in the context of probability measure-preserving actions of groups. Entropy for probability measure-preserving actions of groups was first developed in the case of $\mathbb{Z}$ by [Kolmogorov 1958; Sinaĭ 1959], for amenable groups in [Kieffer 1975; Ornstein and Weiss 1987], and then for sofic groups in [Bowen 2010; Kerr and Li 2011]. See also [Seward 2019] for a potential approach to entropy for arbitrary acting groups, called Rokhlin entropy. A probability measure-preserving action $G \curvearrowright (X, \mu)$ with G sofic is said to have *complete positive entropy* if every nontrivial quotient probability measure-preserving quotient action has positive entropy. Dually, this is equivalent to saying that if B is a G-invariant von Neumann subalgebra B of $L^\infty(X, \mu)$ with $B \neq \mathbb{C}1$, then the action of G on B has positive entropy.

Motivated by the sofic entropy case, one could naively define a tracial von Neumann algebra (M, τ) to have completely positive 1-bounded entropy if any nontrivial subalgebra has positive 1-bounded entropy. However, this will never be satisfied, as any hyperfinite subalgebra will have vanishing 1-bounded entropy. Thus any tracial von Neumann algebra will have many subalgebras with vanishing 1-bounded entropy (these subalgebras can be chosen to be diffuse if M is). Instead, we should think of a result saying that the only subalgebras with vanishing 1-bounded entropy are ones that can be quickly deduced to vanishing have 1-bounded entropy from the properties (P1)–(P12) listed above (e.g., hyperfinite algebras, property Gamma algebras, nonprime algebras, algebras with a Cartan, etc.) as a complete positive entropy result. We may thus think of Theorem 2.6 as a complete positive entropy result for 1-bounded entropy. Since free group factors may be thought of as the free probability analogue of Bernoulli shifts (e.g., because $L(\mathbb{F}_\infty)$ is the crossed product algebra associated to a free Bernoulli shift), Theorem 2.6 should be compared with previous results establishing complete positive entropy of Bernoulli shifts (see [Kerr 2014; Rudolph and Weiss 2000]).

As discussed in [de Santiago et al. 2021, Section 5], Pinsker algebras of measure-preserving dynamical systems are analogous to the maximal rigid subalgebras of s-malleable deformations in that work. This allows for an exchange of ideas and methods between deformation/rigidity theory, free probability theory, and ergodic theory. This will be exploited in Section 4.3, where we adapt arguments in the aforementioned work to show that Pinsker algebras do not have any weak intertwiners between them unless they have corners which are unitarily conjugate.

3. Pinsker algebras of interpolated free group factors

As mentioned before, the combined results of [Hayes 2022] and [Belinschi and Capitaine 2022; Bordenave and Collins 2023] prove that, for $r \in \mathbb{N}$, we have that $Q \leq L(\mathbb{F}_r)$ is amenable if and only if $h(Q : L(\mathbb{F}_r)) = 0$. The main goal of this section is to explain how this automatically generalizes to interpolated free group factors.

Corollary 3.1. *Fix $t > 1$. Then $Q \leq L(\mathbb{F}_t)$ is amenable if and only if $h(Q : L(\mathbb{F}_t)) = 0$.*

By rephrasing the above corollary we obtain the following.

Corollary 3.2. *Fix $t > 1$. Let $P \leq L(\mathbb{F}_t)$. Then $P \leq L(\mathbb{F}_t)$ is Pinsker if and only if P is maximal amenable.*

In order to obtain this result, we will study the relationship between Pinsker algebras and compression. First, we generalize property (P12) to the case where M is not a factor.

Lemma 3.3. *Let (M, τ) be a tracial von Neumann algebra and $P \leq M$. If $h(P : M) \leq 0$, then, for every projection $p \in P$, we have that $h(pPp : pMp) \leq 0$.*

Proof. By decomposing the center of M into atomic and diffuse pieces, we can find a central projection $z_0 \in M$ (potentially zero), a countable set I (potentially empty), and central projections $(z_i)_{i \in I}$ such that

- $1 = z_0 + \sum_i z_i$,
- in the case $z_0 \neq 0$, we have that Mz_0 has diffuse center,
- Mz_i is a factor for all $i \in I$.

For $i \in \{0\} \cup I$, let $P_i = (pPp)z_i$ (even though z_i may not be in P, we still have that P_i is a von Neumann subalgebra of M as z_i is central). Set $\hat{P} = \overline{\sum_i P_i}^{\mathrm{WOT}}$. Then, by (P3) and (P10),

$$\begin{aligned} h(pPp : pMp) \leq h(\hat{P} : pMp) &\leq \tau(pz_0)^2 h(P_0 : (pMp)z_0) + \sum_i \tau(pz_i)^2 h(P_i : (pMp)z_i) \\ &\leq \tau(pz_0)^2 h((pMp)z_0) + \sum_i \tau(pz_i)^2 h(P_i : (pMp)z_i). \end{aligned} \tag{1}$$

So it suffices to show each term on the right-hand side of this inequality is nonpositive.

We first show that $\tau(pz_0)^2 h((pMp)z_0) \leq 0$. If $pz_0 = 0$, the claim is true. Otherwise, since M has diffuse center and $Z(pMp) = pZ(M)p$, we have that $(pMp)z_0$ has diffuse center. Thus $(pMp)z_0 = \mathrm{W}^*(\mathcal{N}_{(pMp)z_0}(pZ(M)pz_0))$, and so the combination of properties (P4) and (P9) implies $h((pMp)z_0) \leq 0$.

Now consider $h(P_i : (pMp)z_i)$ for $i \in I$. By property (P11), we know that

$$h(Pz_i : Mz_i) \leq 0$$

for all $i \in I$. Thus, by property (P12), we have that

$$h(P_i : (pMp)z_i) = \frac{1}{\tau(pz_i)^2} h(Pz_i : Mz_i) \leq 0.$$

Thus we have shown that all terms on the right-hand side of (1) are nonpositive, and this completes the proof. □

We now show that being a Pinsker algebra is preserved under taking corners and amplifications.

Proposition 3.4. *Let (M, τ) be a tracial von Neumann algebra, and suppose that $P \leq M$ is Pinsker. Then*

(i) *we have $Z(M) \subseteq P$,*

(ii) *for any nonzero projection $p \in P$, we have that pPp is a Pinsker subalgebra of pMp,*

(iii) *for any $n \in \mathbb{N}$, we have that $M_n(P)$ is a Pinsker subalgebra of $M_n(M)$.*

Proof. (i) Note that $Z(M) \vee P \subseteq \mathrm{W}^*(\mathcal{N}_M(P))$, and thus, by (P9),

$$h(Z(M) \vee P : M) \leq h(P : M) = 0,$$

and so P being Pinsker forces $Z(M) \vee P \subseteq P$. That is, $Z(M) \subseteq P$.

(ii) Let z be the central support of $p \in M$. Then, by (i), we know that z is under the central support of p in P. So there exists a collection $\{v_i\}_{i \in I}$ of nonzero partial isometries in P such that $v_i^* v_i \leq p$ and $z = \sum_i v_i v_i^*$. Set $p_i = v_i^* v_i$. We may, and will, assume that there is some i_0 such that $v_{i_0} = p$. By Lemma 3.3, $h(pPp : pMp) \leq 0$, and so there exists a Pinsker subalgebra Q of pMp containing pPp. Set

$$\hat{Q} = \overline{Q + \sum_{i \in I \setminus \{i_0\}} v_i v_i^* P v_i v_i^*}^{\mathrm{WOT}}.$$

Thus, by (P10),

$$\begin{aligned} h(Q : M) &\leq h\left(\hat{Q} : \overline{pMp + \sum_{i \in I \setminus \{i_0\}} v_i v_i^* M v_i v_i^*}^{\mathrm{WOT}}\right) \\ &\leq \tau(p)^2 h(Q : pMp) + \sum_{i \in I \setminus \{i_0\}} \tau(v_i v_i^*)^2 h(v_i v_i^* P v_i v_i^* : v_i v_i^* M v_i v_i^*) \\ &\leq \sum_{i \in I \setminus \{i_0\}} \tau(v_i v_i^*)^2 h(v_i v_i^* P v_i v_i^* : v_i v_i^* M v_i v_i^*). \end{aligned}$$

For all $i \in I$, we have that $x \mapsto v_i x v_i^*$ gives a trace-preserving isomorphism from $p_i M p_i$ to $v_i v_i^* M v_i v_i^*$, which takes $p_i P p_i$ to $v_i v_i^* P v_i v_i^*$. Hence, for all $i \in I$,

$$h(v_i v_i^* P v_i v_i^* : v_i v_i^* M v_i v_i^*) = h(p_i P p_i : p_i M p_i) \leq 0$$

by Lemma 3.3. Altogether we have shown that $h(\hat{Q} : M) \leq 0$. Note that

$$\hat{Q} \cap P \supseteq \overline{pPp + \sum_{i \in I \setminus \{i_0\}} v_i v_i^* P v_i v_i^*}^{\mathrm{WOT}},$$

which is diffuse. Hence, by property (P6), we have that $h(\hat{Q} \vee P : M) \leq 0$, and, by maximality, we have that $\hat{Q} \vee P \subseteq P$. It follows that $\hat{Q} \subseteq P$. So

$$Q = p\hat{Q}p \subseteq pPp.$$

(iii) Consider $M \leq M_n(M)$ by identifying it with $M \otimes 1 \leq M \otimes M_n(\mathbb{C}) \cong M_n(M)$. Under this identification, $\mathcal{N}_{M_n(M)}(P) \supseteq \mathcal{U}(M_n(\mathbb{C})) \cup \mathcal{U}(P)$, so $\mathrm{W}^*(\mathcal{N}_{M_n(M)}(P)) \supseteq M_n(P)$. Thus, by properties (P9) and (P3),

$$h(M_n(P) : M_n(M)) \leq h(P : M_n(M)) \leq h(P : M) \leq 0.$$

So we can let Q be the Pinsker algebra of $M_n(M)$ containing $M_n(P)$. Let e_{ij} be the standard matrix units of $M_n(\mathbb{C})$ viewed as elements of $M_n(M)$. Then, by (ii), we have that $e_{11} Q e_{11} = P$. But then, for all $x \in Q$, we have that

$$x = \sum_{i,j} e_{ii} x e_{jj} = \sum_{i,j} e_{i1} e_{11} x e_{11} e_{1j} \in M_n(P).$$

So $Q \leq M_n(P)$. □

Proof of Corollary 3.1. The forward implication is (P4) of the main properties of 1-bounded entropy we listed above. For the reverse implication, suppose for the sake of contradiction that Q is not amenable. Then by Connes–Haagerup characterization of amenability (see Lemma 2.2 in [Haagerup 1985]), there is a projection $p \in Z(Q)$ and $u_1, \ldots, u_r \in \mathcal{U}(Qp)$ such that

$$\left\| \frac{1}{r} \sum_{j=1}^{r} u_j \otimes \bar{u}_j \right\| < 1,$$

where $\bar{u}_j = (u_j^*)^{\mathrm{op}}$ and the norm is computed in $Qp \otimes_{\min} Qp$. Let $P \leq pL(\mathbb{F}_t)p$ be the Pinsker algebra of $L(\mathbb{F}_t)$ which contains Q. By fundamental results of Dykema and Rădulescu (see [Dykema 1994; Rădulescu 1994]), we may choose $s > 0$ such that $(pL(\mathbb{F}_t)p)^s \cong L(\mathbb{F}_2)$. Fix an integer $n > s$, and let q be a projection in $M_n(P)$ such that $\mathrm{Tr} \otimes \tau(q) = s$. Observe that

$$\left\| \frac{1}{r} \sum_{j=1}^{r} (1_n \otimes u_j) \otimes \overline{1_n \otimes u_j} \right\| < 1,$$

where 1_n is the identity of $M_n(\mathbb{C})$. Hence, it follows that $M_n(P)$ also has no nonzero amenable direct summands. We leave it as an exercise to show that this implies that $M_n(P)$ has no nonzero amenable corners. By Proposition 3.4, we know that

$$qM_n(P)q \leq qM_n(pL(\mathbb{F}_t)p)q \cong L(\mathbb{F}_2)$$

is Pinsker. It follows from Theorem 2.6 that $qM_n(P)q$ is amenable. This contradicts our previous observation that $M_n(P)$ has no nonzero amenable corners. □

4. Main results

4.1. ***Orthocomplement bimodule structure for maximal amenable subalgebras.*** We start with the following consequence of Theorem 2.6 on the structure of the orthocomplement bimodule for any maximal amenable $P \leq L(\mathbb{F}_t)$. Note that this verifies the coarseness conjecture, independently due to the first author [Hayes 2018, Conjecture 1.12] and Popa [2021, Conjecture 5.2].

Corollary 4.1. *Let $M = L(\mathbb{F}_t)$ for some $t > 1$. For any maximal amenable $P \leq L(\mathbb{F}_t)$, we have that*

$$_P(L^2(M) \ominus L^2(P))_P \leq [L^2(P) \otimes L^2(P)]^{\oplus \infty}.$$

Proof. As noted in Proposition 2.2, we have that $\mathcal{H}_{\text{anti-c}}(P \leq M)^{\perp}$ embeds into $[L^2(P) \otimes L^2(P)]^{\oplus \infty}$ as a P-P bimodule. Since P is Pinsker by Theorem 2.6, property (P9) implies that

$$\mathcal{H}_{\text{anti-c}}(P \leq M) = L^2(P).$$

Thus,

$$L^2(M) \ominus L^2(P) = \mathcal{H}_{\text{anti-c}}(P \leq M)^{\perp}$$

embeds into $[L^2(P) \otimes L^2(P)]^{\oplus \infty}$. □

Suppose (M, τ) is a tracial von Neumann algebra and $A \leq M$ is a maximal abelian $*$-subalgebra. Write $A = L^\infty(X, \mu)$ for some compact Hausdorff space X and some Borel probability measure μ on X. Let $\pi : C(X) \otimes C(X) \to B(L^2(M) \ominus L^2(A))$ be as in the definition of the left/right measure given in the introduction. Note that if ν is a left/right measure and if $\phi : C(X) \otimes C(X) \to L^\infty(X \times X, \nu)$ is the map sending an element of $C(X) \otimes C(X) \cong C(X \times X)$ to its $L^\infty(\nu)$-equivalence class, then there is a unique normal $*$-isomorphism $\rho : L^\infty(X \times X, \nu) \to \overline{\pi(C(X) \otimes C(X))}^{\mathrm{SOT}}$ such that $\pi = \rho \circ \phi$.

Corollary 4.2. *Let $M = L(\mathbb{F}_t)$ for some $t > 1$. Suppose that $A \leq M$ is abelian and a maximal amenable subalgebra of M. Write $A = L^\infty(X, \mu)$ for some compact metrizable space X and some Borel probability measure μ on X. Then the left/right measure of $A \leq M$ is absolutely continuous with respect to $\mu \otimes \mu$.*

Proof. Let $E : X \times X \to \mathrm{Proj}(L^2(M) \ominus L^2(A))$ be the spectral measure corresponding to the representation π defined as in the paragraph before Corollary 4.2. For a bounded, Borel map $\phi : X \times X \to \mathbb{C}$, we let

$$\tilde{\pi}(\phi) = \int \phi \, dE.$$

By Corollary 4.1, we know that $L^2(M) \ominus L^2(A)$ embeds into an infinite direct sum of $L^2(A) \otimes L^2(A)$ as an A-A bimodule. Thus, for any vector $\xi \in L^2(M) \ominus L^2(A)$, we may find a sequence $(k_n)_n \in L^2(X \times X)$ such that $\sum_n \int \|k_n\|_2^2 < +\infty$ and

$$\langle \pi(\phi)\xi, \xi \rangle = \sum_n \langle \phi k_n, k_n \rangle = \sum_n \iint \phi(x, y)|k_n(x, y)|^2 \, d\mu(x) \, d\mu(y) \quad \text{for all } \phi \in C(X \times X).$$

Set $K = \sum_n |k_n|^2$. Then, for every bounded, Borel $\phi : X \times X \to \mathbb{C}$, we have

$$\langle \tilde{\pi}(\phi)\xi, \xi \rangle = \iint \phi(x, y)K(x, y) \, d\mu(x) \, d\mu(y).$$

In particular, if $B \subseteq \mathbb{C}$ is Borel and $(\mu \otimes \mu)(B) = 0$, then

$$\|\tilde{\pi}(1_B)\xi\|_2^2 = \langle \tilde{\pi}(1_B)\xi, \xi \rangle = 0,$$

the first equality holding as $\tilde{\pi}(1_B)$ is a projection. Since this holds for every $\xi \in L^2(M) \ominus L^2(A)$, we see that $E(B) = \tilde{\pi}(1_B) = 0$. So we have shown that E is absolutely continuous with respect to $\mu \otimes \mu$. □

4.2. ***Generalizations of strong solidity.*** Throughout this section, we will be interested in properties of $\mathrm{W}^*(\mathcal{H}_{\text{anti-c}}(Q \leq M))$ for an inclusion $Q \leq M$ of tracial von Neumann algebras. Since $\mathcal{H}_{\text{anti-c}}(Q \leq M)$ is a subset of $L^2(M)$ and not of M, we need to explain what we mean by $\mathrm{W}^*(\mathcal{H}_{\text{anti-c}}(Q \leq M))$. Every $\xi \in L^2(M)$ may be identified with the densely defined, closed, unbounded operator L_ξ on $L^2(M)$; this L_ξ is the closure of the operator L_ξ^o which has $\mathrm{dom}(L_\xi^o) = M$ and is defined by $L_\xi^o(x) = \xi x$ for all $x \in M$. For $\xi \in L^2(M)$, we let $L_\xi = V_\xi |L_\xi|$ be its polar decomposition. For $X \subseteq L^2(M)$, we then define

$$\mathrm{W}^*(X) = \mathrm{W}^*\big(\{V_\xi : \xi \in X\} \cup \{\phi(|L_\xi|) : \xi \in X, \phi : [0, \infty) \to \mathbb{C} \text{ is bounded and Borel}\}\big).$$

Throughout this section, given a tracial von Neumann algebra (M, τ), we view $M \leq M^\omega$ by identifying it with the image of the constant sequences.

Definition 4.3. Let (M, τ) be a tracial von Neumann algebra. For a free ultrafilter $\omega \in \beta\mathbb{N} \setminus \mathbb{N}$, we say that M is *ω-strongly solid* if $\mathrm{W}^*(N_{M^\omega}(Q)) \cap M$ is amenable for all diffuse, amenable $Q \leq M^\omega$. We say that M is *ultrasolid* if it is ω-strongly solid for every free ultrafilter ω. We say that M is *spectrally solid* if, for any diffuse, amenable $Q \leq M$, we have that $\mathrm{W}^*(\mathcal{H}_{\text{anti-c}}(Q \leq M))$ is amenable. Given a free ultrafilter $\omega \in \beta\mathbb{N} \setminus \mathbb{N}$, we say that M is *spectrally ω-solid* if, for any diffuse, amenable $Q \leq M^\omega$, we have that $\mathrm{W}^*(\mathcal{H}_{\text{anti-c}}(Q \leq M^\omega)) \cap M$ is amenable. We say that M is *spectrally ultrasolid* if it is *spectrally ω-solid* for every free ultrafilter ω.

Corollary 4.4. *Fix $t > 1$.*

(i) *$L(\mathbb{F}_t)$ is spectrally ultrasolid.*

(ii) *If $Q \leq L(\mathbb{F}_t)$, $\omega \in \beta\mathbb{N} \setminus \mathbb{N}$ is a free ultrafilter and $Q' \cap L(\mathbb{F}_t)^\omega$ is diffuse, then Q is amenable.*

Proof. For notational simplicity, set $M = L(\mathbb{F}_t)$.

(i) Fix $\omega \in \beta\mathbb{N} \setminus \mathbb{N}$. Let $Q \leq M^\omega$ be diffuse and amenable. Then, by properties (P8), (P3), (P9), and (P4),

$$\begin{aligned} h(\mathrm{W}^*(\mathcal{H}_{\text{anti-c}}(Q \leq M^\omega)) \cap M : M) &= h(\mathrm{W}^*(\mathcal{H}_{\text{anti-c}}(Q \leq M^\omega)) \cap M : M^\omega) \\ &\leq h(\mathrm{W}^*(\mathcal{H}_{\text{anti-c}}(Q \leq M^\omega)) : M^\omega) = h(Q : M^\omega) \leq 0. \end{aligned}$$

So $h(\mathrm{W}^*(\mathcal{H}_{\text{anti-c}}(Q \leq M^\omega)) \cap M : M) \leq 0$, which implies (Corollary 3.2) that $\mathrm{W}^*(\mathcal{H}_{\text{anti-c}}(Q \leq M^\omega)) \cap M$ is amenable.

(ii) Fix $A \leq Q' \cap M^\omega$ diffuse and abelian. Then $Q \leq W^*(\mathcal{N}_{M^\omega}(A)) \leq \mathrm{W}^*(\mathcal{H}_{\text{anti-c}}(A \leq M^\omega))$, and the result this follows from (i). □

In particular, Corollary 4.4 and Proposition 2.2 imply that if $P \leq L(\mathbb{F}_t)$ is a maximal amenable subalgebra, then $wI_{L(\mathbb{F}_t)}(P, P) \subseteq L^2(P)$, and so P is *strongly malnormal* in the sense of [Popa 2021].

We can take several iterations of this procedure in the ultraproduct setting and it will still have amenable intersection with the diagonal copy of $L(\mathbb{F}_t)$.

Corollary 4.5. *Fix $t > 1$ and a free ultrafilter $\omega \in \beta\mathbb{N} \setminus \mathbb{N}$. Suppose that $Q \leq L(\mathbb{F}_t)^\omega$ is diffuse and amenable. Suppose we are given von Neumann subalgebras Q_α defined for ordinals α which satisfy the following properties*:

- $Q_0 = Q$.
- *If α is a successor ordinal, then $Q_{\alpha-1} \leq Q_\alpha \leq \mathrm{W}^*(\mathcal{H}_{\text{anti-c}}(Q_{\alpha-1} \leq L(\mathbb{F}_t)^\omega))$.*
- *If α is a limit ordinal, then $Q_\alpha = \overline{\bigcup_{\beta<\alpha} Q_\beta}^{\mathrm{SOT}}$.*

Then, for any ordinal α, we have that $Q_\alpha \cap L(\mathbb{F}_t)$ is amenable.

Proof. One applies properties (P8), (P3), (P9), (P7), and transfinite induction to see that

$$h(Q_\alpha : L(\mathbb{F}_t)^\omega) = 0$$

for any α. It then follows by property (P8) that

$$h(Q_\alpha \cap L(\mathbb{F}_t) : L(\mathbb{F}_t) = h(Q_\alpha \cap L(\mathbb{F}_t) : L(\mathbb{F}_t)^\omega) = h(Q_\alpha \cap L(\mathbb{F}_t) : L(\mathbb{F}_t)^\omega) \leq h(Q_\alpha : L(\mathbb{F}_t)^\omega) = 0.$$

The corollary now follows from Theorem 2.6. □

Corollary 4.6. *Fix $t > 1$. Then $L(\mathbb{F}_t)$ has the following Gamma stability property. Fix a free ultrafilter $\omega \in \beta\mathbb{N} \setminus \mathbb{N}$. If $Q \leq L(\mathbb{F}_t)^\omega$ and $Q' \cap L(\mathbb{F}_t)^\omega$ is diffuse, then $Q \cap L(\mathbb{F}_t)$ is amenable.*

Proof. Fix $A \leq Q' \cap L(\mathbb{F}_t)^\omega$ diffuse and abelian. Note that

$$Q \leq \mathrm{W}^*(\mathcal{N}_{L(\mathbb{F}_t)^\omega}(A)) \leq W^*(\mathcal{H}_{\text{anti-c}}(A \leq L(\mathbb{F}_t)^\omega)).$$

Applying Corollary 4.4 with $Q_0 = A$ and $Q_\alpha = Q \vee A$ for all $\alpha \geq 1$, we see that $Q \cap L(\mathbb{F}_t) \leq (A \vee Q) \cap L(\mathbb{F}_t)$ is amenable. □

4.3. ***Intertwining properties for Pinsker algebras.*** In this section, we explore how Pinsker algebras behave with respect to intertwining properties in the sense of [Popa 2006a; 2006c].

Theorem 4.7. *Let (M, τ) be a tracial von Neumann algebra, and let $P, Q \leq M$ be Pinsker. Then exactly one of the following occurs*:

- *either there are nonzero projections $e \in P$, $f \in Q$ and a unitary $u \in \mathcal{U}(M)$ such that $u(ePe)u^* = fQf$,*
- *or $wI_M(Q, P) = \{0\}$.*

Proof. Suppose that $wI_M(Q, P) \neq \{0\}$, so there is a diffuse $Q_0 \leq Q$ with $Q_0 \prec P$. This means there are nonzero projections $f_0 \in Q_0$, $e_0 \in P$, a unital $*$-homomorphism $\Theta : f_0 Q_0 f_0 \to e_0 P e_0$, and a nonzero partial isometry $v \in M$ such that

- $xv = v\Theta(x)$ for all $x \in f_0 Q_0 f_0$,
- $vv^* \in (f_0 Q_0 f_0)' \cap f_0 M f_0$,
- $v^*v \in \Theta(f_0 Q_0 f_0)' \cap e_0 M e_0$.

By property (P9), we have

$$h\big(\mathrm{W}^*(\mathcal{N}_{f_0 M f_0}(f_0 Q_0 f_0)) : f_0 M f_0\big) \leq h(f_0 Q f_0 : f_0 M f_0) \leq 0,$$

and since $f_0 Q f_0 \leq f_0 M f_0$ is Pinsker by Proposition 3.4, we know that $\mathcal{N}_{f_0 M f_0}(f_0 Q_0 f_0) \subseteq \mathcal{U}(f_0 Q f_0)$. Similarly, $\mathcal{N}_{e_0 M e_0}(e_0 P e_0) = \mathcal{U}(e_0 P e_0)$. It then follows as in [de Santiago et al. 2021, Theorem 6.8] that $f = vv^* \in Q$ and $e = v^*v \in P$.

Since $f \in Q$, we have that $v^*(fQf)v$ is a subalgebra of eMe. Moreover, conjugation by v implements an isomorphism between the inclusion $fQf \leq fMf$ and the inclusion $v^*(fQf)v \leq eMe$, which implies that $h(v^*(fQf)v : eMe) \leq 0$. Moreover,

$$v^*(fQf)v \cap ePe \supseteq e\Theta(f_0 Q f_0).$$

Since $\Theta(f_0 Q f_0)$ is the image of a diffuse subalgebra under a normal $*$-homomorphism, it follows that it is diffuse. Since ePe is Pinsker by Proposition 3.4(ii) and $h(v^*(fQf)v : eMe) \leq 0$, this forces $v^*(fQf)v = ePe$ by property (P6). Since M is finite, there is a unitary $u \in \mathcal{U}(M)$ such that $fu = v$. Then $u^*(fQf)u = ePe$, as desired. □

In the case where P and Q are factors, the first option in the dichotomy can be strengthened to saying that P and Q are unitarily conjugate.

Corollary 4.8. *Let (M, τ) be a tracial von Neumann algebra. Let P and Q be Pinsker subalgebras of M such that P and Q are factors. Then, either P and Q are unitarily conjugate, or $wI_M(P, Q) = 0$.*

This follows from the general fact that if M is a tracial von Neumann algebra and P and $Q \leq M$ are factors with unitarily conjugate corners, then they are unitarily conjugate. This is a folklore result and we include the proof here for completeness.

Proposition 4.9. *If Q, P are subalgebras of a tracial von Neumann algebra (M, τ) with unitarily conjugate corners and if P, Q are factors, then P, Q are unitarily conjugate.*

Proof. Choose nonzero projections $p \in P,\ q \in Q$ and a unitary partial isometry $v \in M$, with $v^*v = p$, $vv^* = q$, and

$$vPv^* = qQq.$$

Since P is a factor, we may shrink p, q if necessary to assume that $\tau(p) = 1/n$ for some integer n. Choose projections $p_1, \ldots, p_n$ in P which are pairwise orthogonal with $\tau(p_j) = 1/n$ for all j and with $p_1 = p$. Choose analogous projections $q_1, \ldots, q_n$ in Q with $q_1 = q$. Since P, Q are factors for $2 \leq j \leq n$, we may choose partial isometries a_j, b_j in P, Q such that $a_j^*a_j = p,\ a_ja_j^* = p_j,\ b_j^*b_j = q,\ b_jb_j^* = q_j$, and set $a_1 = p,\ b_1 = q$. Finally, let

$$u = \sum_i b_i v a_i^*.$$

Then u is a unitary, and if $x \in P$, then $u(a_ia_i^*xa_ja_j^*)u^* \in Q$ for all $1 \leq i, j \leq n$. Using that any $x \in P$ is equal to $\sum_{i,j} a_ia_i^*xa_ja_j^*$, we see that $uPu^* \subseteq Q$. By a symmetric argument, $u^*Qu \subseteq P$. □

The combination of the above results allows us to deduce Theorem 1.5 as follows.

Proof of Theorem 1.5. The fact that either (1) or (2) of Theorem 1.5 holds follows from Theorems 4.7 and 2.6. The "in particular" part follows from Corollary 4.8 and Theorem 2.6. □

Example 4.10. Dykema [1993] implies that $L(\mathbb{F}_2) \cong L^\infty[0, 1] * \mathcal{R}$. By Popa [1983a], $L^\infty[0, 1]$ and $\mathcal{R}$ are maximal amenable subalgebras in $L(\mathbb{F}_2)$. Hence, they are Pinsker subalgebras by Theorem 2.6. (Alternatively, [Hayes et al. 2021b] shows directly that they are Pinsker subalgebras.) Therefore, given any automorphism ϕ of $L(\mathbb{F}_2)$, by Theorem 4.7, $L^\infty[0, 1]$ and $\phi(\mathcal{R})$ in $L(\mathbb{F}_2)$ either have zero intertwiners or have unitarily conjugate corners. They do not have isomorphic corners; therefore $wI(L^\infty[0, 1], \phi(\mathcal{R})) = 0$.

Since the free product of any two amenable separable tracial von Neumann algebras is $L(\mathbb{F}_2)$, we can generalize this example quite a bit. To handle these more general examples, we want a strengthening of Theorem 4.7 along the lines of Corollary 4.8 that does not assume that P and Q are factors. First, we use a maximality argument to make the projection e in Theorem 4.7 as large as possible.

Theorem 4.11. *Let (M, τ) be a tracial von Neumann algebra, and let $P, Q \leq M$ be Pinsker subalgebras. There exist projections $e \in P$ and $f \in Q$ and a partial isometry $v \in M$ from e to f such that*

(1) $v(ePe)v^* = fQf$,

(2) *if $e \neq 1$ and u is a unitary such that $v = ue = fu$, then*

$$wI_{(1-f)M(1-f)}((1-f)Q(1-f), u(1-e)P(1-e)u^*) = 0.$$

Proof. Step 1: We show that there exists a choice of e and f satisfying (1) that is maximal, in the sense that no strictly larger projections satisfy (1). To this end, we will apply Zorn's lemma to the set of triples (e, f, v), where $e \in P$ and $f \in Q$ are projections and v is a partial isometry from e to f such that $v(ePe)v^* = fQf$. The partial order on this set will be given by $(e, f, v) \leq (e', f', v')$ if $e \leq e',\ f \leq f'$, and $fv'e = v$. Note that $fv'e = v$ implies that $e = v^*v = e(v')^*fv'e$, which in turn implies

$$|v'e - v|^2 = e + e(v')^*v'e - 2\,\mathrm{Re}(e(v')^*v) = e + e(v')^*v'e - 2\,\mathrm{Re}(e(v')^*fv'e) = e(v')^*v'e - e \leq 0.$$

So $v'e = v$ and similarly $fv' = v$. One checks readily that the above order is a partial order. So it remains to show that every increasing chain $\{(e_\alpha, f_\alpha, v_\alpha)\}_{\alpha\in I}$ has an upper bound. Let $e = \sup_\alpha e_\alpha$ and $f = \sup_\alpha f_\alpha$. Note that, for $\alpha \leq \beta$, we have

$$v_\beta - v_\alpha = v_\beta e_\beta - v_\alpha = v_\beta(e_\beta - e_\alpha),$$

using our previous observation that $\alpha \leq \beta$ implies $v_\alpha = v_\beta e_\alpha$. Since e_α converges to e, it follows that $(e_\alpha)_{\alpha\in I}$ is Cauchy in $L^2(M)$; this in turn implies that $(v_\alpha)_{\alpha\in I}$ is Cauchy in $L^2(M)$ and hence converges to some $v \in L^2(M)$. This v is necessarily also a partial isometry, and $v_\alpha = f_\alpha v e_\alpha$. Moreover, we have $v(ePe)v^* \subseteq fQf$ because, for each $x \in P$, we have $v(e_\alpha x e_\alpha)v^* = v_\alpha(e_\alpha x e_\alpha)v_\alpha^* \in f_\alpha Q f_\alpha \subseteq fQf$. So taking the limit over α, we get $v(exe)v^* \in fQf$. The same reasoning shows that $v^*fQfv \subseteq ePe$, and hence $v(ePe)v^* = fQf$. Hence, by Zorn's lemma, there exists a maximal choice of $e,\ f$, and v.

Step 2: Now we will apply Theorem 4.7 to show that the maximal $e,\ f$, and v satisfy (2). Let u be a unitary such that $v = ue = fu$. Suppose for contradiction that

$$wI_{(1-f)M(1-f)}((1-f)Q(1-f), u(1-e)P(1-e)u^*) \neq \{0\}.$$

By Proposition 3.4 (ii), $(1-f)Q(1-f)$ is Pinsker in $(1-f)M(1-f)$. Note uPu^* is Pinsker in M, and hence $(1-f)uPu^*(1-f) = u(1-e)P(1-e)u^*$ is Pinsker in $(1-f)M(1-f)$. By Theorem 4.7,

$$wI_{(1-f)M(1-f)}((1-f)Q(1-f), u(1-e)P(1-e)u^*) \neq \{0\}$$

implies that there exist projections $e_0 \in u(1-e)P(1-e)u^*$ and $f_0 \in (1-f)Q(1-f)$ and a partial isometry v_0 from e_0 to f_0 that conjugates $e_0u(1-e)P(1-e)u^*e_0$ to $f_0(1-f)Q(1-f)f_0$. Write $e' = u^*e_0u$, so that e' is a projection in P with $e' \leq 1-e$. Let $f' = f_0$, which is a projection in Q with $f' \leq 1-f$. Let $v' = v_0u$, which is a partial isometry in M sending e' to f'. Then

$$v'e'Pe'(v')^* = v_0e_0u(1-e)P(1-e)u^*e_0v_0^* = f_0(1-f)Q(1-f)f_0 = f'Qf'.$$

By Proposition 3.4 (ii), $(f+f')Q(f+f')$ is Pinsker in $(f+f')M(f+f')$, and $(e+e')P(e+e')$ is Pinsker in $(e+e')M(e+e')$, so that

$$(v+v')(e+e')P(e+e')(v+v')^* = (f+f')(v+v')P(v+v')^*(f+f')$$

is Pinsker in $(f+f')M(f+f')$. Now $(f+f')Q(f+f')$ and $(v+v')(e+e')P(e+e')(v+v')^*$ contain the common diffuse subalgebra

$$fQf \oplus f'Qf' = (v+v')[ePe \oplus e'Pe'](v+v')^*.$$

Since $(f+f')Q(f+f')$ and $(v+v')(e+e')P(e+e')(v+v')^*$ are both Pinsker in $(f+f')Q(f+f')$ and intersect diffusely, they must be equal. This contradicts the maximality of (e, f, v) and thus establishes that $wI_{(1-f)M(1-f)}((1-f)Q(1-f), u(1-e)P(1-e)u^*) = \{0\}$. □

Theorem 4.11 implies the following corollary about projections in the Pinsker algebras P and Q. Note that, in the case where P and Q are both factors, (2) below reduces to saying e is either 0 or 1, which yields Corollary 4.8. Thus, the following corollary can be understood as a generalization of Corollary 4.8.

Corollary 4.12. *Let P and Q be Pinsker subalgebras of a tracial von Neumann algebra (M, τ). Let e, f, and v be as in the previous theorem. Then the following hold*:

(1) *Let $e_1, e_2 \in P$ with $e_1 \leq 1-e$, $e_2 \leq e$, and $e_1 \sim_P e_2$. Let f_1 and f_2 satisfy the analogous conditions for Q and f. Then $ve_2v^* \wedge f_2 = 0$.*

(2) *In particular, if Q is a factor, then e is central in P.*

Proof. (1) Suppose $e_1, e_2 \in P$ with $e_1 \leq 1-e$, $e_2 \leq e$, and $e_1 \sim_P e_2$. Let f_1 and f_2 satisfy the analogous conditions for Q and f. Suppose for contradiction that $ve_2v^* \wedge f_2 \neq 0$. Let $f_2' = ve_2v^* \wedge f_2$. Let f_1' be the corresponding subprojection of f_1. Let $e_2' = v^* f_2' v$, and let e_1' be the corresponding subprojection of e_1. Then $e_1'Pe_1'$ is unitarily conjugate to $e_2'Pe_2'$ using the partial isometry that transforms e_1 to e_2, and similarly $f_1'Qf_1'$ is unitarily conjugate to $f_2'Qf_2'$, and $e_2'Pe_2'$ is unitarily conjugate to $f_2'Qf_2'$ using v. This implies that

$$wI_{(1-f)M(1-f)}((1-f)Q(1-f), v(1-e)P(1-e)v^*) \neq \{0\}$$

(or alternatively, it directly contradicts the maximality in Step 2 of the previous proof).

(2) Suppose Q is a factor, and assume for contradiction that e is not central in P. Then there must exist some projections $e_1, e_2 \in P$ with $e_1 \leq 1-e$, $e_2 \leq e$, and $e_1 \sim_P e_2$. Because Q is a factor, there exist projections f_1 and f_2 satisfying the analogous conditions for Q and f and with $f_2 = ve_2v^*$. Hence, we get a contradiction from (1). □

Example 4.13. Dykema [1993, Theorem 4.6] showed that if A and B are SOT-separable diffuse amenable tracial von Neumann algebras, then $A * B \cong L(\mathbb{F}_2)$. Taking two such pairs, there is an isomorphism $\alpha : A_1 * B_1 \to A_2 * B_2 = M$; let α be any such isomorphism. Note that A_1 and A_2 are Pinsker subalgebras by [Hayes et al. 2021b], and hence Theorem 4.7 applies to A_1 and A_2.

- In particular, suppose that $A_1 = \mathcal{R}$ and $A_2 = \mathcal{R}$. Then either $\alpha(A_1)$ and A_2 are unitarily conjugate, or $wI_M(A_2, \alpha(A_1)) = \{0\}$.
- Suppose that $A_1 = \mathcal{R} \oplus \mathcal{R}$ and $A_2 = \mathcal{R}$. Then the projection e from Theorem 4.11 must be central in A_1, and hence there are only four possible choices of e, resulting in a tetrachotomy.
- Generalizing Example 4.10, suppose $A_1 = L^\infty[0, 1]$ and $A_2 = \mathcal{R}$. Then, for any nonzero projections $e \in \alpha(A_1)$ and $f \in A_2$, the von Neumann algebras $e\alpha(A_1)e$ and fA_2f are not isomorphic. Hence the projection e in Theorem 4.11 must be zero, and hence $wI_M(A_2, \alpha(A_1)) = \{0\}$.

Appendix: Invariance of 1-bounded entropy via noncommutative functional calculus

A.1. ***Microstate spaces and definition of h.*** Here we recall definitions of the space of noncommutative laws. Let $\mathbb{C}\langle t_i : i \in I\rangle$ be the algebra of noncommutative complex polynomials in $(t_i)_{i\in I}$ (i.e., the free $\mathbb{C}$-algebra on the set I). We give $\mathbb{C}\langle t_i : i \in I\rangle$ the unique $*$-algebra structure which makes the t_i self-adjoint. If $\mathcal{M}$ is a W*-algebra and $\boldsymbol{x} = (x_i)_{i\in I} \in \mathcal{M}_{\mathrm{sa}}^I$, then there is a unique $*$-homomorphism

$$\mathrm{ev}_{\boldsymbol{x}} : \mathbb{C}\langle t_i : i \in I\rangle \to \mathcal{M}$$

such that $\mathrm{ev}_{\boldsymbol{x}}(t_i) = x_i$. For a noncommutative polynomial $p \in \mathbb{C}\langle t_i : i \in I\rangle$, we define $p(\boldsymbol{x}) = p((x_i)_{i\in I})$ to be $\mathrm{ev}_{\boldsymbol{x}}(p)$.

A tracial *noncommutative law* of a self-adjoint I-tuple is a linear functional $\lambda : \mathbb{C}\langle t_i : i \in I\rangle \to \mathbb{C}$ that is

(1) unital, that is, $\lambda(1) = 1$;

(2) positive, that is, $\lambda(p^*p) \geq 0$;

(3) tracial, that is, $\lambda(pq) = \lambda(qp)$;

(4) exponentially bounded, that is, for some $(R_i)_{i\in I} \in (0, +\infty)^I$, we have

$$|\lambda(t_{i(1)} \cdots t_{i(\ell)})| \leq R_{i(1)} \cdots R_{i(\ell)}$$

for all ℓ and all $i(1), \ldots, i(\ell) \in I$.

Given $\boldsymbol{R} = (R_i)_{i\in I} \in (0, +\infty)^I$, we define

$$\Sigma_{\boldsymbol{R}} = \Sigma_{(R_i)_{i\in I}}$$

to be the set of noncommutative laws satisfying (4) for our given choice of $(R_i)_{i\in I}$. We equip $\Sigma_{\boldsymbol{R}}$ with the topology of pointwise convergence on $\mathbb{C}\langle t_i : i \in I\rangle$, which makes it a compact Hausdorff space.

For a tracial W*-algebra $(\mathcal{M}, \tau)$, a tuple $\boldsymbol{x} = (x_i)_{i\in I} \in \mathcal{M}_{\mathrm{sa}}^I$, and $\boldsymbol{R} = (R_i)_{i\in I} \in (0, +\infty)^I$ satisfying $\|x_i\| \leq R_i$, we define the *noncommutative law of* $\boldsymbol{x}$ as the map

$$\lambda_{\boldsymbol{x}} : \mathbb{C}\langle t_i : i \in I\rangle \to \mathbb{C}, \quad p \mapsto \tau(p(x)).$$

It is straightforward to verify that $\lambda_{\boldsymbol{x}}$ is in $\Sigma_{\boldsymbol{R}}$. Conversely, given any $\lambda \in \Sigma_{\boldsymbol{R}}$, there exists some $(\mathcal{M}, \tau)$ and $\boldsymbol{x} \in \mathcal{M}_{\mathrm{sa}}^I$ such that $\lambda_{\boldsymbol{x}} = \lambda$ and $\|x_i\| \leq R_i$ for all $i \in I$; see either [Belinschi and Nica 2008, Proposition 4.2] or the proof of [Anderson et al. 2010, Proposition 5.2.14 (d)]. We also remark that $(\mathcal{M}, \tau)$ could be $M_n(\mathbb{C})$ with the normalized trace $\tau_n = (1/n)\,\mathrm{Tr}$. Thus, if $\boldsymbol{x} \in M_n(\mathbb{C})_{\mathrm{sa}}^I$, then $\lambda_{\boldsymbol{x}}$ is a well-defined noncommutative law.

The 1-bounded entropy h is defined in terms of Voiculescu's microstate spaces.

Definition A.1 (microstate space). Let $\mathcal{M}$ be a tracial von Neumann algebra and I an index set. Let $\boldsymbol{R} \in (0, +\infty)^I$, let $\boldsymbol{y} \in \mathcal{M}_{\mathrm{sa}}^I$ be a self-adjoint tuple with $\|y_i\| \leq R_i$, and let $\mathcal{O} \subseteq \Sigma_{\boldsymbol{R}}$ be a neighborhood of $\lambda_{\boldsymbol{y}}$. Then we define the microstate space

$$\Gamma_{\boldsymbol{R}}^{(n)}(\mathcal{O}) := \{\boldsymbol{Y} \in M_n(\mathbb{C})_{\mathrm{sa}}^I : \|Y_i\| \leq R_i \text{ for all } i \in I \text{ and } \lambda_{\boldsymbol{Y}} \in \mathcal{O}\}.$$

Definition A.2 (orbital covering numbers). Let I be an index set, and let $\Omega \subset M_n(\mathbb{C})^I_{\mathrm{sa}}$. For $F \subseteq I$ finite and $\varepsilon > 0$, we define the *orbital* (F, ε)*-neighborhood* of Ω as the set

$$N^{\mathrm{orb}}_{F,\varepsilon}(\Omega) = \{\boldsymbol{Y} \in M_n(\mathbb{C})^I_{\mathrm{sa}} : \text{there exists } \boldsymbol{Y}' \in \Omega,\ U \in \mathcal{U}(M_n(\mathbb{C})),\ \|Y_i - UY'_iU^*\|_2 < \varepsilon \text{ for } i \in F\}.$$

Moreover, we define the orbital covering number $K^{\mathrm{orb}}_{F,\varepsilon}(\Omega)$ as the minimal cardinality of a set Ω' such that $\Omega \subseteq N^{\mathrm{orb}}_{F,\varepsilon}(\Omega')$.

Definition A.3. Let (M, τ) be a tracial von Neumann algebra, let I and J be index sets, let $\boldsymbol{R} \in (0, \infty)^I$ and $\boldsymbol{S} \in (0, \infty)^J$, and let $\boldsymbol{x} \in M^I_{\mathrm{sa}}$ and $\boldsymbol{y} \in M^J_{\mathrm{sa}}$, with $\|x_i\| \leq R_i$ for $i \in I$ and $\|y_j\| \leq S_j$ for $j \in J$. For a neighborhood $\mathcal{O}$ of $\lambda_{(\boldsymbol{x},\boldsymbol{y})}$ in $\Sigma_{(\boldsymbol{R},\boldsymbol{S})}$, consider the microstate space $\Gamma^{(n)}_{(\boldsymbol{R},\boldsymbol{S})}(\mathcal{O}) \subseteq M_n(\mathbb{C})^{I\sqcup J}_{\mathrm{sa}}$, and let $\pi_I(\Gamma^{(n)}_{(\boldsymbol{R},\boldsymbol{S})}(\mathcal{O}))$ be its projection onto the I-indexed coordinates. Then define

$$h_{\boldsymbol{R},\boldsymbol{S}}(\boldsymbol{x} : \boldsymbol{y}) = \sup_{\substack{F\subseteq I \text{ finite}\\ \varepsilon>0}} \inf_{\mathcal{O}\ni\lambda_{(\boldsymbol{x},\boldsymbol{y})}} \limsup_{n\to\infty} \frac{1}{n^2} \log K^{\mathrm{orb}}_{F,\varepsilon}(\pi_I(\Gamma^{(n)}_{(\boldsymbol{R},\boldsymbol{S})}(\mathcal{O}))).$$

In this appendix we will give an argument showing, at the same time, that this computation yields the same result for every $\boldsymbol{R}$ and $\boldsymbol{S}$ with $R_i \geq \|x_i\|$ and $S_j \geq \|y_j\|$, and that $h(\boldsymbol{x} : \boldsymbol{y})$ only depends on $\mathrm{W}^*(\boldsymbol{x})$ and $\mathbf{W}^*(\boldsymbol{x}, \boldsymbol{y})$ (and the restriction of the trace to these algebras).

A.2. ***L^2-continuous functional calculus.*** Here we recall the construction of a certain space of noncommutative functions given as L^2-limits of trace polynomials, uniformly over all noncommutative laws. Trace polynomials have been studied in many previous works such as [Cébron 2013; Dabrowski et al. 2021; Driver et al. 2013; Jing 2015; Kemp 2016; 2017; Procesi 1976; Razmyslov 1974; 1985; Rains 1997; Sengupta 2008]. The uniform L^2-completion of trace polynomials was first introduced in [Jekel 2020a; 2020b; 2022], and its relationship with continuous model theory was addressed in [Jekel 2023, §3.5]. Moreover, [Hayes et al. 2021b, Remark 3.5] described how this space is an example of the tracial completions of C*-algebras studied in [Ozawa 2013, p. 351-352] and [Bosa et al. 2019] and implicitly in [Carrión et al. 2023, §6]. Here we follow the version of the construction in [Hayes et al. 2021b, §3].

Definition A.4. Fix an index set I and $\boldsymbol{R} \in (0, +\infty)^I$. Consider the space

$$\mathcal{A}_{\boldsymbol{R}} = C(\Sigma_{\boldsymbol{R}}) \otimes \mathbb{C}\langle t_i : i \in I\rangle.$$

Given $(\mathcal{M}, \tau)$ and $\boldsymbol{x} \in \mathcal{M}^I_{\mathrm{sa}}$ with $\|x_i\| \leq R_i$, we define the evaluation map

$$\mathrm{ev}_{\boldsymbol{x}} : \mathcal{A}_{\boldsymbol{R}} \to \mathcal{M}, \quad \phi \otimes p \mapsto \phi(\lambda_{\boldsymbol{x}})p(\boldsymbol{x}).$$

Then we define a semi-norm on $C(\Sigma_{\boldsymbol{R}}) \otimes \mathbb{C}\langle t_i : i \in I\rangle$ by

$$\|f\|_{\boldsymbol{R},2} = \sup_{(\mathcal{M},\tau),\boldsymbol{x}} \|\mathrm{ev}_{\boldsymbol{x}}(f)\|_{L^2(\mathcal{M})},$$

where the supremum is over all tracial W*-algebras $(\mathcal{M}, \tau)$ and all $\boldsymbol{x} \in \mathcal{M}^I_{\mathrm{sa}}$ with $\|x_i\| \leq R_i$. Denote by $\mathcal{F}_{\boldsymbol{R},2}$ the completion of $\mathcal{A}_{\boldsymbol{R}}/\{f \in \mathcal{A}_{\boldsymbol{R}} : \|f\|_{\boldsymbol{R},2} = 0\}$.

It is immediate that, for every $(\mathcal{M}, \tau)$, for every self-adjoint tuple $\boldsymbol{x} \in \mathcal{M}_{\mathrm{sa}}^I$ with $\|x_i\| \leq R_i$, the evaluation map $\mathrm{ev}_{\boldsymbol{x}} : \mathcal{A}_{\boldsymbol{R}} \to \mathcal{M}$ passes to a well-defined map $\mathcal{F}_{\boldsymbol{R},2} \to L^2(\mathcal{M})$, which we continue to denote by $\mathrm{ev}_{\boldsymbol{x}}$, and we will also define $f(\boldsymbol{x}) = \mathrm{ev}_{\boldsymbol{x}}(f)$. Moreover, it is clear that $f(\boldsymbol{x}) = \mathrm{ev}_{\boldsymbol{x}}(f)$ always lies in $L^2(\mathrm{W}^*(\boldsymbol{x}))$ because this holds when f is a simple tensor.

It will be convenient often to restrict our attention to elements of $\mathcal{F}_{\boldsymbol{R},2}$ that are bounded in operator norm. For $f \in \mathcal{F}_{\boldsymbol{R},2}$, let us define

$$\|f\|_{\boldsymbol{R},\infty} = \sup_{(\mathcal{M},\tau),\boldsymbol{x}} \|\mathrm{ev}_{\boldsymbol{x}}(f)\|$$

and then set

$$\mathcal{F}_{\boldsymbol{R},\infty} = \{f \in \mathcal{F}_{\boldsymbol{R},2} : \|f\|_{\boldsymbol{R},\infty} < +\infty\}.$$

It was shown in [Hayes et al. 2021b, Lemma 3.3] that $\mathcal{F}_{\boldsymbol{R},\infty}$ is a C*-algebra with respect to the norm $\|\cdot\|_{\boldsymbol{R},\infty}$ and the multiplication and $*$-operation arising from the natural ones on simple tensors.

One of the most useful properties of $\mathcal{F}_{\boldsymbol{R},\infty}$ is that it allows all the elements of a von Neumann algebra to be expressed as a function of the generators. More precisely, [Hayes et al. 2021b, Proposition 3.4] showed the following.

Proposition A.5. *Given $(\mathcal{M}, \tau)$ and $\boldsymbol{x} \in \mathcal{M}_{\mathrm{sa}}^I$ with $\|x_i\| \leq R_i$, the evaluation map $\mathrm{ev}_{\boldsymbol{x}} : \mathcal{F}_{\boldsymbol{R},2} \to L^2(\mathrm{W}^*(\boldsymbol{x}))$ is surjective. It also restricts to a surjective $*$-homomorphism $\mathcal{F}_{\boldsymbol{R},\infty} \to \mathrm{W}^*(\boldsymbol{x})$.*

This surjectivity property on its own is not too significant because, for instance, the double dual of the C*-universal free product of $C[-R_i, R_i]$ over $i \in I$ can be used to define a functional calculus that is surjective. The benefit of the construction used here is that it achieves surjectivity at the *same time* as relatively strong continuity properties.

First, we show that the noncommutative law of the output will depend continuously on the noncommutative law of the input. As we will see later, this property allows these functions to transform between microstate spaces for different generators of a von Neumann algebra. To fix notation, let I and I' be index sets. Let $\boldsymbol{R} \in (0, +\infty)^I$ and $\boldsymbol{R}' \in (0, +\infty)^{I'}$. We define

$$\mathcal{F}_{\boldsymbol{R},\boldsymbol{R}'} = \{\boldsymbol{f} = (f_i)_{i \in I'} \in (\mathcal{F}_{\boldsymbol{R},\infty})_{\mathrm{sa}}^{I'} : \|f_i\|_{\boldsymbol{R},\infty} \leq R_i' \text{ for all } i \in I'\}.$$

Proposition A.6 [Hayes et al. 2021b, Proposition 3.7]. *Let $\boldsymbol{R} \in (0, +\infty)^I$ and $\boldsymbol{R}' \in (0, +\infty)^{I'}$. Let $\boldsymbol{f} = (f_i)_{i \in I'} \in \mathcal{F}_{\boldsymbol{R},\boldsymbol{R}'}$.*

(1) *Given $(\mathcal{M}, \tau)$ and $\boldsymbol{x} \in \mathcal{M}_{\mathrm{sa}}^I$ with $\|x_i\| \leq R_i$, we set $\boldsymbol{f}(\boldsymbol{x}) = (f_i(\boldsymbol{x}))_{i \in I'}$. Then $\lambda_{\boldsymbol{f}(\boldsymbol{x})}$ is uniquely determined by $\lambda_{\boldsymbol{x}}$.*

(2) *Let $\boldsymbol{f}_*$ be the "push-forward" mapping $\Sigma_{\boldsymbol{R}} \to \Sigma_{\boldsymbol{R}'}$ defined by $\boldsymbol{f}_* \lambda_{\boldsymbol{x}} = \lambda_{\boldsymbol{f}(\boldsymbol{x})}$ for all such tuples $\boldsymbol{x}$. Then $\boldsymbol{f}_*$ is continuous.*

Another consequence of this is that these spaces of functions are closed under composition.

Proposition A.7. *Fix index sets I, I', and I'' and corresponding tuples $\boldsymbol{R}$, $\boldsymbol{R}'$, and $\boldsymbol{R}''$ from $(0, \infty)$. Let $\boldsymbol{f} \in \mathcal{F}_{\boldsymbol{R},\boldsymbol{R}'}$ and $\boldsymbol{g} \in \mathcal{F}_{\boldsymbol{R}',\boldsymbol{R}''}$. Then there exists a unique $\boldsymbol{h} \in \mathcal{F}_{\boldsymbol{R},\boldsymbol{R}'}$ such that $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))$ for all M and $\boldsymbol{x} \in M^I$ with $\|x_i\| \leq R_i$.*

Proof. First, we consider $g \in \mathcal{F}_{\boldsymbol{R}',2}$ and show that $g \circ \boldsymbol{f}$ is a well-defined element of $\mathcal{F}_{\boldsymbol{R}',2}$. If g is a simple tensor $\phi \otimes p$, then $\phi(\lambda_{\boldsymbol{f}(\boldsymbol{x})}) = \phi \circ \boldsymbol{f}_* \lambda_{\boldsymbol{x}}$ defines a continuous function on the space of laws by the previous proposition. Also, since $\mathcal{F}_{\mathcal{R},\infty}$ is a C*-algebra, $p \circ \boldsymbol{f} \in \mathcal{F}_{\boldsymbol{R},\infty}$ and hence so is $g \circ \boldsymbol{f} = (\phi \circ \boldsymbol{f}_* \otimes 1)(p \circ \boldsymbol{f})$.

Next, if g is a linear combination of simple tensors, one can check directly that $\|g \circ \boldsymbol{f}\|_{\boldsymbol{R},2} \leq \|g\|_{\boldsymbol{R}',2}$ by considering evaluations on all possible tuples. This estimate allows us to pass to the completion, showing that if $g \in \mathcal{F}_{\boldsymbol{R}',2}$, then $g \circ \boldsymbol{f} \in \mathcal{F}_{\boldsymbol{R},2}$.

Again, by evaluating on points, we see that $\|g \circ \boldsymbol{f}\|_{\boldsymbol{R},\infty} \leq \|g\|_{\boldsymbol{R}',\infty}$. Replacing the single function g by an I''-tuple yields the asserted result. □

The second continuity property that we need for $\mathcal{F}_{\boldsymbol{R},\boldsymbol{R}'}$ is uniform continuity with respect to the L^2 norm. This will allow us to use the functions in $\mathcal{F}_{\boldsymbol{R},\boldsymbol{R}'}$ to "push forward" ε-coverings from one microstate space to another.

Proposition A.8 [Hayes et al. 2021b, Proposition 3.9]. *Let I be an index set and $\boldsymbol{R} \in (0,+\infty)^I$, and let $f \in \mathcal{F}_{\boldsymbol{R},2}$. Then, for every $\varepsilon > 0$, there exists a finite $F \subseteq I$ and a $\delta > 0$ such that, for every $(\mathcal{M}, \tau)$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}_{\mathrm{sa}}^I$ with $\|x_i\|, \|y_i\| \leq R_i$, if $\|x_i - y_i\|_2 < \delta$ for all $i \in F$, then $\|f(\boldsymbol{x}) - f(\boldsymbol{y})\|_2 < \varepsilon$.*

A.3. *Proof of monotonicity and invariance.*

Theorem A.9. *Let (M, τ) be a tracial von Neumann algebra. Let I and J be index sets, let $\boldsymbol{R} \in (0,\infty)^I$ and $\boldsymbol{S} \in (0,\infty)^J$, and let $\boldsymbol{x} \in M_{\mathrm{sa}}^I$ and $\boldsymbol{y} \in M_{\mathrm{sa}}^J$, with $\|x_i\| \leq R_i$ for $i \in I$ and $\|y_j\| \leq S_j$ for $j \in J$. Let I', J', $\boldsymbol{R}'$, $\boldsymbol{S}'$, $\boldsymbol{x}' \in M_{\mathrm{sa}}^{I'}$, $\boldsymbol{y}' \in M_{\mathrm{sa}}^{J'}$ satisfy similar conditions. Suppose that*

$$\mathrm{W}^*(\boldsymbol{x}, \boldsymbol{y}) \supseteq \mathrm{W}^*(\boldsymbol{x}', \boldsymbol{y}') \quad \text{and} \quad \mathrm{W}^*(\boldsymbol{x}) \subseteq \mathrm{W}^*(\boldsymbol{x}').$$

Then

$$h_{\boldsymbol{R},\boldsymbol{S}}(\boldsymbol{x} : \boldsymbol{y}) \leq h_{\boldsymbol{R}',\boldsymbol{S}'}(\boldsymbol{x}' : \boldsymbol{y}').$$

Proof. Unwinding the suprema and infima in the definition of h, it suffices to show that, for every $F \subseteq I$ finite and $\varepsilon > 0$, there exists $F' \subseteq I'$ finite and $\varepsilon' > 0$ such that, for every neighborhood $\mathcal{O}'$ of $\lambda_{(\boldsymbol{x}',\boldsymbol{y}')}$ in $\Sigma_{(\boldsymbol{R}',\boldsymbol{S}')}$, there exists a neighborhood $\mathcal{O}$ of $\lambda_{(\boldsymbol{x},\boldsymbol{y})}$ in $\Sigma_{(\boldsymbol{R},\boldsymbol{S})}$ such that

$$K^{\mathrm{orb}}_{F,\varepsilon}(\pi_I(\Gamma^{(n)}_{(\boldsymbol{R},\boldsymbol{S})}(\mathcal{O}))) \leq K^{\mathrm{orb}}_{F',\varepsilon'}(\pi_{I'}(\Gamma^{(n)}_{(\boldsymbol{R}',\boldsymbol{S}')}(\mathcal{O}'))).$$

Fix F and ε. Because $\mathrm{W}^*(\boldsymbol{x}) \subseteq \mathrm{W}^*(\boldsymbol{x}')$, there exists $\boldsymbol{f} \in \mathcal{F}_{\boldsymbol{R}',\boldsymbol{R}}$ such that $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}')$. By Proposition A.8, there exists $\varepsilon' > 0$ and $F' > 0$ such that, for all tracial von Neumann algebras N and all $\boldsymbol{z}, \boldsymbol{w} \in N^{I'}$ with $\|z_i\|, \|w_i\| \leq R_i'$, we have that $\max_{i \in F'} \|z_i - w_i\|_2 < 2\varepsilon'$ implies $\max_{i \in F} \|f_i(\boldsymbol{z}) - f_i(\boldsymbol{w})\|_2 < \frac{1}{2}\varepsilon$.

Now fix a neighborhood $\mathcal{O}'$ of $\lambda_{\boldsymbol{x}',\boldsymbol{y}'}$ in $\Sigma_{(\boldsymbol{R}',\boldsymbol{S}')}$. We claim that

$$K^{\mathrm{orb}}_{F,\varepsilon/2}(\boldsymbol{f}(\pi_{I'}(\Gamma^{(n)}_{\boldsymbol{R}',\boldsymbol{S}'}(\mathcal{O}')))) \leq K^{\mathrm{orb}}_{F',\varepsilon'}(\pi_{I'}(\Gamma^{(n)}_{\boldsymbol{R}',\boldsymbol{S}'}(\mathcal{O}'))).$$

Indeed, let Ω be a set of cardinality $K^{\mathrm{orb}}_{F',\varepsilon'}(\pi_I(\Gamma^{(n)}_{\boldsymbol{R},\boldsymbol{S}}(\mathcal{O}')))$ such that $\pi_I(\Gamma^{(n)}_{\boldsymbol{R},\boldsymbol{S}}(\mathcal{O}')) \subseteq N^{\mathrm{orb}}_{F',\varepsilon'}(\Omega)$. Let $\Omega' \subseteq \pi_I(\Gamma^{(n)}_{\boldsymbol{R},\boldsymbol{S}}(\mathcal{O}'))$ be chosen to have one element within ε' of each element of Ω, so that

$$\pi_I(\Gamma^{(n)}_{\boldsymbol{R},\boldsymbol{S}}(\mathcal{O}')) \subseteq N^{\mathrm{orb}}_{F',2\varepsilon'}(\Omega').$$

Then each $\boldsymbol{X}' \in \Omega'$, and more generally in $\pi_I(\Gamma^{(n)}_{\boldsymbol{R},\boldsymbol{S}}(\mathcal{O}'))$ satisfies $\|X_i'\| \le R_i$, so that it is valid to apply the uniform continuity estimate for $\boldsymbol{f}$ to such points $\boldsymbol{X}'$. The choice of (F, ε) thus implies that

$$\boldsymbol{f}(\pi_{I'}(\Gamma^{(n)}_{\boldsymbol{R}',\boldsymbol{S}'}(\mathcal{O}'))) \subseteq N^{\mathrm{orb}}_{F',\varepsilon'/2}(\boldsymbol{f}(\Omega')),$$

which proves our claim about the covering numbers.

Next, we describe how to choose $\mathcal{O}$. Since $\mathrm{W}^*(\boldsymbol{x}', \boldsymbol{y}') \subseteq \mathrm{W}^*(\boldsymbol{x}, \boldsymbol{y})$, there exists $\boldsymbol{g} \in \mathcal{F}_{(\boldsymbol{R},\boldsymbol{S}),(\boldsymbol{R}',\boldsymbol{S}')}$ such that $(\boldsymbol{x}', \boldsymbol{y}') = \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y})$. By continuity of $\boldsymbol{g}_* : \Sigma_{(\boldsymbol{R},\boldsymbol{S})} \to \Sigma_{(\boldsymbol{R}',\boldsymbol{S}')}$, the set

$$\mathcal{O}_1 = (\boldsymbol{g}_*)^{-1}(\mathcal{O}')$$

is open. Let

$$\mathcal{O}_2 = \big\{\lambda_{(\boldsymbol{z},\boldsymbol{w})} \in \Sigma_{(\boldsymbol{R},\boldsymbol{S})} : \max_{i\in F} \|f_i \circ \pi_{I'} \circ \boldsymbol{g}(\boldsymbol{z}, \boldsymbol{w}) - z_i\|_2 < \tfrac{1}{2}\varepsilon\big\}.$$

The set $\mathcal{O}_2$ is open using Propositions A.7 and A.6. It also contains $\lambda_{(\boldsymbol{x},\boldsymbol{y})}$ because

$$\boldsymbol{f}(\pi_{I'}(\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}))) = \boldsymbol{f}(\pi_{I'}(\boldsymbol{x}', \boldsymbol{y}')) = \boldsymbol{f}(\boldsymbol{x}') = \boldsymbol{x}.$$

Let $\mathcal{O} = \mathcal{O}_1 \cap \mathcal{O}_2$. We claim that

$$\pi_I \Gamma^{(n)}_{\boldsymbol{R},\boldsymbol{S}}(\mathcal{O}) \subseteq N^{\mathrm{orb}}_{F,\varepsilon/2}(\boldsymbol{f}(\pi_{I'}(\Gamma^{(n)}_{\boldsymbol{R}',\boldsymbol{S}'}(\mathcal{O}')))).$$

Indeed, if $\boldsymbol{X}$ is in the set on the left-hand side, then there exists $\boldsymbol{Y}$ such that $\lambda_{(\boldsymbol{X},\boldsymbol{Y})} \in \mathcal{O}$. In particular, this means that $\lambda_{\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Y})} \in \mathcal{O}'$, or in other words $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Y}) \in \Gamma^{(n)}_{\boldsymbol{R}',\boldsymbol{S}'}(\mathcal{O}')$. Moreover,

$$\max_{i\in F} \|X_i - f_i \circ \pi_{I'} \circ \boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Y})\|_2 < \tfrac{1}{2}\varepsilon.$$

Therefore, $\boldsymbol{X}$ is in the $(F, \varepsilon/2)$-neighborhood of $\boldsymbol{f} \circ \pi_{I'}$ of some point in $\Gamma^{(n)}_{\boldsymbol{R}',\boldsymbol{S}'}(\mathcal{O}')$, which proves the claimed inclusion.

This inclusion $\pi_I \Gamma^{(n)}_{\boldsymbol{R},\boldsymbol{S}}(\mathcal{O}) \subseteq N^{\mathrm{orb}}_{F,\varepsilon/2}(\boldsymbol{f}(\pi_{I'}(\Gamma^{(n)}_{\boldsymbol{R}',\boldsymbol{S}'}(\mathcal{O}'))))$ in turn implies that

$$K^{\mathrm{orb}}_{F,\varepsilon}(\pi_I \Gamma^{(n)}_{\boldsymbol{R},\boldsymbol{S}}(\mathcal{O})) \le K^{\mathrm{orb}}_{F,\varepsilon/2}(\boldsymbol{f}(\pi_{I'}(\Gamma^{(n)}_{\boldsymbol{R}',\boldsymbol{S}'}(\mathcal{O}')))) \le K^{\mathrm{orb}}_{F',\varepsilon'}(\pi_{I'}(\Gamma^{(n)}_{(\boldsymbol{R}',\boldsymbol{S}')}(\mathcal{O}'))),$$

where the second inequality is the earlier claim that we proved. □

This theorem implies the following:

- In the case where $\boldsymbol{x} = \boldsymbol{x}'$ and $\boldsymbol{y} = \boldsymbol{y}'$, the theorem shows that $h_{\boldsymbol{S},\boldsymbol{R}}(\boldsymbol{x} : \boldsymbol{y})$ is independent of $\boldsymbol{R}$ and $\boldsymbol{S}$ so long as $\|x_i\| \le R_i$ for $i \in I$ and $\|y_j\| \le S_j$ for $j \in J$. Thus, we may unambiguously write $h(\boldsymbol{x} : \boldsymbol{y})$.
- In the case where $\mathrm{W}^*(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{W}^*(\boldsymbol{x}', \boldsymbol{y}')$ and $\mathrm{W}^*(\boldsymbol{x}) = \mathrm{W}^*(\boldsymbol{x}')$, the theorem shows that $h(\boldsymbol{x} : \boldsymbol{y}) = h(\boldsymbol{x}' : \boldsymbol{y}')$. Hence, for $N \le M$, we may unambiguously define $h(N : M)$ as $h(\boldsymbol{x} : \boldsymbol{y})$ for some tuples $\boldsymbol{x}$ and $\boldsymbol{y}$ such that $N = \mathrm{W}^*(\boldsymbol{x})$ and $M = \mathrm{W}^*(\boldsymbol{x}, \boldsymbol{y})$.
- Now suppose that $P \le N \le M$. Applying the theorem in the case where $\mathrm{W}^*(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{W}^*(\boldsymbol{x}', \boldsymbol{y}') = M$ and $P = \mathrm{W}^*(\boldsymbol{x}) \subseteq \mathrm{W}^*(\boldsymbol{x}') = N$, we obtain $h(P : M) \le h(N : M)$.
- Again, suppose $P \le N \le M$. Applying the theorem in the case where $\mathrm{W}^*(\boldsymbol{x}, \boldsymbol{y}) = M \supseteq N = \mathrm{W}^*(\boldsymbol{x}', \boldsymbol{y}')$ and $P = \mathrm{W}^*(\boldsymbol{x}) = \mathrm{W}^*(\boldsymbol{x}')$, we obtain $h(P : M) \le h(P : N)$.

Acknowledgements

We thank Enes Kurt, Sorin Popa, Thomas Sinclair, and Stefaan Vaes for providing many helpful comments that improved the exposition.

B. Hayes gratefully acknowledges support from the NSF grant DMS-2000105. D. Jekel gratefully acknowledges support from the NSF grant DMS-2002826. S. Kunnawalkam Elayavalli gratefully acknowledges support from the Simons Postdoctoral Fellowship.

References

[Anderson et al. 2010] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An introduction to random matrices*, Cambridge Stud. in Adv. Math. **118**, Cambridge Univ. Press, 2010. MR Zbl

[Bandeira et al. 2023] A. S. Bandeira, M. T. Boedihardjo, and R. van Handel, "Matrix concentration inequalities and free probability", *Invent. Math.* **234**:1 (2023), 419–487. MR Zbl

[Belinschi and Capitaine 2022] S. Belinschi and M. Capitaine, "Strong convergence of tensor products of independent G.U.E. matrices", preprint, 2022. arXiv 2205.07695

[Belinschi and Nica 2008] S. T. Belinschi and A. Nica, "η-series and a Boolean Bercovici–Pata bijection for bounded k-tuples", *Adv. Math.* **217**:1 (2008), 1–41. MR Zbl

[Bordenave and Collins 2023] C. Bordenave and B. Collins, "Norm of matrix-valued polynomials in random unitaries and permutations", preprint, 2023. arXiv 2304.05714

[Bosa et al. 2019] J. Bosa, N. P. Brown, Y. Sato, A. Tikuisis, S. White, and W. Winter, *Covering dimension of* C**-algebras and* 2*-coloured classification*, Mem. Amer. Math. Soc. **1233**, Amer. Math. Soc., Providence, RI, 2019. MR Zbl

[Boutonnet and Popa 2023] R. Boutonnet and S. Popa, "Maximal amenable MASAs of radial type in the free group factors", 2023. To appear in *Proc. Amer. Math. Soc.* arXiv 2302.13355

[Bowen 2010] L. Bowen, "Measure conjugacy invariants for actions of countable sofic groups", *J. Amer. Math. Soc.* **23**:1 (2010), 217–245. MR Zbl

[Brothier and Wen 2016] A. Brothier and C. Wen, "The cup subalgebra has the absorbing amenability property", *Int. J. Math.* **27**:2 (2016), art. id. 1650013. MR Zbl

[Cameron et al. 2010] J. Cameron, J. Fang, M. Ravichandran, and S. White, "The radial masa in a free group factor is maximal injective", *J. Lond. Math. Soc.* (2) **82**:3 (2010), 787–809. MR Zbl

[Carrión et al. 2023] J. R. Carrión, J. Gabe, C. Schafhauser, A. Tikuisis, and S. White, "Classifying $*$-homomorphisms, I: Unital simple nuclear C^*-algebras", preprint, 2023. arXiv 2307.06480

[Cébron 2013] G. Cébron, "Free convolution operators and free Hall transform", *J. Funct. Anal.* **265**:11 (2013), 2645–2708. MR Zbl

[Charlesworth et al. 2023] I. Charlesworth, R. de Santiago, B. Hayes, D. Jekel, S. Kunnawalkam Elayavalli, and B. Nelson, "Strong 1-boundedness, L^2-Betti numbers, algebraic soficity, and graph products", 2023. To appear in *Kyoto J. Math.* arXiv 2305.19463

[Chifan and Sinclair 2013] I. Chifan and T. Sinclair, "On the structural theory of II_1 factors of negatively curved groups", *Ann. Sci. École Norm. Sup.* (4) **46**:1 (2013), 1–33. MR Zbl

[Chifan et al. 2023] I. Chifan, A. Ioana, and S. Kunnawalkam Elayavalli, "An exotic II_1 factor without property gamma", *Geom. Funct. Anal.* **33**:5 (2023), 1243–1265. MR Zbl

[Collins et al. 2022] B. Collins, A. Guionnet, and F. Parraud, "On the operator norm of non-commutative polynomials in deterministic matrices and iid GUE matrices", *Camb. J. Math.* **10**:1 (2022), 195–260. MR Zbl

[Connes 1976] A. Connes, "Classification of injective factors: cases II_1, II_∞, III_λ, $\lambda \neq 1$", *Ann. of Math.* (2) **104**:1 (1976), 73–115. MR Zbl

[Dabrowski et al. 2021] Y. Dabrowski, A. Guionnet, and D. Shlyakhtenko, "Free transport for convex potentials", *New Zealand J. Math.* **52** (2021), 259–359. MR Zbl

[Ding and Kunnawalkam Elayavalli 2024] C. Ding and S. Kunnawalkam Elayavalli, "Structure of relatively biexact group von Neumann algebras", *Comm. Math. Phys.* **405**:4 (2024), art. id. 104. MR Zbl

[Ding et al. 2023] C. Ding, S. Kunnawalkam Elayavalli, and J. Peterson, "Properly proximal von Neumann algebras", *Duke Math. J.* **172**:15 (2023), 2821–2894. MR Zbl

[Driver et al. 2013] B. K. Driver, B. C. Hall, and T. Kemp, "The large-N limit of the Segal–Bargmann transform on $\mathbb{U}_N$", *J. Funct. Anal.* **265**:11 (2013), 2585–2644. MR Zbl

[Dykema 1993] K. Dykema, "Free products of hyperfinite von Neumann algebras and free dimension", *Duke Math. J.* **69**:1 (1993), 97–119. MR Zbl

[Dykema 1994] K. Dykema, "Interpolated free group factors", *Pacific J. Math.* **163**:1 (1994), 123–135. MR Zbl

[Dykema and Mukherjee 2013] K. Dykema and K. Mukherjee, "Measure-multiplicity of the Laplacian masa", *Glasg. Math. J.* **55**:2 (2013), 285–292. MR Zbl

[Dykema et al. 2006] K. J. Dykema, A. M. Sinclair, and R. R. Smith, "Values of the Pukánszky invariant in free group factors and the hyperfinite factor", *J. Funct. Anal.* **240**:2 (2006), 373–398. MR Zbl

[Feldman and Moore 1977] J. Feldman and C. C. Moore, "Ergodic equivalence relations, cohomology, and von Neumann algebras, II", *Trans. Amer. Math. Soc.* **234**:2 (1977), 325–359. MR Zbl

[Galatan and Popa 2017] A. Galatan and S. Popa, "Smooth bimodules and cohomology of II_1 factors", *J. Inst. Math. Jussieu* **16**:1 (2017), 155–187. MR Zbl

[Ge 1998] L. Ge, "Applications of free entropy to finite von Neumann algebras, II", *Ann. of Math.* (2) **147**:1 (1998), 143–157. MR Zbl

[Ge and Popa 1998] L. Ge and S. Popa, "On some decomposition properties for factors of type II_1", *Duke Math. J.* **94**:1 (1998), 79–101. MR Zbl

[Haagerup 1985] U. Haagerup, "Injectivity and decomposition of completely bounded maps", pp. 170–222 in *Operator algebras and their connections with topology and ergodic theory* (Buşteni, Romania, 1983), edited by H. Araki et al., Lecture Notes in Math. **1132**, Springer, 1985. MR Zbl

[Hayes 2018] B. Hayes, "1-bounded entropy and regularity problems in von Neumann algebras", *Int. Math. Res. Not.* **2018**:1 (2018), 57–137. MR Zbl

[Hayes 2022] B. Hayes, "A random matrix approach to the Peterson–Thom conjecture", *Indiana Univ. Math. J.* **71**:3 (2022), 1243–1297. MR Zbl

[Hayes et al. 2021a] B. Hayes, D. Jekel, and S. Kunnawalkam Elayavalli, "Property (T) and strong 1-boundedness for von Neumann algebras", 2021. To appear in *J. Inst. Math. Jussieu*. arXiv 2107.03278

[Hayes et al. 2021b] B. Hayes, D. Jekel, B. Nelson, and T. Sinclair, "A random matrix approach to absorption in free products", *Int. Math. Res. Not.* **2021**:3 (2021), 1919–1979. MR Zbl

[Hayes et al. 2024] B. Hayes, D. Jekel, and S. Kunnawalkam Elayavalli, "Vanishing first cohomology and strong 1-boundedness for von Neumann algebras", *J. Noncommut. Geom.* **18**:2 (2024), 383–409. MR Zbl

[Houdayer 2015] C. Houdayer, "Gamma stability in free product von Neumann algebras", *Comm. Math. Phys.* **336**:2 (2015), 831–851. MR Zbl

[Ioana et al. 2008] A. Ioana, J. Peterson, and S. Popa, "Amalgamated free products of weakly rigid factors and calculation of their symmetry groups", *Acta Math.* **200**:1 (2008), 85–153. MR Zbl

[Izumi et al. 1998] M. Izumi, R. Longo, and S. Popa, "A Galois correspondence for compact groups of automorphisms of von Neumann algebras with a generalization to Kac algebras", *J. Funct. Anal.* **155**:1 (1998), 25–63. MR Zbl

[Jekel 2020a] D. Jekel, "An elementary approach to free entropy theory for convex potentials", *Anal. PDE* **13**:8 (2020), 2289–2374. MR Zbl

[Jekel 2020b] D. A. Jekel, *Evolution equations in non-commutative probability*, Ph.D. thesis, University of California, Los Angeles, 2020, available at https://www.proquest.com/docview/2415413745.

[Jekel 2022] D. Jekel, "Conditional expectation, entropy, and transport for convex Gibbs laws in free probability", *Int. Math. Res. Not.* **2022**:6 (2022), 4514–4619. MR Zbl

[Jekel 2023] D. Jekel, "Covering entropy for types in tracial W*-algebras", *J. Log. Anal.* **15** (2023), art. id. 2. MR Zbl

[Jing 2015] N. Jing, "Unitary and orthogonal equivalence of sets of matrices", *Linear Algebra Appl.* **481** (2015), 235–242. MR Zbl

[Jung 2007] K. Jung, "Strongly 1-bounded von Neumann algebras", *Geom. Funct. Anal.* **17**:4 (2007), 1180–1200. MR Zbl

[Kemp 2016] T. Kemp, "The large-N limits of Brownian motions on $\mathbb{GL}_N$", *Int. Math. Res. Not.* **2016**:13 (2016), 4012–4057. MR Zbl

[Kemp 2017] T. Kemp, "Heat kernel empirical laws on $\mathbb{U}_N$ and $\mathbb{GL}_N$", *J. Theoret. Probab.* **30**:2 (2017), 397–451. MR Zbl

[Kerr 2014] D. Kerr, "Bernoulli actions of sofic groups have completely positive entropy", *Israel J. Math.* **202**:1 (2014), 461–474. MR Zbl

[Kerr and Li 2011] D. Kerr and H. Li, "Entropy and the variational principle for actions of sofic groups", *Invent. Math.* **186**:3 (2011), 501–558. MR Zbl

[Kieffer 1975] J. C. Kieffer, "A generalized Shannon–McMillan theorem for the action of an amenable group on a probability space", *Ann. Probab.* **3**:6 (1975), 1031–1037. MR Zbl

[Kolmogorov 1958] A. N. Kolmogorov, "A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces", *Dokl. Akad. Nauk SSSR* (*N.S.*) **119** (1958), 861–864. In Russian. MR Zbl

[Kunnawalkam Elayavalli 2023] S. Kunnawalkam Elayavalli, "Remarks on the diagonal embedding and strong 1-boundedness", *Doc. Math.* **28**:3 (2023), 671–681. MR Zbl

[Mukherjee 2013] K. Mukherjee, "Singular masas and measure-multiplicity invariant", *Houston J. Math.* **39**:2 (2013), 561–598. MR Zbl

[Murray and von Neumann 1936] F. J. Murray and J. von Neumann, "On rings of operators", *Ann. of Math.* (2) **37**:1 (1936), 116–229. MR Zbl

[Neshveyev and Størmer 2002] S. Neshveyev and E. Størmer, "Ergodic theory and maximal abelian subalgebras of the hyperfinite factor", *J. Funct. Anal.* **195**:2 (2002), 239–261. MR Zbl

[Ornstein and Weiss 1987] D. S. Ornstein and B. Weiss, "Entropy and isomorphism theorems for actions of amenable groups", *J. Anal. Math.* **48** (1987), 1–141. MR Zbl

[Ozawa 2009] N. Ozawa, "An example of a solid von Neumann algebra", *Hokkaido Math. J.* **38**:3 (2009), 557–561. MR Zbl

[Ozawa 2013] N. Ozawa, "Dixmier approximation and symmetric amenability for C*-algebras", *J. Math. Sci. Univ. Tokyo* **20**:3 (2013), 349–374. MR Zbl

[Ozawa and Popa 2010a] N. Ozawa and S. Popa, "On a class of II_1 factors with at most one Cartan subalgebra", *Ann. of Math.* (2) **172**:1 (2010), 713–749. MR Zbl

[Ozawa and Popa 2010b] N. Ozawa and S. Popa, "On a class of II_1 factors with at most one Cartan subalgebra, II", *Amer. J. Math.* **132**:3 (2010), 841–866. MR Zbl

[Parekh et al. 2018] S. Parekh, K. Shimada, and C. Wen, "Maximal amenability of the generator subalgebra in q-Gaussian von Neumann algebras", *J. Operator Theory* **80**:1 (2018), 125–152. MR Zbl

[Parraud 2023] F. Parraud, "Asymptotic expansion of smooth functions in polynomials in deterministic matrices and iid GUE matrices", *Comm. Math. Phys.* **399**:1 (2023), 249–294. MR Zbl

[Peterson 2009] J. Peterson, "L^2-rigidity in von Neumann algebras", *Invent. Math.* **175**:2 (2009), 417–433. MR Zbl

[Peterson and Thom 2011] J. Peterson and A. Thom, "Group cocycles and the ring of affiliated operators", *Invent. Math.* **185**:3 (2011), 561–592. MR Zbl

[Pimsner and Popa 1986] M. Pimsner and S. Popa, "Entropy and index for subfactors", *Ann. Sci. École Norm. Sup.* (4) **19**:1 (1986), 57–106. MR Zbl

[Popa 1983a] S. Popa, "Maximal injective subalgebras in factors associated with free groups", *Adv. Math.* **50**:1 (1983), 27–48. MR Zbl

[Popa 1983b] S. Popa, "Orthogonal pairs of $*$-subalgebras in finite von Neumann algebras", *J. Operator Theory* **9**:2 (1983), 253–268. MR Zbl

[Popa 1999] S. Popa, "Some properties of the symmetric enveloping algebra of a subfactor, with applications to amenability and property T", *Doc. Math.* **4** (1999), 665–744. MR Zbl

[Popa 2005] S. Popa, "Deformation-rigidity theory", NCGOA conference mini-course, Vanderbilt University, 2005.

[Popa 2006a] S. Popa, "On a class of type II_1 factors with Betti numbers invariants", *Ann. of Math.* (2) **163**:3 (2006), 809–899. MR Zbl

[Popa 2006b] S. Popa, "Some computations of 1-cohomology groups and construction of non-orbit-equivalent actions", *J. Inst. Math. Jussieu* **5**:2 (2006), 309–332. MR Zbl

[Popa 2006c] S. Popa, "Strong rigidity of II_1 factors arising from malleable actions of w-rigid groups, I", *Invent. Math.* **165**:2 (2006), 369–408. MR Zbl

[Popa 2007] S. Popa, "On Ozawa's property for free group factors", *Int. Math. Res. Not.* **2007**:11 (2007), art. id. rnm036. MR Zbl

[Popa 2019] S. Popa, "Constructing MASAs with prescribed properties", *Kyoto J. Math.* **59**:2 (2019), 367–397. MR Zbl

[Popa 2021] S. Popa, "Coarse decomposition of II_1 factors", *Duke Math. J.* **170**:14 (2021), 3073–3110. MR Zbl

[Popa and Vaes 2014a] S. Popa and S. Vaes, "Unique Cartan decomposition for II_1 factors arising from arbitrary actions of free groups", *Acta Math.* **212**:1 (2014), 141–198. MR Zbl

[Popa and Vaes 2014b] S. Popa and S. Vaes, "Unique Cartan decomposition for II_1 factors arising from arbitrary actions of hyperbolic groups", *J. Reine Angew. Math.* **694** (2014), 215–239. MR Zbl

[Procesi 1976] C. Procesi, "The invariant theory of $n \times n$ matrices", *Adv. Math.* **19**:3 (1976), 306–381. MR Zbl

[Pukánszky 1960] L. Pukánszky, "On maximal abelian subrings of factors of type II_1", *Canad. J. Math.* **12** (1960), 289–296. MR Zbl

[Rains 1997] E. M. Rains, "Combinatorial properties of Brownian motion on the compact classical groups", *J. Theoret. Probab.* **10**:3 (1997), 659–679. MR Zbl

[Razmyslov 1974] Y. P. Razmyslov, "Trace identities of full matrix algebras over a field of characteristic zero", *Izv. Akad. Nauk SSSR Ser. Mat.* **38** (1974), 723–756. In Russian; translated in *Math. USSR-Izv.* **8**:4 (1974), 727–760. MR Zbl

[Razmyslov 1985] Y. P. Razmyslov, "Trace identities and central polynomials in the matrix superalgebras $M_{n,k}$", *Mat. Sb.* (*N.S.*) **128(170)**:2 (1985), 194–215. In Russian; translated in *Math. USSR-Sb.* **56**:1 (1987), 187–206. MR Zbl

[Robertson and Steger 2010] G. Robertson and T. Steger, "Malnormal subgroups of lattices and the Pukánszky invariant in group factors", *J. Funct. Anal.* **258**:8 (2010), 2708–2713. MR Zbl

[Rădulescu 1991] F. Rădulescu, "Singularity of the radial subalgebra of $\mathcal{L}(F_N)$ and the Pukánszky invariant", *Pacific J. Math.* **151**:2 (1991), 297–306. MR Zbl

[Rădulescu 1994] F. Rădulescu, "Random matrices, amalgamated free products and subfactors of the von Neumann algebra of a free group, of noninteger index", *Invent. Math.* **115**:2 (1994), 347–389. MR Zbl

[Rudolph and Weiss 2000] D. J. Rudolph and B. Weiss, "Entropy and mixing for amenable group actions", *Ann. of Math.* (2) **151**:3 (2000), 1119–1150. MR Zbl

[de Santiago et al. 2021] R. de Santiago, B. Hayes, D. J. Hoff, and T. Sinclair, "Maximal rigid subalgebras of deformations and L^2-cohomology", *Anal. PDE* **14**:7 (2021), 2269–2306. MR Zbl

[Sengupta 2008] A. N. Sengupta, "Traces in two-dimensional QCD: the large-N limit", pp. 193–212 in *Traces in number theory, geometry and quantum fields*, edited by S. Albeverio et al., Aspects Math. **E38**, Vieweg, Wiesbaden, Germany, 2008. MR Zbl

[Seward 2019] B. Seward, "Krieger's finite generator theorem for actions of countable groups, I", *Invent. Math.* **215**:1 (2019), 265–310. MR Zbl

[Sinaĭ 1959] Y. Sinaĭ, "On the concept of entropy for a dynamic system", *Dokl. Akad. Nauk SSSR* **124** (1959), 768–771. In Russian. MR Zbl

[Sinclair 2011] T. Sinclair, "Strong solidity of group factors from lattices in $\mathrm{SO}(n, 1)$ and $\mathrm{SU}(n, 1)$", *J. Funct. Anal.* **260**:11 (2011), 3209–3221. MR Zbl

[Sinclair and Smith 2005] A. M. Sinclair and R. R. Smith, "The Pukánszky invariant for masas in group von Neumann factors", *Illinois J. Math.* **49**:2 (2005), 325–343. MR Zbl

[Voiculescu 1991] D. Voiculescu, "Limit laws for random matrices and free products", *Invent. Math.* **104**:1 (1991), 201–220. MR Zbl

[Voiculescu 1994] D. Voiculescu, "The analogues of entropy and of Fisher's information measure in free probability theory, II", *Invent. Math.* **118**:3 (1994), 411–440. MR Zbl

[Voiculescu 1995] D. Voiculescu, "Free probability theory: random matrices and von Neumann algebras", pp. 227–241 in *Proceedings of the International Congress of Mathematicians*, *I* (Zürich, 1994), edited by S. D. Chatterji, Birkhäuser, Basel, 1995. MR Zbl

[Voiculescu 1996] D. Voiculescu, "The analogues of entropy and of Fisher's information measure in free probability theory, III: The absence of Cartan subalgebras", *Geom. Funct. Anal.* **6**:1 (1996), 172–199. MR Zbl

[Wen 2016] C. Wen, "Maximal amenability and disjointness for the radial masa", *J. Funct. Anal.* **270**:2 (2016), 787–801. MR Zbl

[White 2008] S. White, "Values of the Pukánszky invariant in McDuff factors", *J. Funct. Anal.* **254**:3 (2008), 612–631. MR Zbl

Received 8 Dec 2023. Accepted 20 Jul 2024.

BEN HAYES: brh5c@virginia.edu
Department of Mathematics, University of Virginia, Charlottesville, VA, United States

DAVID JEKEL: daj@math.ku.dk
Department of Mathematics, University of Copenhagen, Copenhagen, Denmark

SRIVATSAV KUNNAWALKAM ELAYAVALLI: skunnawalkamelayaval@ucsd.edu
Department of Mathematics, University of California, San Diego, La Jolla, CA, United States

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at msp.org/apde.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in APDE are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

White space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

ANALYSIS & PDE

Volume 18 No. 7 2025