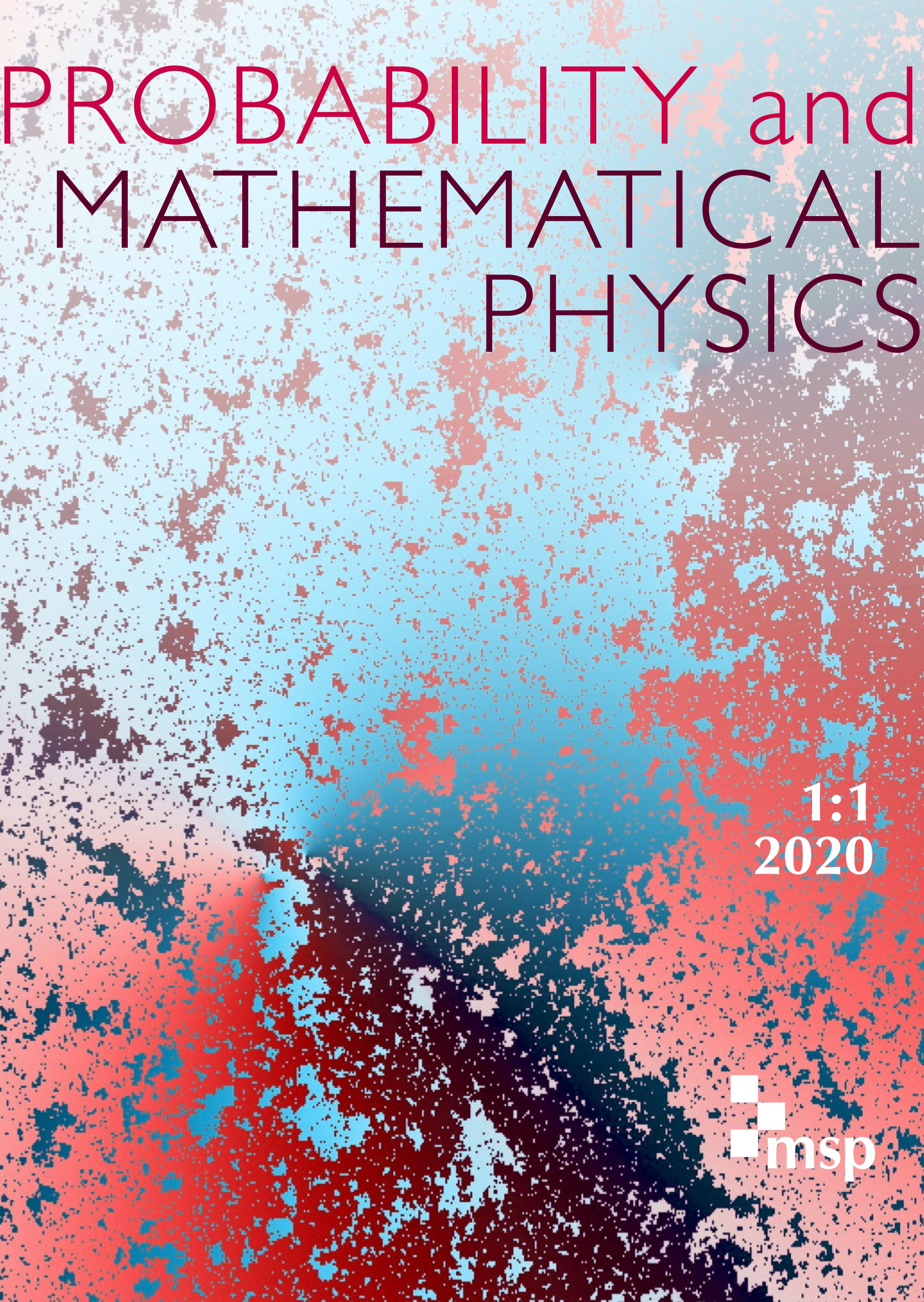


PROBABILITY and MATHEMATICAL PHYSICS

1:1
2020

PROBABILITY and
MATHEMATICAL
PHYSICS msp.org/pmp

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

Probability and Mathematical Physics (ISSN 2690-1005 electronic, 2690-0998 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840 is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

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

PUBLISHED BY
mathematical sciences publishers
nonprofit scientific publishing
https://msp.org/

PROBABILITY and
MATHEMATICAL
PHYSICS

Vol. 1, No. 1, 2020
https://doi.org/10.2140/pmp.2020.1.1

INTRODUCING PMP

We are very pleased to present the first issue of *Probability and Mathematical Physics* (PMP), a new journal devoted to publishing the highest quality papers in all topics of mathematics relevant to physics, with an emphasis on probability and analysis and a particular eye towards developments in their intersection.

The interface between mathematical physics, probability and analysis is indeed an active and exciting topic of current research. We are thinking, for example, of progress on statistical mechanics models and renormalization methods, SPDEs related to mathematical physics, random matrix theory, integrable probability, quantum gravity and SLEs, kinetic theory and the microscopic derivation of macroscopic laws, many-body quantum mechanics, fluid mechanics and turbulence theory, general relativity, semiclassical analysis, Anderson localization, and much else. The journal's scope therefore encompasses the traditional topics of mathematical physics as well as topics in probability and the analysis of PDEs that are most relevant to physics.

In this first issue, we are proud to publish six excellent papers which serve as perfect examples — in terms of quality, scope as well as breadth — of the kind of mathematics we intend to publish in PMP.

Our intention in founding PMP was to create a nonprofit journal published and controlled by the mathematical community that would be at the highest level in its field. Thanks to MSP and the other members of the editorial board, who immediately believed in our project, this vision has been quickly realized.

In the first year since we began accepting submissions, the community has responded very positively and submitted many papers, surpassing our expectations. We currently have a robust pipeline with many promising articles, and we expect to continue to build momentum over the next several years by publishing the most exciting and diverse mathematical research motivated by physics.

ALEXEI BORODIN (borodin@math.mit.edu), Massachusetts Institute of Technology
HUGO DUMINIL-COPIN (duminil@ihes.fr), Institut des Hautes Études Scientifiques
ROBERT SEIRINGER (robert.seiringer@ist.ac.at), IST Austria
SYLVIA SERFATY (serfaty@cims.nyu.edu), Courant Institute of Mathematical Sciences
Editors-in-Chief

PROBABILITY and
MATHEMATICAL
PHYSICS

Vol. 1, No. 1, 2020
https://doi.org/10.2140/pmp.2020.1.3

SHARP SPECTRAL ASYMPTOTICS FOR NONREVERSIBLE METASTABLE DIFFUSION PROCESSES

DORIAN LE PEUTREC AND LAURENT MICHEL

Let $U_h : \mathbb{R}^d \to \mathbb{R}^d$ be a smooth vector field and consider the associated overdamped Langevin equation

$$dX_t = -U_h(X_t)dt + \sqrt{2h}dB_t$$

in the low temperature regime $h \to 0$. In this work, we study the spectrum of the associated diffusion $L = -h\Delta + U_h \cdot \nabla$ under the assumptions that $U_h = U_0 + h\nu$, where the vector fields

$$U_0 : \mathbb{R}^d \to \mathbb{R}^d \quad \text{and} \quad \nu : \mathbb{R}^d \to \mathbb{R}^d$$

are independent of $h \in (0, 1]$, and that the dynamics admits $e^{-\frac{V}{h}}dx$ as an invariant measure for some smooth function $V : \mathbb{R}^d \to \mathbb{R}$. Assuming additionally that V is a Morse function admitting n_0 local minima, we prove that there exists $\epsilon > 0$ such that in the limit $h \to 0$, L admits exactly n_0 eigenvalues in the strip $\{0 \le \mathrm{Re}(z) < \epsilon\}$, which have moreover exponentially small moduli. Under a generic assumption on the potential barriers of the Morse function V, we also prove that the asymptotic behaviors of these small eigenvalues are given by Eyring–Kramers type formulas.

MSC2010: 35P15, 35Q82, 60J60, 81Q12, 81Q20.

Keywords: nonreversible overdamped Langevin dynamics, metastability, spectral theory, semiclassical analysis, Eyring–Kramers formulas.

1. Introduction

1A. ***Setting and motivation.*** Let $U_h : \mathbb{R}^d \to \mathbb{R}^d,\ d \geq 2$, be a smooth vector field depending on a small parameter $h \in (0, 1]$, and consider the associated overdamped Langevin equation

$$dX_t = -U_h(X_t)dt + \sqrt{2h}dB_t, \tag{1-1}$$

where $X_t \in \mathbb{R}^d$ and $(B_t)_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$. The associated Kolmogorov (backward) and Fokker–Planck equations are then the evolution equations

$$\partial_t u + L(u) = 0 \quad \text{and} \quad \partial_t \rho + L^\dagger(\rho) = 0. \tag{1-2}$$

Here, the elliptic differential operator

$$L = -h\Delta + U_h \cdot \nabla$$

is the infinitesimal generator of the process (1-1),

$$L^\dagger = -\operatorname{div} \circ (h\nabla + U_h)$$

denotes the formal adjoint of L, and for $x \in \mathbb{R}^d$ and $t \geq 0,\ u(t, x) = \mathbb{E}^x[f(X_t)]$ is the expected value of the observable $f(X_t)$ when $X_0 = x$ and $\rho(t, \cdot)$ is the probability density (with respect to the Lebesgue measure on $\mathbb{R}^d$) of the presence of $(X_t)_{t\geq 0}$. In this setting, the Fokker–Planck equation, that is, the second equation of (1-2), is also known as the Kramers–Smoluchowski equation.

Throughout this paper, we assume that the vector field U_h decomposes as

$$U_h = U_0 + h\nu$$

for some real smooth vector fields U_0 and ν independent of h. Moreover, we consider the case where the above overdamped Langevin dynamics admit a specific stationary distribution satisfying the following assumption:

Assumption 1. There exists a smooth function $V : \mathbb{R}^d \to \mathbb{R}$ such that $L^\dagger(e^{-V/h}) = 0$ for every $h \in (0, 1]$.

A straightforward computation shows that Assumption 1 is satisfied if and only if the vector field $U_h = U_0 + h\nu$ satisfies the following relations, where we denote $b := U_0 - \nabla V$:

$$b \cdot \nabla V = 0, \quad \operatorname{div}(\nu) = 0, \quad \text{and} \quad \operatorname{div}(b) = \nu \cdot \nabla V. \tag{1-3}$$

Using this decomposition, the generator L writes

$$L_{V,b,\nu} := L = -h\Delta + \nabla V \cdot \nabla + b_h \cdot \nabla, \tag{1-4}$$

where

$$b_h := b + h\nu = U_0 - \nabla V + h\nu = U_h - \nabla V. \tag{1-5}$$

Note moreover that the two following particular cases enter in the framework of Assumption 1:

(1) The case where

$$b \cdot \nabla V = 0, \quad \operatorname{div} b = 0 \quad \text{and} \quad \nu = 0, \tag{1-6}$$

which is in particular satisfied when $\nu = 0$ and b is the matrix product $b = J\nabla V$, where J is a smooth map from $\mathbb{R}^d$ into the set of real antisymmetric matrices of size d such that $\operatorname{div}(J\nabla V)=0$. For instance, this latter condition holds if $J(x)=\tilde{J}\circ V(x)$ for some antisymmetric matrices $\tilde{J}(y)$ depending smoothly on $y\in\mathbb{R}$.

(2) The case where

$$b=J\nabla V \quad \text{and} \quad \nu=\Big(\sum_{i=1}^{d}\partial_i J_{ij}\Big)_{1\le j\le d}, \tag{1-7}$$

where J is a smooth map from $\mathbb{R}^d$ into the set of real antisymmetric matrices of size d.

In the case of (1-7), $L_{V,b,\nu}$ has in particular the supersymmetric-type structure

$$L_{V,b,\nu}=-he^{\frac{V}{h}}\operatorname{div}\circ(e^{-\frac{V}{h}}(I_d-J)\nabla), \tag{1-8}$$

and both cases coincide when b_h has the form $b_h=b=J\nabla V$ for some constant antisymmetric matrix J. In the case of (1-6), the structure (1-8) fails to be true in general and we refer to [Michel 2016] for more details on these questions. Let us also point out that under Assumption 1, the vector field b_h defined in (1-5) is very close to the transverse vector field introduced in [Bouchet and Reygner 2016] and then used in [Landim et al. 2019].

In this paper, we are interested in the spectral analysis of the operator $L_{V,b,\nu}$ and in its connections with the long-time behavior of the dynamics (1-1) when $h\to 0$. In this regime, the process $(X_t)_{t\ge 0}$ solution to (1-1) is typically metastable, which is characterized by a very slow return to equilibrium. We refer especially in this connection to the related works [Bouchet and Reygner 2016; Landim et al. 2019] dealing with the mean transition times between the different wells of the potential V for the process $(X_t)_{t\ge 0}$. Our setting is also motivated by the question of accelerating the convergence to equilibrium, which is of interest for computational purposes. It is known that nongradient perturbations of the overdamped gradient Langevin dynamics

$$dX_t=-\nabla V(X_t)dt+\sqrt{2h}dB_t \tag{1-9}$$

which preserve the invariant measure $e^{-V/h}dx$ cannot converge slower to equilibrium than the associated gradient dynamics (1-9). See [Lelièvre et al. 2013], where the authors considered linear drifts and computed the optimal rate of return to equilibrium in this case.

1B. ***Preliminary analysis.*** In view of Assumption 1, we look at $L_{V,b,\nu}$ acting in the natural weighted Hilbert space $L^2(\mathbb{R}^d,m_h)$, where

$$m_h(dx):=Z_h^{-1}e^{-\frac{V(x)}{h}}dx \quad \text{and} \quad Z_h:=\int_{\mathbb{R}^d}e^{-\frac{V(x)}{h}}dx. \tag{1-10}$$

Note that we assume here that $e^{-V/h}\in L^1(\mathbb{R}^d)$ for every $h\in(0,1]$, which will be a simple consequence of our further hypotheses. In this setting, a first important consequence of (1-3) is the following identity, easily deduced from the relation $\operatorname{div}(b_he^{-V/h})=0$:

$$\forall u,v\in\mathcal{C}_c^\infty(\mathbb{R}^d),\quad \langle L_{V,b,\nu}u,v\rangle_{L^2(m_h)}=\langle u,L_{V,-b,-\nu}v\rangle_{L^2(m_h)}.$$

In particular, using (1-4),

$$\operatorname{Re}\langle L_{V,b,\nu}u, u\rangle_{L^2(m_h)} = \langle(-h\Delta + \nabla V\cdot\nabla)u, u\rangle_{L^2(m_h)} = h\|\nabla u\|^2_{L^2(m_h)} \geq 0 \tag{1-11}$$

for all $u \in \mathcal{C}_c^\infty(\mathbb{R}^d)$, and the operator $L_{V,b,\nu}$ acting on $\mathcal{C}_c^\infty(\mathbb{R}^d)$ in $L^2(\mathbb{R}^d, m_h)$ is hence accretive.

Let us now introduce the confining assumptions at infinity on the functions V, b, and ν that we will consider in the rest of this work.

Assumption 2. There exist $C > 0$ and a compact set $K \subset \mathbb{R}^d$ such that

$$V \geq -C \quad \text{on } \mathbb{R}^d$$

and, for all $x \in \mathbb{R}^d \setminus K$,

$$|\nabla V(x)| \geq \frac{1}{C} \quad \text{and} \quad |\operatorname{Hess} V(x)| \leq C|\nabla V(x)|^2. \tag{1-12}$$

Moreover, there exists $C > 0$ such that the vector fields $b = U_0 - \nabla V$ and ν satisfy the following estimate for all $x \in \mathbb{R}^d$:

$$|b(x)| + |\nu(x)| \leq C(1 + |\nabla V(x)|). \tag{1-13}$$

One can show that when V is bounded from below and the first estimate of (1-12) is satisfied, $V(x) \geq C|x|$ outside a compact set, for some $C > 0$ (see, for example, [Menz and Schlichting 2014, Lemma 3.14]). In particular, when Assumption 2 is satisfied, then $e^{-V/h} \in L^1(\mathbb{R}^d)$ for all $h \in (0, 1]$ (which justifies the definition of Z_h in (1-10)).

In order to study the operator $L_{V,b,\nu}$ in $L^2(\mathbb{R}^d, m_h)$, it is often useful to work with its counterpart in the flat space $L^2(\mathbb{R}^d, dx)$ by using the unitary transformation

$$\mathsf{U} : L^2(\mathbb{R}^d, dx) \to L^2(\mathbb{R}^d, m_h), \quad \mathsf{U}(u) = m_h^{-\frac{1}{2}}u = Z_h^{\frac{1}{2}}e^{\frac{V}{2h}}u.$$

Defining $\phi := V/2$, we then have the unitary equivalence

$$\mathsf{U}^* hL_{V,b,\nu}\mathsf{U} = -h^2\Delta + |\nabla\phi|^2 - h\Delta\phi + b_h \cdot d_{\phi,h} = \Delta_\phi + b_h \cdot d_\phi, \tag{1-14}$$

where

$$d_\phi := d_{\phi,h} := h\nabla + \nabla\phi = he^{-\frac{\phi}{h}}\nabla e^{\frac{\phi}{h}} \tag{1-15}$$

and

$$\Delta_\phi := \Delta_{\phi,h} := -h^2\Delta + |\nabla\phi|^2 - h\Delta\phi = -h^2 e^{\frac{\phi}{h}} \operatorname{div} e^{-\frac{\phi}{h}} d_\phi$$

denotes the usual semiclassical Witten Laplacian acting on functions. It is thus equivalent to study $L_{V,b,\nu}$ acting in the weighted space $L^2(\mathbb{R}^d, m_h)$ or

$$P_\phi := P_{\phi,b,\nu} := \Delta_\phi + b_h \cdot d_\phi \tag{1-16}$$

acting in the flat space $L^2(\mathbb{R}^d, dx)$.

The Witten Laplacian $\Delta_\phi = P_{\phi,0,0}$, which is the counterpart of the weighted Laplacian

$$L_{V,0,0} = -h\Delta + \nabla V \cdot \nabla = h\nabla^*\nabla$$

(the adjoint is considered here with respect to m_h) acting in the flat space $L^2(\mathbb{R}^d, dx)$, is moreover essentially self-adjoint on $\mathcal{C}_c^\infty(\mathbb{R}^n)$ (see [Helffer 2013, Theorem 9.15]). We still denote by Δ_ϕ its unique self-adjoint extension and by $D(\Delta_\phi)$ the domain of this extension. In addition, it is clear that for every $h \in (0,1]$, $\Delta_\phi e^{-\phi/h} = 0$ in the distribution sense. Hence, under Assumption 2, since $\phi = V/2$ satisfies the relation (1-12), $e^{-\phi/h} \in L^2(\mathbb{R}^d)$ and the essential self-adjointness of Δ_ϕ then implies that $e^{-\phi/h} \in D(\Delta_\phi)$ so that $0 \in \operatorname{Ker} \Delta_\phi$. It follows moreover from (1-12) and from [Helffer et al. 2004, Proposition 2.2] that there exists $h_0 > 0$ and $c_0 > 0$ such that for all $h \in (0, h_0]$,

$$\sigma_{\text{ess}}(\Delta_\phi) \subset [c_0, +\infty[.$$

Coming back to the more general operator $P_\phi = P_{\phi,b,\nu}$ defined in (1-16), or equivalently to the operator $L_{V,b,\nu}$ according to the relation (1-14), the following proposition gathers some of its basic properties which specify in particular the preceding properties of Δ_ϕ (and their equivalents concerning the weighted Laplacian $L_{V,0,0}$). It will be proven in Section 2A.

Proposition 1.1. *Under Assumption 1, the operator P_ϕ with domain $\mathcal{C}_c^\infty(\mathbb{R}^d)$ is accretive. Moreover, assuming in addition Assumption 2, there exists $h_0 \in (0,1]$ such that the following hold true for every $h \in (0, h_0]$:*

(i) *The closure of $(P_\phi, \mathcal{C}_c^\infty(\mathbb{R}^d))$, that we still denote by P_ϕ, is maximal accretive, and hence its unique maximal accretive extension.*

(ii) *The operator P_ϕ^* is maximal accretive and $\mathcal{C}_c^\infty(\mathbb{R}^d)$ is a core for P_ϕ^*. We have moreover the inclusions*

$$D(\Delta_\phi) \subset D(P_\phi) \cap D(P_\phi^*) \subset D(P_\phi) \cup D(P_\phi^*) \subset \{u \in L^2(\mathbb{R}^d),\ d_\phi u \in L^2(\mathbb{R}^d)\},$$

where, for any unbounded operator A, $D(A)$ denotes the domain of A. In addition, for $\boldsymbol{P}_\phi \in \{P_\phi, P_\phi^*\}$, we have the equality

$$\forall\, u \in D(\boldsymbol{P}_\phi), \quad \operatorname{Re}\langle \boldsymbol{P}_\phi u, u\rangle = \|d_\phi u\|^2.$$

(iii) *There exists $\Lambda_0 > 0$ such that, defining*

$$\Gamma_{\Lambda_0} := \Big\{\operatorname{Re}(z) \geq 0 \quad \textit{and} \quad |\operatorname{Im} z| \leq \Lambda_0 \max\big(\operatorname{Re}(z), \sqrt{\operatorname{Re}(z)}\,\big)\Big\} \subset \mathbb{C},$$

the spectrum $\sigma(P_\phi)$ of P_ϕ is included in Γ_{Λ_0} and

$$\forall\, z \in \Gamma_{\Lambda_0}^c \cap \{\operatorname{Re}(z) > 0\}, \quad \|(P_\phi - z)^{-1}\|_{L^2 \to L^2} \leq \frac{1}{\operatorname{Re}(z)}.$$

(iv) *There exists $c_1 > 0$ such that the map $z \mapsto (P_\phi - z)^{-1}$ is meromorphic in $\{\operatorname{Re}(z) < c_1\}$ with finite rank residues. In particular, the spectrum of P_ϕ in $\{\operatorname{Re}(z) < c_1\}$ is made of isolated eigenvalues with finite algebraic multiplicities.*

(v) $\operatorname{Ker} P_\phi = \operatorname{Ker} P_\phi^* = \operatorname{Span}\{e^{-\phi/h}\}$ *and 0 is an isolated eigenvalue of P_ϕ (and then of P_ϕ^*) with algebraic multiplicity one.*

From (1-14) and the last item of Proposition 1.1, note that $\operatorname{Ker} L_{V,b,\nu} = \operatorname{Span}\{1\}$ and that 0 is an isolated eigenvalue of $L_{V,b,\nu}$ with algebraic multiplicity one. Moreover, according to Proposition 1.1 and to the Hille–Yosida theorem, the operators $L_{V,b,\nu}$ and its adjoint $L_{V,b,\nu}^*$ (in $L^2(\mathbb{R}^d, m_h)$) generate, for every $h > 0$ small enough, contraction semigroups $(e^{-tL_{V,b,\nu}})_{t\geq 0}$ and $(e^{-tL^*_{V,b,\nu}})_{t\geq 0}$ on $L^2(\mathbb{R}^d, m_h)$ which permit us to solve (1-2).

1C. *Generic Morse-type hypotheses and labeling procedure.* In order to describe precisely, in particular by stating Eyring–Kramers type formulas, the spectrum around 0 of $L_{V,b,\nu}$ (or equivalently of P_ϕ) in the regime $h \to 0$, we will assume from now on that V is a Morse function:

Assumption 3. The function V is a Morse function.

Under Assumption 3 and thanks to Assumption 2, the set $\mathcal{U}$ made of the critical points of V is finite. In the following, the critical points of V with index 0 and with index 1, that is, its local minima and its saddle points, will play a fundamental role, and we will respectively denote by $\mathcal{U}^{(0)}$ and $\mathcal{U}^{(1)}$ the sets made of these points. Throughout the paper, we will moreover denote

$$n_0 := \operatorname{card}(\mathcal{U}^{(0)}).$$

From the pioneer work by Witten [1982], it is well-known that for every $h \in (0, 1]$ small enough, there is a correspondence between the small eigenvalues of Δ_ϕ and the local minima of $\phi = V/2$. More precisely, we have the following result (see [Helffer and Sjöstrand 1985; Helffer 1988; Helffer et al. 2004; Michel and Zworski 2018]).

Proposition 1.2. *Assume that (1-12) and Assumption 3 hold true. Then, there exist $\epsilon_0 > 0$ and $h_0 > 0$ such that for every $h \in (0, h_0]$, Δ_ϕ has precisely n_0 eigenvalues (counted with multiplicity) in the interval $[0, \epsilon_0 h]$. Moreover, these eigenvalues are actually exponentially small, that is, they live in an interval $[0, Che^{-2\frac{S}{h}}]$ for some $C, S > 0$ independent of $h \in (0, h_0]$.*

Since the operator $P_\phi = \Delta_\phi + b_h \cdot d_\phi$ is not self-adjoint (when $b_h \neq 0$), the analysis of its spectrum is more complicated than that of the spectrum of Δ_ϕ. The following result states a counterpart of Proposition 1.2 in this setting. In this statement and in the sequel, for any $a \in \mathbb{C}$ and $r > 0$, we will denote by $D(a, r) \subset \mathbb{C}$ the open disk of center a and radius r.

Theorem 1.3. *Assume that Assumptions 1–3 hold true, and let $\epsilon_0 > 0$ be given by Proposition 1.2. Then, for every $\epsilon_1 \in (0, \epsilon_0)$, there exists $h_0 > 0$ such that for all $h \in (0, h_0]$, the set $\sigma(P_\phi) \cap \{\operatorname{Re} z < \epsilon_1 h\}$ is finite and consists of*

$$n_0 = \operatorname{card}(\sigma(\Delta_\phi) \cap \{\operatorname{Re} z < \epsilon_0 h\})$$

eigenvalues counted with algebraic multiplicity. Moreover, there exists $C > 0$ such that for all $h \in (0, h_0]$,

$$\sigma(P_\phi) \cap \{\operatorname{Re} z < \epsilon_1 h\} \subset D(0, Ch^{\frac{1}{2}} e^{-\frac{S}{h}}),$$

where S is given by Proposition 1.2. In addition, for every $\epsilon \in (0, \epsilon_1)$, one has, uniformly with respect to z,

$$\forall z \in \{\operatorname{Re} z < \epsilon_1 h\} \cap \{|z| > \epsilon h\}, \quad \|(P_\phi - z)^{-1}\|_{L^2 \to L^2} = \mathcal{O}(h^{-1}).$$

Lastly, all the above conclusions also hold for P^*_ϕ.

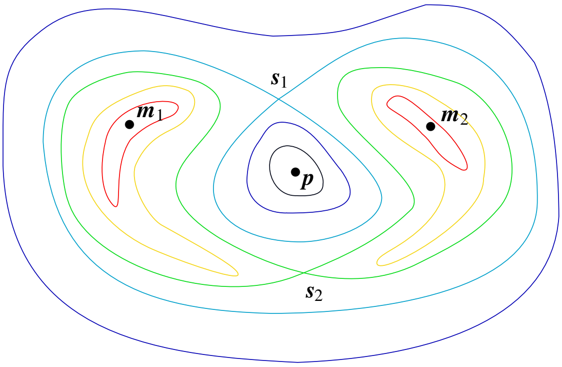


Figure 1. Some level sets of a Morse function V such that $V(x) \to +\infty$ when $|x| \to +\infty$ and admitting five critical points: two local minima $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$, one local maximum $\boldsymbol{p}$, and two saddle points $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. The point $\boldsymbol{s}_1$ is nonseparating, whereas $\boldsymbol{s}_2$ is separating.

This theorem will be proved in Section 2B using Proposition 1.2 and a finite-dimensional reduction. In order to give sharp asymptotics of the small eigenvalues of P_ϕ, that is, the ones in $D(0, Ch^{1/2}e^{-S/h})$, we will introduce some additional, but generic, topological assumptions on the Morse function V (see Assumption 4 below). To this end, we first recall the general labeling of [Hérau et al. 2011] (see, in particular, Definition 4.1 there) generalizing the labeling of [Helffer et al. 2004] (and of [Bovier et al. 2004; 2005]). The main ingredient is the notion of separating saddle point, from Definition 1.5 (see also an illustration in Figure 1), after the following observation. Here and in the sequel, we define, for $a \in \mathbb{R}$,

$$\{V < a\} := V^{-1}((-\infty, a)) \quad \text{and} \quad \{V \leq a\} := V^{-1}((-\infty, a]),$$

and $\{V > a\}$ and $\{V \geq a\}$ in a similar way. The following lemma recalls the local structure of the sublevel sets of a Morse function. A proof can be found in [Helffer et al. 2004].

Lemma 1.4. *Let $z \in \mathbb{R}^d$ and $V : \mathbb{R}^d \to \mathbb{R}$ be a Morse function. For any $r > 0$, we denote by $B(z, r) \subset \mathbb{R}^d$ the open ball of center z and radius r. Then, for every $r > 0$ small enough, $B(z, r) \cap \{V < V(z)\}$ has at least two connected components if and only if z is a saddle point of V, i.e., if and only if $z \in \mathcal{U}^{(1)}$. In this case, $B(z, r) \cap \{V < V(z)\}$ has precisely two connected components.*

Definition 1.5. (i) We say that the saddle point $\boldsymbol{s} \in \mathcal{U}^{(1)}$ is a separating saddle point of V if, for every $r > 0$ small enough, the two connected components of $B(\boldsymbol{s}, r) \cap \{V < V(\boldsymbol{s})\}$ (see Lemma 1.4) are contained in different connected components of $\{V < V(\boldsymbol{s})\}$. We will denote by $\mathcal{V}^{(1)}$ the set made of these points.

(ii) We say that $\sigma \in \mathbb{R}$ is a separating saddle value of V if it has the form $\sigma = V(\boldsymbol{s})$ for some $\boldsymbol{s} \in \mathcal{V}^{(1)}$.

(iii) Moreover, we say that $E \subset \mathbb{R}^d$ is a critical component of V if there exists $\sigma \in V(\mathcal{V}^{(1)})$ such that E is a connected component of $\{V < \sigma\}$ satisfying $\partial E \cap \mathcal{V}^{(1)} \neq \varnothing$.

Let us now describe the general labeling procedure of [Hérau et al. 2011]. We will omit details when associating local minima and separating saddle points below, but the following proposition (see [Di Gesù et al. 2020, Proposition 18]) may be helpful to understand the construction.

Proposition 1.6. *Assume that V is a Morse function with a finite number of critical points and such that $V(x) \to +\infty$ when $|x| \to +\infty$. Let $\lambda \in \mathbb{R}$ and let $\mathcal{C}$ be a connected component of $\{V < \lambda\}$. Then,*

$$\mathcal{C} \cap \mathcal{V}^{(1)} \neq \varnothing \iff \operatorname{card}(\mathcal{C} \cap \mathcal{U}^{(0)}) \geq 2.$$

Let us also define

$$\sigma := \max_{\mathcal{C} \cap \mathcal{V}^{(1)}} V$$

with the convention $\sigma := \min_{\mathcal{C}} V$ when $\mathcal{C} \cap \mathcal{V}^{(1)} = \varnothing$. Then:

(i) *For every $\mu \in (\sigma, \lambda]$, the set $\mathcal{C} \cap \{V < \mu\}$ is a connected component of $\{V < \mu\}$.*

(ii) *If $\mathcal{C} \cap \mathcal{V}^{(1)} \neq \varnothing$, then $\mathcal{C} \cap \mathcal{U}^{(0)} \subset \{V < \sigma\}$ and all the connected components of $\mathcal{C} \cap \{V < \sigma\}$ are critical.*

Under the hypotheses of Proposition 1.6, $V(\mathcal{V}^{(1)})$ is finite. We moreover assume that $n_0 \geq 2$, so that, under the hypotheses of Proposition 1.1 and of Theorem 1.3, 0 is not the only exponentially small eigenvalue of P_ϕ (or equivalently of $L_{V,b,\nu}$) and $\mathcal{V}^{(1)} \neq \varnothing$ by Proposition 1.6. We then denote the elements of $V(\mathcal{V}^{(1)})$ by $\sigma_2 > \sigma_3 > \cdots > \sigma_N$, where $N \geq 2$. For convenience, we also introduce a fictive infinite saddle value $\sigma_1 = +\infty$. Starting from σ_1, we will recursively associate to each σ_i a finite family of local minima $(\boldsymbol{m}_{i,j})_j$ and a finite family of critical components $(E_{i,j})_j$ (see Definition 1.5).

Let $N_1 := 1$, $\underline{\boldsymbol{m}} = \boldsymbol{m}_{1,1}$ be a global minimum of V (arbitrarily chosen if there are more than one), and $E_{1,1} := \mathbb{R}^d$. We now proceed in the following way:

- Let us denote, for some $N_2 \geq 1$, by $E_{2,1}, \ldots, E_{2,N_2}$ the connected components of $\{V < \sigma_2\}$ which do not contain $\boldsymbol{m}_{1,1}$. They are all critical by the preceding proposition and we associate to each $E_{2,j}$, where $j \in \{1, \ldots, N_2\}$, some global minimum $\boldsymbol{m}_{2,j}$ of $V|_{E_{2,j}}$ (arbitrarily chosen if there are more than one).
- Let us then consider, for some $N_3 \geq 1$, the connected components $E_{3,1}, \ldots, E_{3,N_3}$ of $\{V < \sigma_3\}$ which do not contain the local minima of V previously labeled. These components are also critical and included in the $E_{2,j} \cap \{V < \sigma_3\}$'s, $j \in \{1, \ldots, N_2\}$, such that $E_{2,j} \cap \{V = \sigma_3\} \cap \mathcal{V}^{(1)} \neq \varnothing$ (and $\sigma_3 = \max_{E_{2,j} \cap \mathcal{V}^{(1)}} V$ for such a j). We then again associate to each $E_{3,j}$, $j \in \{1, \ldots, N_3\}$, some global minimum $\boldsymbol{m}_{3,j}$ of $V|_{E_{3,j}}$.
- We continue this process until having considered the connected components of $\{V < \sigma_N\}$, after which all the local minima of V have been labeled.

Next, we define two mappings

$$E : \mathcal{U}^{(0)} \to \mathcal{P}(\mathbb{R}^d) \quad \text{and} \quad \boldsymbol{j} : \mathcal{U}^{(0)} \to \mathcal{P}(\mathcal{V}^{(1)} \cup \{\boldsymbol{s}_1\}),$$

where, for any set A, $\mathcal{P}(A)$ denotes the power set of A, and $\boldsymbol{s}_1$ is a fictive saddle point such that $V(\boldsymbol{s}_1) = \sigma_1 = +\infty$, as follows: for every $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, N_i\}$,

$$E(\boldsymbol{m}_{i,j}) := E_{i,j} \tag{1-17}$$

and

$$\boldsymbol{j}(\underline{\boldsymbol{m}}) := \{\boldsymbol{s}_1\} \quad \text{and, when } i \geq 2, \quad \boldsymbol{j}(\boldsymbol{m}_{i,j}) := \partial E_{i,j} \cap \mathcal{V}^{(1)} \neq \varnothing. \tag{1-18}$$

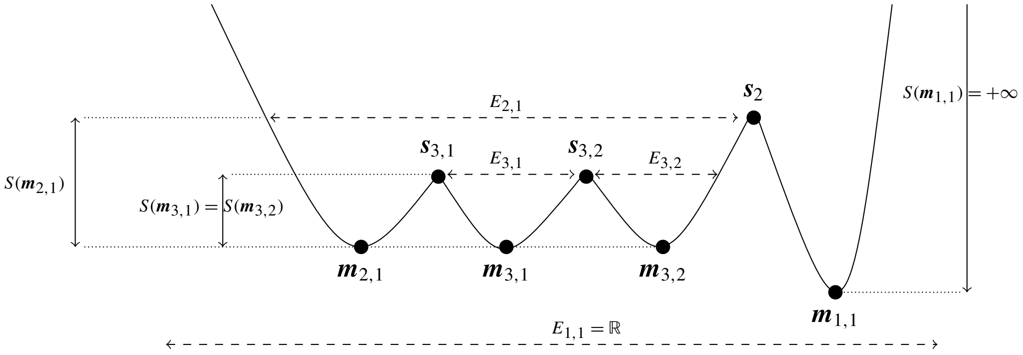


Figure 2. A 1-D labeling example when V admits four local minima. In this example, $V(\boldsymbol{m}_{1,1}) < V(\boldsymbol{m}_{2,1}) = V(\boldsymbol{m}_{3,1}) = V(\boldsymbol{m}_{3,2})$, $\boldsymbol{j}(\boldsymbol{m}_{2,1}) = \{\boldsymbol{s}_2\}$, $\boldsymbol{j}(\boldsymbol{m}_{3,1}) = \{\boldsymbol{s}_{3,1}, \boldsymbol{s}_{3,2}\}$, and $\boldsymbol{j}(\boldsymbol{m}_{3,2}) = \{\boldsymbol{s}_{3,2}\}$. Note that other choices of construction of the maps $\boldsymbol{j}$ and E are possible here since $\operatorname{argmin}_{E_{2,1}} V = \{\boldsymbol{m}_{2,1}, \boldsymbol{m}_{3,1}, \boldsymbol{m}_{3,2}\}$.

In particular, $E(\underline{\boldsymbol{m}}) = \mathbb{R}^d$ and

$$\forall\, i \in \{1, \dots, N\},\ \forall\, j \in \{1, \dots, N_i\}, \quad \varnothing \neq \boldsymbol{j}(\boldsymbol{m}_{i,j}) \subset \{V = \sigma_i\}.$$

Lastly, we define the mappings

$$\boldsymbol{\sigma} : \mathcal{U}^{(0)} \to V(\mathcal{V}^{(1)}) \cup \{\sigma_1\} \quad \text{and} \quad S : \mathcal{U}^{(0)} \to (0, +\infty]$$

by

$$\forall\, \boldsymbol{m} \in \mathcal{U}^{(0)}, \quad \boldsymbol{\sigma}(\boldsymbol{m}) := V(\boldsymbol{j}(\boldsymbol{m})) \quad \text{and} \quad S(\boldsymbol{m}) := \boldsymbol{\sigma}(\boldsymbol{m}) - V(\boldsymbol{m}), \tag{1-19}$$

where, with a slight abuse of notation, we have identified the set $V(\boldsymbol{j}(\boldsymbol{m}))$ with its unique element. Note that $S(\boldsymbol{m}) = +\infty$ if and only if $\boldsymbol{m} = \underline{\boldsymbol{m}}$. An example of the preceding labeling is given in Figure 2.

Our generic topological assumption is the following one. Assume that V is a Morse function with a finite number $n_0 \geq 2$ of critical points such that $V(x) \to +\infty$ when $|x| \to +\infty$, and let $E : \mathcal{U}^{(0)} \to \mathcal{P}(\mathbb{R}^d)$ and $\boldsymbol{j} : \mathcal{U}^{(0)} \to \mathcal{P}(\mathcal{V}^{(1)} \cup \{\boldsymbol{s}_1\})$ be the mappings defined in (1-17) and in (1-18).

Assumption 4. For every $\boldsymbol{m} \in \mathcal{U}^{(0)}$, the following hold true:

(i) The local minimum $\boldsymbol{m}$ is the unique global minimum of $V|_{E(\boldsymbol{m})}$,

(ii) For all $\boldsymbol{m}' \in \mathcal{U}^{(0)} \setminus \{\boldsymbol{m}\}$, $\boldsymbol{j}(\boldsymbol{m}) \cap \boldsymbol{j}(\boldsymbol{m}') = \varnothing$.

In particular, V uniquely attains its global minimum, at $\underline{\boldsymbol{m}} \in \mathcal{U}^{(0)}$.

Note that the example of Figure 2 does not satisfy Assumption 4 since neither item (i) nor (ii) holds there. See also Figure 3 for a similar example satisfying Assumption 4.

Let us moreover underline that this assumption is a little more general than the one considered in the generic case in [Helffer et al. 2004; Hérau et al. 2011] (see also [Bovier et al. 2004; 2005]) where, for instance, each set $\boldsymbol{j}(\boldsymbol{m})$, $\boldsymbol{m} \in \mathcal{U}^{(0)} \setminus \{\underline{\boldsymbol{m}}\}$, is assumed to only contain one element.

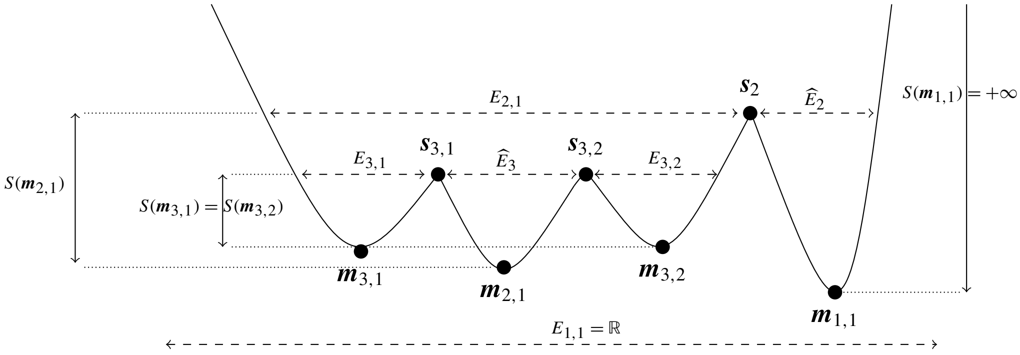


Figure 3. A 1-D example when V admits four local minima and satisfies Assumption 4. Here, $V(\boldsymbol{m}_{1,1}) < V(\boldsymbol{m}_{2,1}) < V(\boldsymbol{m}_{3,1}) = V(\boldsymbol{m}_{3,2})$, $\boldsymbol{j}(\boldsymbol{m}_{2,1}) = \{\boldsymbol{s}_2\}$, $\boldsymbol{j}(\boldsymbol{m}_{3,1}) = \{\boldsymbol{s}_{3,1}\}$, and $\boldsymbol{j}(m_{3,2}) = \{\boldsymbol{s}_{3,2}\}$. Moreover, $\widehat{E}_2$ and $\widehat{E}_3$ denote respectively the sets $\widehat{E}(\boldsymbol{m}_{2,1})$ and $\widehat{E}(\boldsymbol{m}_{3,1}) = \widehat{E}(\boldsymbol{m}_{3,2})$ introduced in Remark 1.7.

Remark 1.7. One can also show that Assumption 4 implies that for every $\boldsymbol{m} \in \mathcal{U}^{(0)}$ such that $\boldsymbol{m} \neq \underline{\boldsymbol{m}}$, there is precisely one connected component $\widehat{E}(\boldsymbol{m}) \neq E(\boldsymbol{m})$ of $\{f < \boldsymbol{\sigma}(\boldsymbol{m})\}$ such that $\widehat{E}(\boldsymbol{m}) \cap \overline{E(\boldsymbol{m})} \neq \varnothing$. In other words, there exists a connected component $\widehat{E}(\boldsymbol{m}) \neq E(\boldsymbol{m})$ of $\{f < \boldsymbol{\sigma}(\boldsymbol{m})\}$ such that $\boldsymbol{j}(\boldsymbol{m}) \subset \partial \widehat{E}(\boldsymbol{m})$. Moreover, the global minimum $\boldsymbol{m}'$ of $V|_{\widehat{E}(\boldsymbol{m})}$ is unique and satisfies $\boldsymbol{\sigma}(\boldsymbol{m}') > \boldsymbol{\sigma}(\boldsymbol{m})$ and $V(\boldsymbol{m}') < V(\boldsymbol{m})$ (see examples of such sets in Figure 3). We refer to [Michel 2019] or [Di Gesù et al. 2020] for more details on the geometry of the sublevel sets of a Morse function.

1D. *Main results and comments.*

In order to state our main results, we also need the following lemma which is fundamental in our analysis.

Lemma 1.8. *For $x \in \mathbb{R}^d$, let $B(x) := \mathrm{Jac}_x b$ denote the Jacobian matrix of $b = U_0 - \nabla V$ at x, and consider a saddle point $\boldsymbol{s} \in \mathcal{U}^{(1)}$.*

(i) *The matrix* $\operatorname{Hess} V(\boldsymbol{s}) + B^*(\boldsymbol{s}) \in \mathcal{M}_d(\mathbb{R})$ *admits precisely one negative eigenvalue $\mu = \mu(\boldsymbol{s})$, which has moreover geometric multiplicity one.*

(ii) *Denote by $\xi = \xi(\boldsymbol{s})$ one of the two (real) unitary eigenvectors of* $\operatorname{Hess} V(\boldsymbol{s}) + B^*(\boldsymbol{s})$ *associated with μ. The real symmetric matrix*

$$M_V := \operatorname{Hess} V(\boldsymbol{s}) + 2|\mu|\xi\xi^*$$

is then positive definite and its determinant satisfies

$$\det M_V = -\det \operatorname{Hess} V(\boldsymbol{s}).$$

(iii) *Lastly, denoting by $\lambda_1 = \lambda_1(\boldsymbol{s})$ the negative eigenvalue of* $\operatorname{Hess} V(\boldsymbol{s})$, $|\mu| \geq |\lambda_1|$, *with equality if and only if $B^*(\boldsymbol{s})\xi = 0$, and*

$$\langle (\operatorname{Hess} V(\boldsymbol{s}))^{-1}\xi, \xi \rangle = \frac{1}{\mu} < 0.$$

Note that the real matrix $\operatorname{Hess} V(\boldsymbol{s}) + B^*(\boldsymbol{s})$ of Lemma 1.8 is in general nonsymmetric. Let us also point out that the statements of Lemma 1.8 already appeared in the related work [Landim et al. 2019] (see in particular the beginning of Section 8 there), and in [Landim and Seo 2018], where proofs are given (see Section 4.1 there). We will nevertheless give a proof in Section 3 for the sake of completeness.

We can now state our main results.

Theorem 1.9. *Suppose that Assumptions 1–4 hold true, and let* $\epsilon_0 > 0$ *be given by Proposition 1.2. Then, for all* $\epsilon_1 \in (0, \epsilon_0)$, *there exists* $h_0 > 0$ *such that for all* $h \in (0, h_0]$, *one has, counting the eigenvalues with algebraic multiplicity,*

$$\sigma(L_{V,b,\nu}) \cap \{\mathrm{Re} z < \epsilon_1\} = \{\lambda(\boldsymbol{m}, h),\ \boldsymbol{m} \in \mathcal{U}^{(0)}\},$$

where, denoting by $\underline{\boldsymbol{m}}$ *the unique absolute minimum of* V, $\lambda(\underline{\boldsymbol{m}}, h) = 0$ *and, for all* $\boldsymbol{m} \neq \underline{\boldsymbol{m}}$, $\lambda(\boldsymbol{m}, h)$ *satisfies the Eyring–Kramers type formula*

$$\lambda(\boldsymbol{m}, h) = \zeta(\boldsymbol{m}) e^{-\frac{S(\boldsymbol{m})}{h}} \big(1 + \mathcal{O}\big(\sqrt{h}\,\big)\big), \tag{1-20}$$

where $S : \mathcal{U}^{(0)} \to (0, +\infty]$ *is defined in (1-19) and, for every* $\boldsymbol{m} \in \mathcal{U}^{(0)} \setminus \{\underline{\boldsymbol{m}}\}$,

$$\zeta(\boldsymbol{m}) := \frac{\det \mathrm{Hess}\, V(\boldsymbol{m})^{\frac{1}{2}}}{2\pi} \Bigg(\sum_{\boldsymbol{s} \in \boldsymbol{j}(\boldsymbol{m})} \frac{|\mu(\boldsymbol{s})|}{|\det \mathrm{Hess}\, V(\boldsymbol{s})|^{\frac{1}{2}}} \Bigg), \tag{1-21}$$

where $\boldsymbol{j} : \mathcal{U}^{(0)} \to \mathcal{P}(\mathcal{V}^{(1)} \cup \{\boldsymbol{s}_1\})$ *is defined in (1-18) and the* $\mu(\boldsymbol{s})$*'s are defined in Lemma 1.8. In addition,*

$$\sigma(L_{V,-b,-\nu}) \cap \{\mathrm{Re} z < \epsilon_1\} = \sigma(L^*_{V,b,\nu}) \cap \{\mathrm{Re} z < \epsilon_1\} = \{\overline{\lambda(\boldsymbol{m}, h)},\ \boldsymbol{m} \in \mathcal{U}^{(0)}\}.$$

Remark 1.10. In the case where V has precisely two minima $\underline{\boldsymbol{m}}$ and $\boldsymbol{m}$ such that $V(\underline{\boldsymbol{m}}) = V(\boldsymbol{m})$, the above result can be easily generalized. In this case, using the definitions of S and $\boldsymbol{j}$ given in (1-19) and in (1-18) (note that the choice of $\underline{\boldsymbol{m}}$ among the two minima of V is arbitrary in this case), we have, counting the eigenvalues with algebraic multiplicity, for every $h > 0$ small enough,

$$\sigma(L_{V,b,\nu}) \cap \{\mathrm{Re} z < \epsilon_1\} = \{0, \lambda(\boldsymbol{m}, h)\},$$

where

$$\lambda(\boldsymbol{m}, h) = \zeta(\boldsymbol{m}) e^{-\frac{S(\boldsymbol{m})}{h}} \big(1 + \mathcal{O}\big(\sqrt{h}\,\big)\big)$$

with

$$\zeta(\boldsymbol{m}) = \frac{\det \mathrm{Hess}\, V(\boldsymbol{m})^{\frac{1}{2}} + \det \mathrm{Hess}\, V(\underline{\boldsymbol{m}})^{\frac{1}{2}}}{2\pi} \Bigg(\sum_{\boldsymbol{s} \in \boldsymbol{j}(\boldsymbol{m})} \frac{|\mu(\boldsymbol{s})|}{|\det \mathrm{Hess}\, V(\boldsymbol{s})|^{\frac{1}{2}}} \Bigg).$$

Moreover, since $\sigma(L_{V,b,\nu}) = \overline{\sigma(L_{V,b,\nu})}$, the eigenvalue $\lambda(\boldsymbol{m}, h)$ is real.

Let us make a few comments on Theorem 1.9.

First, observe that if we assume that $U_h = \nabla V$, that is, that $b_h = 0$ (see (1-5)), we obtain the precise asymptotics of the small eigenvalues of $L_{V,0,0}$ (or equivalently of Δ_ϕ after multiplication by $1/h$; see (1-14)) and hence recover the results already proved in this reversible setting in [Bovier et al. 2005; Helffer et al. 2004] (see also [Menz and Schlichting 2014] for an extension to logarithmic Sobolev inequalities). In this case, for every saddle point $\boldsymbol{s}$ appearing in (1-21), the real number $\mu(\boldsymbol{s})$ is indeed the negative eigenvalue of $\mathrm{Hess}\, V(\boldsymbol{s})$ according to the first item of Lemma 1.8. We also point out that under the hypotheses made in [Bovier et al. 2005; Helffer et al. 2004], the set $\boldsymbol{j}(\boldsymbol{m})$ actually contains one unique element for every $\boldsymbol{m} \in \mathcal{U}^{(0)} \setminus \{\underline{\boldsymbol{m}}\}$. Moreover, our analysis permits us in this case to recover that the error term $\mathcal{O}(\sqrt{h})$ is actually of order $\mathcal{O}(h)$, as proven in [Helffer et al. 2004]. However, it does not permit us to prove that this $\mathcal{O}(h)$ actually admits a full asymptotic expansion in h as proven in [Helffer et al. 2004].

To the best of our knowledge, the above theorem is the first result giving sharp asymptotics of the small eigenvalues of the generator $L_{V,b,\nu}$ in the nonreversible case. Similar results were obtained by Hérau, Hitrik and Sjöstrand for the Kramers–Fokker–Planck (KFP) equation in [Hérau et al. 2011]. They deal with non-self-adjoint and nonelliptic operators, which makes the analysis more complicated than in our framework. However, the KFP equation enjoys several symmetries which are crucial in their analysis. First of all, the KFP operator has a supersymmetric structure (for a nonsymmetric skew-product $\langle\cdot,\cdot\rangle_{\mathrm{KFP}}$) which permits them to write the interaction matrix associated with the small eigenvalues as a square $M = A^*A$, where the adjoint A^* is taken with respect to $\langle\cdot,\cdot\rangle_{\mathrm{KFP}}$. Using this square structure, the authors can then follow the strategy of [Helffer et al. 2004] to construct accurate approximations of the matrices A and A^*. However, since $\langle\cdot,\cdot\rangle_{\mathrm{KFP}}$ is not a scalar product, they cannot identify the squares of the singular values of A with the eigenvalues of M. This difficulty is solved by using an extra symmetry (the PT-symmetry), which permits the modification of the skew-product $\langle\cdot,\cdot\rangle_{\mathrm{KFP}}$ into a new product $\langle\cdot,\cdot\rangle_{\mathrm{KFPS}}$, which is a scalar product when restricted to the “small spectral subspace”, and for which the identity $M = A^*A$ remains true with an adjoint taken with respect to $\langle\cdot,\cdot\rangle_{\mathrm{KFPS}}$. This permits a conclusion as in [Helffer et al. 2004], using in particular the Fan inequalities to estimate the singular values of A.

In the present case, neither of these two symmetries are available in general ($L_{V,b,\nu}$, or equivalently P_ϕ, enjoys a supersymmetric structure when b and ν satisfy the relation (1-7) however; see (1-8) or Remark 3.2). We developed an alternative approach based on the construction of very accurate quasimodes and partly inspired by [Di Gesù and Le Peutrec 2017] (see also the related constructions made in [Bovier et al. 2004; Landim et al. 2019; Le Peutrec and Nectoux 2019]). This permits the construction of the interaction matrix M as above. However, since we cannot write $M = A^*A$ and use the Fan inequalities as in [Helffer et al. 2004; Hérau et al. 2011] (and, e.g., in [Helffer and Nier 2006; Le Peutrec 2010; Michel 2019; Di Gesù et al. 2020; Le Peutrec and Nectoux 2019]), we have to compute directly the eigenvalues of M. To this end, we use crucially the Schur complement method. This leads to Theorem A.4 in the Appendix, which permits us to replace the use of the Fan inequalities to perform the final analysis in our setting. We believe that these two arguments are quite general and may be used in other contexts.

Though it is generic, one may ask if Assumption 4 is necessary to get Eyring–Kramers type formulas as in Theorem 1.9. In the reversible setting, the full general (Morse) case was recently treated in [Michel 2019], but applying the methods developed there to our nonreversible setting was not straightforward and we decided to postpone this analysis to future works. In connection with this, let us point out that in the general (Morse) case, some tunneling effect between the characteristic wells of V defined by the mapping E (see (1-17)) mixes their corresponding prefactors; see Remark 1.10, or [Michel 2019] for more intricate situations in the reversible setting.

Note that Theorem 1.9 does not state that the operator $L_{V,b,\nu}$ is diagonalizable when restricted to the spectral subspace associated with its small eigenvalues. Indeed, since $L_{V,b,\nu}$ is not self-adjoint, we cannot exclude the existence of Jordan's blocks. We cannot exclude the existence of nonreal eigenvalues either, but the spectrum of $L_{V,b,\nu}$ is obviously stable by complex conjugation since $L_{V,b,\nu}$ is a partial differential operator with real coefficients. However, in the case where for every $\boldsymbol{m} \in \mathcal{U}^{(0)} \setminus \{\underline{\boldsymbol{m}}\}$, the prefactors $\zeta(\boldsymbol{m}')$ defined in (1-21) are all distinct for $\boldsymbol{m}' \in S^{-1}(S(\boldsymbol{m}))$, the $\lambda(\boldsymbol{m}, h)$'s, $\boldsymbol{m} \in \mathcal{U}^{(0)}$, are then real eigenvalues of multiplicity one of $L_{V,b,\nu}$, and its restriction to its small spectral subspace is diagonalizable.

Coming back to the contraction semigroups $(e^{-tL_{V,b,\nu}})_{t\geq 0}$ and $(e^{-tL^*_{V,b,\nu}})_{t\geq 0}$ on $L^2(\mathbb{R}^d, m_h)$ introduced just after Proposition 1.1, Theorem 1.9 has the following consequences on the rate of convergence to equilibrium for the process (1-1).

Theorem 1.11. *Assume that the hypotheses of Theorem 1.9 hold and let $\boldsymbol{m}^* \in \mathcal{U}^{(0)} \setminus \{\underline{\boldsymbol{m}}\}$ be such that*

$$S(\boldsymbol{m}^*) = \max_{\boldsymbol{m}\in\mathcal{U}^{(0)}} S(\boldsymbol{m}) \quad \text{and} \quad \zeta(\boldsymbol{m}^*) = \min_{\boldsymbol{m}\in S^{-1}(S(\boldsymbol{m}^*))} \zeta(\boldsymbol{m}), \tag{1-22}$$

where the prefactors $\zeta(\boldsymbol{m})$, $\boldsymbol{m} \in \mathcal{U}^{(0)} \setminus \{\underline{\boldsymbol{m}}\}$, are defined in (1-21), and $S : \mathcal{U}^{(0)} \to (0, +\infty]$ is defined in (1-19). Let us then define, for any $h > 0$,

$$\lambda(h) := \zeta(\boldsymbol{m}^*)e^{-\frac{S(\boldsymbol{m}^*)}{h}}.$$

Then, there exist $h_0 > 0$ and $C > 0$ such that for every $h \in (0, h_0]$,

$$\forall\, t \geq 0, \quad \|e^{-tL_{V,b,\nu}} - \Pi_0\|_{L^2(m_h)\to L^2(m_h)} \leq Ce^{-\lambda(h)(1-C\sqrt{h})t}, \tag{1-23}$$

where Π_0 denotes the orthogonal projector on $\operatorname{Ker} L_{V,b,\nu} = \operatorname{Span}\{1\}$:

$$\forall\, u \in L^2(m_h), \qquad \Pi_0 u = \langle u, 1\rangle_{L^2(m_h)} = \int_{\mathbb{R}^d} u\, dm_h.$$

Assume moreover that $(X_t)_{t\geq 0}$ is a solution to (1-1) and that the probability distribution ϱ_0 of X_0 admits a density $\mu_0 \in L^2(\mathbb{R}^d, m_h)$ with respect to the probability measure m_h. Then, for every $t \geq 0$, the probability distribution ϱ_t of X_t admits the density $\mu_t = e^{-tL^*_{V,b,\nu}}\mu_0 \in L^2(\mathbb{R}^d, m_h)$ with respect to m_h, and for every $h \in (0, h_0]$,

$$\forall\, t \geq 0, \quad \|\varrho_t - \nu_h\|_{TV} \leq C\|\mu_0 - 1\|_{L^2(m_h)}e^{-\lambda(h)(1-C\sqrt{h})t}, \tag{1-24}$$

where $\|\cdot\|_{TV}$ denotes the total variation distance.

Finally, when there exists one unique $\boldsymbol{m}^*$ satisfying (1-22), the eigenvalue $\lambda(\boldsymbol{m}^*, h)$ associated with $\boldsymbol{m}^*$ (see (1-20)) is real and simple, and the estimates (1-23) and (1-24) hold if one replaces $\lambda(h)(1 - C\sqrt{h})$ by $\lambda(\boldsymbol{m}^*, h)$ in the exponential terms.

Theorems 1.9 and 1.11 describe the metastable behavior of the dynamics (1-1) from a spectral perspective.

Concerning the question of accelerating the convergence to equilibrium mentioned at the end of Section 1A, the exponential rate of convergence to equilibrium appearing in the estimates (1-23) and (1-24) is generically strictly larger than the optimal rate for the associated gradient dynamics (1-9). To be more precise, let us assume, as in the last part of the statement of Theorem 1.11, that there exists one unique $\boldsymbol{m}^*$ satisfying (1-22). The exponential rate of return to equilibrium appearing in (1-23) and (1-24) is then given by the spectral gap $\lambda(\boldsymbol{m}^*, h)$ of $L_{V,b,\nu}$. Moreover, denoting by $\lambda^\nabla(\boldsymbol{m}^*, h)$ the spectral gap of the generator $L_{V,0,0}$ of the associated gradient dynamics (1-9), that is, the optimal rate of return to equilibrium in the gradient setting, it follows from Theorem 1.9 and item (iii) in Lemma 1.8 that, as soon as $B^*(\boldsymbol{s}^*) \neq 0$ for at least one $\boldsymbol{s}^* \in \boldsymbol{j}(\boldsymbol{m}^*)$, the ratio of the rates $\lambda(\boldsymbol{m}^*, h)/\lambda^\nabla(\boldsymbol{m}^*, h)$ converges to some constant $c > 1$ when $h \to 0$.

In addition, it is not difficult to see that playing with b_h, one can make $\lim_{h\to 0}(\lambda(\boldsymbol{m}^*, h)/\lambda^\nabla(\boldsymbol{m}^*, h))$ arbitrarily big. Taking, for example, $b_h = b = aJ\nabla V$ around $\boldsymbol{s}^*$ for $a \in \mathbb{R}$ and some constant antisymmetric and invertible matrix J,

$$\lim_{a\to\infty}\lim_{h\to 0}\frac{\lambda(\boldsymbol{m}^*, h)}{\lambda^\nabla(\boldsymbol{m}^*, h)} = +\infty.$$

Nevertheless, making this limit too big will deteriorate the constant C appearing in (1-23) and (1-24), as well as the interval $(0, h_0] \ni h$ for which these estimates remain relevant. A more interesting problem is the computation of the optimal rate when h_0 is small but fixed, that is, when the preceding J has a constant size (see [Lelièvre et al. 2013] in the case of linear drifts). We did not perform the whole computation, but a partial one seems to indicate that the optimal (or at least a reasonable) choice for J is given when it sends the unstable direction of Hess $V(\boldsymbol{s}^*)$ onto one of its stable directions corresponding to a maximal eigenvalue.

A closely related point of view to ours is to study the mean transition times between the different wells of the potential V for the process $(X_t)_{t\geq 0}$ solution to (1-1). In the nonreversible case, this question has been studied recently, e.g., in [Bouchet and Reygner 2016; Landim et al. 2019], to which we also refer for more details and references on this subject.

In [Bouchet and Reygner 2016], an Eyring–Kramers type formula (for the mean transition times) is derived from formal computations relying on the study of the appropriate quasipotential. In the case of a double-well potential V and under the assumption that $U_h = \nabla V + b$ (that is, that $\nu = 0$; see (1-5)) for some vector field b only satisfying $b \cdot \nabla V = 0$ (that is, without assuming $\operatorname{div} b = 0$ as we do when $\nu = 0$; see (1-3)), Bouchet and Reygner derived their formula (5.65), which, compared to a formula such as (the inverse of) (1-20), contains some extra term in the prefactor which measures the non-Gibbsianness of their situation.

In this general setting, the measure m_h is indeed invariant for the dynamics if and only if $\operatorname{div} b = 0$, and this extra term involves the integral of the function $F := \operatorname{div}(b)$ along the so-called instanton trajectory. Under the additional assumption that m_h is invariant (that is, that $F = 0$), this extra term equals 1, which leads to the formula (5.66) in [Bouchet and Reygner 2016], which is similar to (the inverse of) (1-20) in Theorem 1.9 (see more precisely Corollary 1.12, which clarifies the relation between eigenvalues of $L_{V,b,\nu}$ and mean transition times). In the present paper, we restrict ourselves to the Gibbsian case, so that our formulas do not contain any extra prefactor as discussed above. It would be of great interest to study the general case of a drift of the form $\nabla V + b$, where $b \cdot \nabla V = 0$ but without assuming $\operatorname{div} b = 0$, by mixing our approach and quasi-potential constructions.

In [Landim et al. 2019], the authors use a potential theoretic approach to prove an Eyring–Kramers type formula similar to the formula (5.66) of [Bouchet and Reygner 2016] in the case of a double-well potential V, when b and ν satisfy the relation (1-7) in such a way that $L_{V,b,\nu}$ has the form (1-8). Though the mathematical objects considered in [Landim et al. 2019] and in the present paper are not the same, these two works share some similarities. Nevertheless, we would like to emphasize that our approach permits us to go beyond the supersymmetric assumption (1-7) and to treat the case of multiple-well potentials.

To be more precise on the connections between the present paper and [Landim et al. 2019] (and also [Bouchet and Reygner 2016]), we conclude this introduction with the corollary below, which combines the results given by Theorem 1.9 when V is a double-well potential and [Landim et al. 2019, Theorem 5.2 and Remarks 5.3 and 5.6]. This result generalizes in particular, in this nonreversible double-well setting, the results obtained in the reversible case in [Bovier et al. 2004; 2005] on the relations between the small eigenvalues of $L_{V,b,\nu}$ and the mean transition times of (1-1) when $b=\nu=0$.

Corollary 1.12. *Assume that the hypotheses of Theorem 1.9 hold, with*

$$\lim_{|x|\to+\infty}\frac{x}{|x|}\cdot\nabla V(x)=+\infty \quad \textit{and} \quad \lim_{|x|\to+\infty}|\nabla V(x)|-2\Delta V(x)=+\infty,$$

and that V admits precisely two local minima $\underline{\boldsymbol{m}}$ and $\boldsymbol{m}$ such that $V(\underline{\boldsymbol{m}})<V(\boldsymbol{m})$ (hence $\mathcal{V}^{(1)}=\boldsymbol{j}(\boldsymbol{m})$). Assume in addition that b and ν satisfy the relation (1-7), and hence that $b=J\nabla V$ for some smooth map J from $\mathbb{R}^d$ into the set of real antisymmetric matrices of size d, and that J is uniformly bounded on $\mathbb{R}^d$. Let $\mathcal{O}(\underline{\boldsymbol{m}})$ be a smooth open connected set containing $\underline{\boldsymbol{m}}$ such that $\overline{\mathcal{O}(\underline{\boldsymbol{m}})}\subset\{V<\boldsymbol{\sigma}(\boldsymbol{m})\}$. Let then $(X_t)_{t\geq 0}$ be the solution to (1-1) such that $X_0=\boldsymbol{m}$ and let

$$\tau_{\mathcal{O}(\underline{\boldsymbol{m}})}:=\inf\{t\geq 0,\ X_t\in\mathcal{O}(\underline{\boldsymbol{m}})\}$$

be the first hitting time of $\mathcal{O}(\underline{\boldsymbol{m}})$. The expectation of $\tau_{\mathcal{O}(\underline{\boldsymbol{m}})}$ and the nonzero small eigenvalue $\lambda(\boldsymbol{m},h)$ of $L_{V,b,\nu}$ are then related by the following formula in the limit $h\to 0$:

$$\mathbb{E}(\tau_{\mathcal{O}(\underline{\boldsymbol{m}})})=\frac{1}{\lambda(\boldsymbol{m},h)}\big(1+\mathcal{O}\big(\sqrt{h|\ln h|^3}\,\big)\big).$$

Let us mention here that the hypotheses of Corollary 1.12 are simply the minimal hypotheses permitting the simultaneous application of Theorem 1.9 and [Landim et al. 2019, Theorem 5.2] in its refinement specified in [Landim et al. 2019, Remark 5.6].

2. General spectral estimates

2A. ***Proof of Proposition 1.1.*** The unbounded operator $(P_\phi,\mathcal{C}_c^\infty(\mathbb{R}^d))$ is accretive, since, according to (1-11), one has

$$\forall\, u\in\mathcal{C}_c^\infty(\mathbb{R}^d),\qquad \operatorname{Re}\langle P_\phi u,u\rangle=\langle\Delta_\phi u,u\rangle=\|d_\phi u\|^2\geq 0. \tag{2-1}$$

In order to prove that its closure is maximal accretive, it then suffices to show that $\operatorname{Ran}(P_\phi+1)$ is dense in $L^2(\mathbb{R}^d)$ (see, for example, [Helffer 2013, Theorem 13.14]). The proof of this fact is rather standard but we give it for the sake of completeness (see the proof of [Helffer and Nier 2005, Proposition 5.5] for a similar proof). Suppose that $f\in L^2(\mathbb{R}^d)$ is orthogonal to $\operatorname{Ran}(P_\phi+1)$. Then $(P_\phi^*+1)f=0$ in the distribution sense and, since P_ϕ is real, one can assume that f is real. In particular, since $P_\phi^* = \Delta_\phi - b_h \cdot d_\phi$ is elliptic with smooth coefficients, f belongs to $\mathcal{C}^\infty(\mathbb{R}^d)$. Thus, for every $\zeta \in \mathcal{C}_c^\infty(\mathbb{R}^d, \mathbb{R})$, one has

$$h^2\langle\nabla(\zeta f), \nabla(\zeta f)\rangle + \int \zeta^2(|\nabla\phi|^2 - h\Delta\phi + 1) f^2 = \langle(P_\phi^* + 1)\zeta f, \zeta f\rangle = h^2\int|\nabla\zeta|^2 f^2 - h\int(b_h \cdot d\zeta)\zeta f^2.$$

Take now ζ such that $0 \le \zeta \le 1$, $\zeta = 1$ on $B(0, 1)$ and $\operatorname{supp}\zeta \subset B(0, 2)$, and define $\zeta_k := \zeta\left(\frac{\cdot}{k}\right)$ for $k \in \mathbb{N}^*$. According to (1-13) and to the above relation, there exists $C > 0$ such that for every $k \in \mathbb{N}^*$,

$$\begin{aligned}\int \zeta_k^2(|\nabla\phi|^2 - h\Delta\phi + 1) f^2 &\le C\frac{h^2}{k^2}\|f\|^2 + C\frac{h}{k}\|f\|\|(1 + |\nabla\phi|)\zeta_k f\| \\ &\le C\Big(1 + \frac{1}{2\epsilon}\Big)\frac{h^2}{k^2}\|f\|^2 + \frac{\epsilon}{2}C\|(1 + |\nabla\phi|)\zeta_k f\|^2,\end{aligned}$$

where $\epsilon > 0$ is arbitrary. Choosing $\epsilon = 1/(2C)$ and using (1-12), it follows that for every $h > 0$ small enough,

$$\tfrac{1}{4}\|\zeta_k f\|^2 \le \int \zeta_k^2\big(\tfrac{1}{2}|\nabla\phi|^2 - h\Delta\phi + \tfrac{1}{2}\big) f^2 \le C\Big(1 + \frac{1}{2\epsilon}\Big)\frac{h^2}{k^2}\|f\|^2,$$

which implies, taking the limit $k \to +\infty$, that $f = 0$. Hence, the closure of P_ϕ, that we still denote by P_ϕ, is maximal accretive. Note moreover, that (2-1) implies that $D(P_\phi) \subset \{u \in L^2(\mathbb{R}^d),\ d_\phi u \in L^2(\mathbb{R}^d)\}$ and that $\operatorname{Re}\langle P_\phi u, u\rangle = \|d_\phi u\|^2$ for every $u \in D(P_\phi)$.

Let us now prove that $D(\Delta_\phi) \subset D(P_\phi)$, which amounts to showing that for every $u \in D(\Delta_\phi)$, there exists a sequence $(u_n)_{n\in\mathbb{N}}$ of $\mathcal{C}_c^\infty(\mathbb{R}^d)$ such that $u_n \to u$ in $L^2(\mathbb{R}^d)$ and $(P_\phi u_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. Since $(\Delta_\phi, \mathcal{C}_c^\infty(\mathbb{R}^d))$ is essentially self-adjoint, for any such u, there exists a sequence $(u_n)_{n\in\mathbb{N}}$ in $\mathcal{C}_c^\infty(\mathbb{R}^d)$ such that $u_n \to u$ in $L^2(\mathbb{R}^d)$ and $(\Delta_\phi u_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, and it thus suffices to show that $(b_h \cdot d_\phi u_n)_{n\in\mathbb{N}}$ is also a Cauchy sequence. For this purpose, we introduce the exterior derivative d acting from 0-forms into 1-forms and the twisted semiclassical derivative $d_\phi = e^{-\phi/h} \circ hd \circ e^{\phi/h}$. Note that the notation d_ϕ has actually already been defined in (1-15) with a different meaning; we are thus using here a slight abuse of notation, by identifying the exterior derivative d acting on functions with ∇. Thanks to (1-12) and (1-13), there exists $C > 0$ such that for every $h > 0$ small enough and every $u \in \mathcal{C}_c^\infty(\mathbb{R}^d)$, one has

$$\|b_h \cdot d_\phi u\|^2 \le \int |b_h|^2 |d_\phi u|^2 \le C\langle|\nabla\phi|^2\, d_\phi u, d_\phi u\rangle + C\|d_\phi u\|^2 \le 2C\langle\Delta_\phi^{(1)} d_\phi u, d_\phi u\rangle + 2C\|d_\phi u\|^2,$$

where $\Delta_\phi^{(1)}$ denotes the Witten Laplacian acting on 1-forms, that is,

$$\Delta_\phi^{(1)} = \Delta_\phi^{(0)} \otimes \mathrm{Id} + 2h\operatorname{Hess}\phi = (-h^2\Delta + |\nabla\phi|^2 - h\Delta\phi) \otimes \mathrm{Id} + 2h\operatorname{Hess}\phi.$$

Combined with the intertwining relation $\Delta_\phi^{(1)} d_\phi = d_\phi \Delta_\phi^{(0)}$, we get

$$\|b_h \cdot d_\phi u\|^2 \le 2C(\|\Delta_\phi^{(0)} u\|^2 + \|d_\phi u\|^2) \le 2C\|\Delta_\phi^{(0)} u\|(\|\Delta_\phi^{(0)} u\| + \|u\|) \tag{2-2}$$

for every $u \in \mathcal{C}_c^\infty(\mathbb{R}^d)$. This implies that for any Cauchy sequence $(u_n)_{n\in\mathbb{N}}$ in $L^2(\mathbb{R}^d)$ such that $(\Delta_\phi u_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, $(b_h \cdot d_\phi u_n)_{n\in\mathbb{N}}$ is also a Cauchy sequence, and thus that $D(\Delta_\phi) \subset D(P_\phi)$.

The statement about P_ϕ^* is then a straightforward consequence of the above analysis. Indeed, since $P_\phi^* = \Delta_\phi - b_h \cdot d_\phi$ on $\mathcal{C}_c^\infty(\mathbb{R}^d)$, the above arguments imply that the closure of $(P_\phi^*, \mathcal{C}_c^\infty(\mathbb{R}^d))$ is maximal accretive and that its domain contains $D(\Delta_\phi)$. Moreover, P_ϕ^* is maximal accretive since P_ϕ is, and hence coincides with the closure of $(P_\phi^*, \mathcal{C}_c^\infty(\mathbb{R}^d))$.

Let us now prove the statement on the spectrum of P_ϕ. Throughout, we will denote $\mathbb{C}_+ = \{\mathrm{Re}(z) \geq 0\}$. It follows from (1-12) and from (1-13) that for every $u \in \mathcal{C}_c^\infty(\mathbb{R}^d)$, it holds that, for some $C > 0$ and every $h > 0$ small enough,

$$|\langle b_h \cdot d_\phi u, u\rangle| \leq \|d_\phi u\| \|b_h u\| \leq C(\|d_\phi u\|^2 + \|u\| \|d_\phi u\|). \tag{2-3}$$

Set $\Lambda_0 = 5C$ for some $C \geq 1$ satisfying (2-3), and let $z \in \mathbb{C}_+$ be such that $|\mathrm{Im}(z)| \geq \Lambda_0 \max(\mathrm{Re}(z), \sqrt{\mathrm{Re}(z)}\,)$. Suppose first that $\mathrm{Re}(z)\|u\|^2 \geq \frac{1}{2}\|d_\phi u\|^2$. Then, thanks to the estimate (2-3), we have

$$\begin{aligned} |\langle (b_h \cdot d_\phi - i\,\mathrm{Im}(z))u, u\rangle| &\geq \big(|\mathrm{Im}(z)| - C\big(2\,\mathrm{Re}(z) + \sqrt{2\,\mathrm{Re}(z)}\,\big)\big)\|u\|^2 \\ &\geq C \max\big(\mathrm{Re}(z), \sqrt{\mathrm{Re}(z)}\,\big)\|u\|^2 \geq \mathrm{Re}(z)\|u\|^2. \end{aligned}$$

Since $|\langle (b_h \cdot d_\phi - i\,\mathrm{Im}(z))u, u\rangle| \leq |\langle (P_\phi - z)u, u\rangle|$, this implies that

$$|\langle (P_\phi - z)u, u\rangle| \geq \mathrm{Re}(z)\|u\|^2. \tag{2-4}$$

Suppose now that $\mathrm{Re}(z)\|u\|^2 \leq \frac{1}{2}\|d_\phi u\|^2$. One then directly obtains

$$|\langle (P_\phi - z)u, u\rangle| \geq \langle (\Delta_\phi - \mathrm{Re}(z))u, u\rangle \geq \mathrm{Re}(z)\|u\|^2,$$

which, combined with (2-4), implies that

$$\|(P_\phi - z)u\| \geq \mathrm{Re}(z)\|u\| \tag{2-5}$$

for every $z \in \mathbb{C}_+ \setminus \Gamma_{\Lambda_0}$ and $u \in \mathcal{C}_c^\infty(\mathbb{R}^d)$. Since P_ϕ is closed, it follows that $P_\phi - z$ is injective with closed range, and hence semi-Fredholm, for every $z \in \mathbb{C}_+ \setminus \Gamma_{\Lambda_0}$ such that $\mathrm{Re}(z) \neq 0$. Assume now for a while that the fourth item in Proposition 1.1, which is proved independently just below, is satisfied, and let $\lambda \in \mathbb{R}$ be such that $i\lambda \in \sigma(P_\phi)$. By assumption, $i\lambda$ is then an eigenvalue of P_ϕ and there exists some $u \in D(P_\phi) \setminus \{0\}$ such that $P_\phi u = i\lambda u$. In particular,

$$0 = \mathrm{Re}\langle P_\phi u, u\rangle = \|d_\phi u\|^2 = h^2 \|e^{-\frac{\phi}{h}} \nabla(e^{\frac{\phi}{h}} u)\|^2,$$

which implies $u \in \mathrm{Span}\{e^{-\phi/h}\}$ and then $\lambda = 0$. This shows that $\sigma(P_\phi) \cap i\mathbb{R} \subset \{0\}$ and thus, P_ϕ being maximal accretive, that $\sigma(P_\phi) \cap \{\mathrm{Re}(z) \leq 0\} \subset \{0\}$. It follows that $P_\phi - z$ is semi-Fredholm for every $z \in \mathbb{C} \setminus \Gamma_{\Lambda_0}$, and has index 0 on $\{\mathrm{Re}\, z \leq 0\} \setminus \{0\}$. But the open set $\mathbb{C} \setminus \Gamma_{\Lambda_0}$ being connected, the index of $P_\phi - z$ is constant, and then equal to 0, on $\mathbb{C} \setminus \Gamma_{\Lambda_0}$ (see [Kato 1995, Theorem 5.17 in Chapter 4]). Hence, $P_\phi - z$ being injective on $\mathbb{C} \setminus \Gamma_{\Lambda_0}$, it is invertible from $D(P_\phi)$ onto $L^2(\mathbb{R}^d)$ on $\mathbb{C} \setminus \Gamma_{\Lambda_0}$ and the resolvent estimate stated in Proposition 1.1 becomes a direct consequence of (2-5).

Let us now prove the fourth item of Proposition 1.1. Thanks to (1-12), there exist $c > 0$ and $R > 0$ such that

$$\forall\, |x| \geq R, \quad |\nabla\phi(x)|^2 \geq c.$$

Take $c_1 \in (0, c)$ and let W be a nonnegative smooth function such that $\mathrm{supp}(W) \subset B(0, R)$ and $W(x) + |\nabla\phi(x)|^2 \geq \frac{1}{2}(c + c_1)$ for all $x \in \mathbb{R}^d$. There exists consequently $h_0 > 0$ such that for all $h \in (0, h_0]$, one has

$$\tilde{W} := W + |\nabla\phi|^2 - h\Delta\phi \geq c_1$$

on $\mathbb{R}^d$. Introduce the operator

$$\tilde{P}_\phi = P_\phi + W = -h^2\Delta + \tilde{W} + b_h d_\phi$$

with domain $D(P_\phi)$. Since P_ϕ is maximal accretive and $W \in \mathcal{C}_c^\infty(\mathbb{R}^d, \mathbb{R}^+)$, $\tilde{P}_\phi$ is also maximal accretive (see, for example, [Helffer 2013, Theorem 13.25]). Moreover, for every $u \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ and then for every $u \in D(P_\phi)$, one has

$$\mathrm{Re}\langle \tilde{P}_\phi u, u\rangle = \langle(-h^2\Delta + \tilde{W})u, u\rangle \geq c_1\|u\|^2,$$

which implies as above that for every z in $\{\mathrm{Re}(z) < c_1\}$, $\tilde{P}_\phi - z$ is invertible from $D(P_\phi)$ onto $L^2(\mathbb{R}^d)$. Hence, for every z in $\{\mathrm{Re}(z) < c_1\}$, we can write

$$P_\phi - z = \tilde{P}_\phi - z - W = (\mathrm{Id} - W(\tilde{P}_\phi - z)^{-1})(\tilde{P}_\phi - z).$$

Of course, $z \mapsto (\tilde{P}_\phi - z)^{-1}$ is holomorphic on $\{\mathrm{Re}\, z < c_1\}$ and thanks to the analytic Fredholm theorem, it then suffices to prove that

$$K(z) := W(\tilde{P}_\phi - z)^{-1} : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$$

is compact for every z in $\{\mathrm{Re}(z) < c_1\}$. This follows from the compactness of the embedding $H_R^1 \subset L^2(\mathbb{R}^d)$ and from the fact that for every $z \in \{\mathrm{Re}\, z < c_1\}$, $K(z)$ acts continuously from $L^2(\mathbb{R}^d)$ into H_R^1, where

$$H_R^1 := \{u \in H^1(\mathbb{R}^d),\ \mathrm{supp}(u) \subset B(0, R)\}.$$

Indeed, for any z in $\{\mathrm{Re}(z) < c_1\}$, the operator $d_\phi(\tilde{P}_\phi - z)^{-1} : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ is continuous thanks to (2-1) and hence, since W is smooth and supported in $B(0, R)$, $K(z) : L^2(\mathbb{R}^d) \to H_R^1$ is also continuous.

To conclude, it remains to prove the last statement of Proposition 1.1. To this end, note first that $P_\phi e^{-\phi/h} = 0$ according to (1-16) and let us recall that, according to (1-12), $e^{-\phi/h} \in D(\Delta_\phi) \subset D(P_\phi)$. Thus, $\mathrm{Span}\{e^{-\phi/h}\} \subset \mathrm{Ker}\, P_\phi$ and 0 is an eigenvalue of P_ϕ. It has moreover finite algebraic multiplicity according to the preceding analysis. Conversely, the relation

$$\forall\, u \in D(P_\phi), \quad \mathrm{Re}\langle P_\phi u, u\rangle = \|d_\phi u\|^2 = h^2\|e^{-\frac{\phi}{h}}\nabla(e^{\frac{\phi}{h}}u)\|^2$$

leads to $\mathrm{Ker}\, P_\phi \subset \mathrm{Span}\{e^{-\phi/h}\}$ and the same arguments also show that $\mathrm{Ker}\, P_\phi^* = \mathrm{Span}\{e^{-\phi/h}\}$. This implies that 0 is an eigenvalue of P_ϕ with algebraic multiplicity one. Indeed, if this were not the case, there would exist $u \in D(P_\phi)$ such that $u \notin \mathrm{Ker}\, P_\phi$ and $P_\phi u = e^{-\phi/h}$, and hence such that

$$0 < \langle P_\phi u, e^{-\frac{\phi}{h}}\rangle = \langle u, P_\phi^* e^{-\frac{\phi}{h}}\rangle = 0.$$

2B. ***Spectral analysis near the origin.*** Let us denote by $(e_k^W)_{k\geq 1}$ the eigenfunctions of Δ_ϕ associated with the nondecreasing sequence of eigenvalues $(\lambda_k^W)_{k\geq 1}$. Let ϵ_0 and $h_0 > 0$ be given by Proposition 1.2.

We recall that for every $h \in (0, h_0]$,

$$\operatorname{card}(\sigma(\Delta_\phi) \cap \{\operatorname{Re} z < \epsilon_0 h\}) = n_0,$$

where n_0 is the number of local minima of ϕ. We define

$$R_- : \mathbb{C}^{n_0} \to L^2(\mathbb{R}^d), \qquad (\alpha_k) \mapsto \sum_{k=1}^{n_0} \alpha_k e_k^W$$

and $R_+ := R_-^*$, i.e.,

$$R_+ : L^2(\mathbb{R}^d) \to \mathbb{C}^{n_0}, \qquad u \mapsto (\langle u, e_k^W \rangle)_{k=1,\dots,n_0}.$$

Note in particular the relations

$$R_+ R_- = \mathrm{Id}_{\mathbb{C}^{n_0}} \quad \text{and} \quad R_- R_+ = \Pi, \tag{2-6}$$

where Π denotes the orthogonal projection onto $\operatorname{Ran}(R_-) = \operatorname{Span}(e_k^W, k \in \{1, \dots, n_0\})$. We also define the spectral projector

$$\hat{\Pi} := 1 - \Pi.$$

For $z \in \mathbb{C}$, let us then consider on the Hilbert space $\hat{E} := \operatorname{Ran}(\hat{\Pi})$ the following unbounded operator which will be useful in the rest of this section:

$$\hat{P}_{\phi,z} := \hat{\Pi}(P_\phi - z)\hat{\Pi} \quad \text{with domain } D(\hat{P}_{\phi,z}) := \hat{\Pi}(D(P_\phi)). \tag{2-7}$$

Hence $D(\hat{P}_{\phi,z})$ is dense in $\hat{E}$ and, since $\operatorname{Ran}\Pi \subset D(\Delta_\phi) \subset D(P_\phi)$, it holds that $\hat{\Pi}(D(P_\phi)) \subset D(P_\phi)$ and $\hat{P}_{\phi,z}$ is well and densely defined.

Lemma 2.1. *Let ϵ_0 and $h_0 > 0$ be given by Proposition 1.2. Then, for every $h \in (0, h_0]$, the operator $\hat{P}_{\phi,z} : D(\hat{P}_{\phi,z}) \to \hat{E}$ defined in (2-7) is invertible on $\{\operatorname{Re} z < \epsilon_0 h\}$. Moreover, for any $\epsilon_1 \in (0, \epsilon_0)$,*

$$\forall\, z \in \{\operatorname{Re} z < \epsilon_1 h\}, \quad \|\hat{P}_{\phi,z}^{-1}\|_{\hat{E} \to \hat{E}} = \mathcal{O}(h^{-1}),$$

uniformly with respect to z.

Proof. We begin with the following observation: the unbounded operator

$$\hat{\Pi}(P_\phi^* - z)\hat{\Pi} \quad \text{with domain } \hat{\Pi}(D(P_\phi^*)) \subset D(P_\phi^*)$$

is well and densely defined on $\hat{E}$, and satisfies

$$\hat{\Pi}(P_\phi^* - z)\hat{\Pi} = \hat{P}_{\phi,z}^*.$$

Indeed, the relation $\langle \hat{\Pi}(P_\phi - z)\hat{\Pi} v, w \rangle = \langle v, \hat{\Pi}(P_\phi^* - z)\hat{\Pi} w \rangle$, valid for every $v \in D(P_\phi)$ and $w \in D(P_\phi^*)$, implies that $\hat{\Pi}(P_\phi^* - z)\hat{\Pi} \subset \hat{P}_{\phi,z}^*$. Moreover, for every $v \in D(P_\phi)$ and $w \in D(\hat{P}_{\phi,z}^*)$, one has

$$\begin{aligned} \langle (P_\phi - z)v, w \rangle &= \langle (P_\phi - z)\Pi v, w \rangle + \langle (P_\phi - z)\hat{\Pi} v, w \rangle \\ &= \langle (P_\phi - z)\Pi v, w \rangle + \langle \hat{\Pi} v, \hat{P}_{\phi,z}^* w \rangle. \end{aligned}$$

Since $P_\phi \Pi$ is continuous, Π being continuous with finite rank, one has $|\langle P_\phi \Pi v, w\rangle| \le C\|v\|\|w\|$ for some $C > 0$ independent of (v, w), which implies that $w \in D(P_\phi^*)$. Hence $D(\hat{P}^*_{\phi,z}) \subset D(P_\phi^*)$ and since $\mathrm{Ran}(\Pi) \subset D(\Delta_\phi) \subset D(P_\phi^*)$, this implies $\hat{\Pi}(P_\phi^* - z)\hat{\Pi} = \hat{P}^*_{\phi,z}$.

Let us now consider z in $\{\mathrm{Re}\, z < \epsilon_0 h\}$ and let us prove that $\hat{P}_{\phi,z}$ is invertible from $D(\hat{P}_{\phi,z})$ onto $\hat{E}$. First, according to Proposition 1.2, we have for every $u \in D(\Delta_\phi)$,

$$\mathrm{Re}\langle (P_\phi - z)\hat{\Pi}u, \hat{\Pi}u\rangle = \langle (\Delta_\phi - \mathrm{Re}(z))\hat{\Pi}u, \hat{\Pi}u\rangle \ge (\epsilon_0 h - \mathrm{Re}\, z)\|\hat{\Pi}u\|^2, \tag{2-8}$$

and the inequality (2-8) is also true when $u \in D(P_\phi)$. Indeed, for any $u \in D(P_\phi)$, there exists a sequence $(u_n)_{n\in\mathbb{N}}$ in $D(\Delta_\phi)$ such that $u_n \to u$ and $P_\phi u_n \to P_\phi u$ in $L^2(\mathbb{R}^d)$. Hence $\hat{\Pi}u_n \to \hat{\Pi}u$ and, since $P_\phi \Pi$ is continuous, it also holds that $P_\phi \hat{\Pi} u_n \to P_\phi \hat{\Pi} u$. In particular, it follows that $\hat{P}_{\phi,z}$ is injective. Note that a similar analysis shows that $\hat{P}^*_{\phi,z}$ is also injective.

Second, let us show that $\hat{P}_{\phi,z}$ is closed, which will in particular imply that $\mathrm{Ran}(\hat{P}_{\phi,z})$ is closed according to (2-8). For brevity, we denote $\hat{P} = \hat{P}_{\phi,z}$ and $P = P_\phi$. Suppose that $(u_n)_{n\in\mathbb{N}}$ is a sequence in $D(\hat{P}) \subset D(P)$ such that $u_n \to u$ and $\hat{P}u_n \to v$ in $\hat{E}$. Since $\mathrm{Ran}\,\Pi \subset D(\Delta_\phi) \subset D(P^*)$,

$$\Pi P u_n = \sum_{k=1}^{n_0} \langle P u_n, e_k^W\rangle e_k^W = \sum_{k=1}^{n_0} \langle u_n, P^* e_k^W\rangle e_k^W \xrightarrow[n\to+\infty]{} \sum_{k=1}^{n_0} \langle u, P^* e_k^W\rangle e_k^W,$$

so $(P-z)u_n = \hat{P}u_n + \Pi(P-z)u_n$ converges. Since P is closed, this implies $u \in D(P) \cap \mathrm{Ran}\,\hat{\Pi} = \hat{\Pi}(D(P))$ and that

$$(P - z)u = v + g \quad \text{with } g \in \mathrm{Ran}\,\Pi.$$

Multiplying this relation by $\hat{\Pi}$, we get $v = \hat{P}u$, which proves that $\hat{P}$ is closed.

To prove that $\hat{P}$ is invertible from $D(\hat{P})$ onto $\hat{E}$, it is thus enough to prove that $\mathrm{Ran}(\hat{P})$ is dense in $\hat{E}$. Let then $v \in \hat{E}$ be such that $\langle \hat{P}u, v\rangle = 0$ for all $u \in D(\hat{P})$. Then $v \in D(\hat{P}^*)$ and $\hat{P}^* v = 0$. Thus, by injectivity of $\hat{P}^*$, $v = 0$, which proves the invertibility of $\hat{P} : D(\hat{P}_{\phi,z}) \to \hat{E}$.

The relation (2-8) then implies that for all $z \in \{\mathrm{Re}\, z \le \epsilon_1 h\}$, one has

$$\mathrm{Re}\langle (P_\phi - z)\hat{\Pi}u, \hat{\Pi}u\rangle \ge \delta h \|\hat{\Pi}u\|^2$$

with $\delta = \epsilon_0 - \epsilon_1 > 0$. Hence, for the operator norm on $\hat{E} \subset L^2(\mathbb{R}^d)$,

$$\hat{P}^{-1}_{\phi,z} = \mathcal{O}(h^{-1}),$$

uniformly with respect to $z \in \{\mathrm{Re}\, z < \epsilon_1 h\}$. □

For $z \in \mathbb{C}$, we now consider the Grushin operator $\mathcal{P}_\phi(z) : D(P_\phi) \times \mathbb{C}^{n_0} \to L^2(\mathbb{R}^d) \times \mathbb{C}^{n_0}$ defined by

$$\mathcal{P}_\phi(z) = \begin{pmatrix} P_\phi - z & R_- \\ R_+ & 0 \end{pmatrix}. \tag{2-9}$$

Lemma 2.2. *Let ϵ_0 and $h_0 > 0$ be given by Proposition 1.2. Then, the operator $\mathcal{P}_\phi(z)$ is invertible on $\{\mathrm{Re}\, z < \epsilon_0 h\}$. More precisely, for every $z \in \{\mathrm{Re}\, z < \epsilon_0 h\}$, $(u, u_-) \in D(P_\phi) \times \mathbb{C}^{n_0}$ and $(f, y) \in L^2(\mathbb{R}^d) \times \mathbb{C}^{n_0}$,*

$$\mathcal{P}_\phi(z)(u, u_-) = (f, y)$$

if and only if

$$(u, u_-) = (R_- y + v,\ R_+ f - R_+(P_\phi - z)R_- y - R_+ P_\phi v),$$

where

$$v := \hat{P}_{\phi,z}^{-1}\hat{\Pi} f - \hat{P}_{\phi,z}^{-1}\hat{\Pi} P_\phi R_- y \in \hat{\Pi}(D(P_\phi)).$$

Proof. Let $(f, y) \in L^2(\mathbb{R}^d) \times \mathbb{C}^{n_0}$ and assume that $(u, u_-) \in D(P_\phi) \times \mathbb{C}^{n_0}$ satisfies

$$\begin{cases} (P_\phi - z)u + R_- u_- = f, \\ R_+ u = y. \end{cases} \tag{2-10}$$

Applying R_+ to the first equation and R_- to the second one, we get, according to (2-6):

$$u_- = R_+ f - R_+(P_\phi - z)u \quad \text{and} \quad u = R_- y + v,$$

with $v \in \operatorname{Ran} \hat{\Pi} \cap D(P_\phi) = \hat{\Pi}(D(P_\phi))$ a solution to

$$(P_\phi - z)R_- y + (P_\phi - z)v + R_- u_- = f.$$

Then, applying $\hat{\Pi}$ to the latter equation, we get, using $\hat{\Pi} R_- = 0$,

$$\hat{\Pi}(P_\phi - z)\hat{\Pi} v = \hat{\Pi} f - \hat{\Pi}(P_\phi - z)R_- y - \hat{\Pi} R_- u_- = \hat{\Pi} f - \hat{\Pi} P_\phi R_- y. \tag{2-11}$$

Conversely, note that if $v \in \operatorname{Ran} \hat{\Pi} \cap D(P_\phi)$ is a solution to (2-11), then according to (2-6),

$$\big(u = R_- y + v, u_- = R_+ f - R_+(P_\phi - z)(R_- y + v)\big) \in D(P_\phi) \times \mathbb{C}^{n_0}$$

is a solution to (2-10).

Hence, the statement of Lemma 2.2 simply follows from Lemma 2.1 which implies that, for every $z \in \{\operatorname{Re} z < \epsilon_0 h\}$,

$$v = \hat{P}_{\phi,z}^{-1}\hat{\Pi} f - \hat{P}_{\phi,z}^{-1}\hat{\Pi} P_\phi R_- y \in \hat{\Pi}(D(P_\phi))$$

is the unique solution to (2-11). □

Proof of Theorem 1.3. Let ϵ_0 and h_0 be as in Lemmata 2.1 and 2.2, and take $\epsilon_1 \in (0, \epsilon_0)$. For $z \in \{\operatorname{Re} z < \epsilon_0 h\}$, let $\mathcal{E}_\phi(z) = \mathcal{P}_\phi(z)^{-1}$. According to Lemma 2.2, it thus holds that

$$\mathcal{E}_\phi(z) = \begin{pmatrix} E(z) & E_+(z) \\ E_-(z) & E_{-+}(z) \end{pmatrix},$$

where E, E_-, E_+, E_{-+} are holomorphic in $\{\operatorname{Re} z < \epsilon_0 h\}$ and satisfy the formulas

$$E_+(z) = R_- - \hat{P}_{\phi,z}^{-1}\hat{\Pi} P_\phi R_-, \qquad E_-(z) = R_+ - R_+ P_\phi \hat{P}_{\phi,z}^{-1}\hat{\Pi}, \tag{2-12}$$

$$E_{-+}(z) = -R_+(P_\phi - z)R_- + R_+ P_\phi \hat{P}_{\phi,z}^{-1}\hat{\Pi} P_\phi R_- \tag{2-13}$$

and

$$E(z) = \hat{P}_{\phi,z}^{-1}\hat{\Pi}. \tag{2-14}$$

Moreover, $P_\phi - z$ is invertible if and only if $E_{-+}(z)$ is, in which case,

$$(P_\phi - z)^{-1} = E(z) - E_+(z)E_{-+}(z)^{-1}E_-(z). \tag{2-15}$$

We refer to [Sjöstrand and Zworski 2007] for more details.

We now want to use these formulas to compute the number of poles of $(P_\phi - z)^{-1}$. Thanks to (2-2), one has, for some $C > 0$ and all $k \in \{1, \dots, n_0\}$,

$$\|b_h \cdot d_\phi e_k^W\| \le C(\|\Delta_\phi e_k^W\| + \|d_\phi e_k^W\|) \le C\Big(\lambda_k^W + \sqrt{\lambda_k^W}\Big).$$

Using the bound $\lambda_k^W \le Che^{-2S/h}$ given by Proposition 1.2, this yields the existence of some $C > 0$ such that for every $k \in \{1, \dots, n_0\}$,

$$\|b_h \cdot d_\phi e_k^W\| \le C\sqrt{h}e^{-\frac{S}{h}} \quad \text{and} \quad \|P_\phi e_k^W\| \le C\sqrt{h}e^{-\frac{S}{h}}. \tag{2-16}$$

This shows that $R_+\Delta_\phi R_- = \mathcal{O}(he^{-2S/h})$ and $R_+ b\cdot d_\phi R_- = \mathcal{O}(\sqrt{h}e^{-S/h})$. Hence, for all $z \in \mathbb{C}$,

$$R_+(P_\phi - z)R_- = R_+P_\phi R_- - z\,\mathrm{Id}_{\mathbb{C}^{n_0}} = -z\,\mathrm{Id}_{\mathbb{C}^{n_0}} + \mathcal{O}(\sqrt{h}e^{-\frac{S}{h}}). \tag{2-17}$$

On the other hand, we deduce from (2-16) and from the related relation

$$\langle P_\phi u, e_k^W\rangle = \langle u, \Delta_\phi e_k^W\rangle - \langle u, b_h \cdot d_\phi e_k^W\rangle = \mathcal{O}(he^{-2\frac{S}{h}} + \sqrt{h}e^{-\frac{S}{h}})\|u\|,$$

valid for any $u \in D(P_\phi)$ and $k \in \{1, \dots, n_0\}$, that

$$P_\phi R_- = \mathcal{O}(h^{\frac{1}{2}}e^{-\frac{S}{h}}) \quad \text{and} \quad R_+P_\phi = \mathcal{O}(h^{\frac{1}{2}}e^{-\frac{S}{h}}). \tag{2-18}$$

Moreover, we know from Lemma 2.1 that $\hat{P}_{\phi,z}^{-1} = \mathcal{O}(h^{-1})$ uniformly on $\{\mathrm{Re}\, z < \epsilon_1 h\}$. Therefore, injecting this estimate and (2-17) and (2-18) into (2-13) and (2-12), respectively, we obtain uniformly on $\{\mathrm{Re}\, z < \epsilon_1 h\}$,

$$E_{-+}(z) = z\,\mathrm{Id}_{\mathbb{C}^{n_0}} + \mathcal{O}(h^{\frac{1}{2}}e^{-\frac{S}{h}}) \tag{2-19}$$

and

$$E_+(z) = R_- + \mathcal{O}(h^{-\frac{1}{2}}e^{-\frac{S}{h}}) \quad \text{and} \quad E_-(z) = R_+ + \mathcal{O}(h^{-\frac{1}{2}}e^{-\frac{S}{h}}). \tag{2-20}$$

According to (2-19), $E_{-+}(z)$ is then invertible when $z \in \{\mathrm{Re}\, z < \epsilon_1 h\}$ satisfies $|z| \ge Ch^{\frac{1}{2}}e^{-S/h}$ for C large enough and the spectrum of P_ϕ in $\{\mathrm{Re}\, z < \epsilon_1 h\}$ is then of order $\mathcal{O}(h^{\frac{1}{2}}e^{-S/h})$. Moreover, for $|z| = \frac{1}{2}\epsilon_1 h$,

$$E_{-+}(z) = z(\mathrm{Id}_{\mathbb{C}^{n_0}} + \mathcal{O}(h^{-\frac{1}{2}}e^{-\frac{S}{h}})) \tag{2-21}$$

and injecting (2-21) and (2-20) into (2-15) shows that

$$(P_\phi - z)^{-1} = E(z) - \frac{1}{z}(\Pi + \mathcal{O}(h^{-\frac{1}{2}}e^{-\frac{S}{h}})).$$

Thus, the spectral projector on the open disk $D\big(0, \frac{1}{2}\epsilon_1 h\big)$ satisfies

$$\Pi_{D(0,\frac{\epsilon_1}{2}h)} := -\frac{1}{2\pi i}\int_{\partial D(0,\frac{\epsilon_1}{2}h)} (P_\phi - z)^{-1}\,dz = \Pi + \mathcal{O}(h^{-\frac{1}{2}}e^{-\frac{S}{h}}),$$

where we recall that Π is a projector of rank n_0. This implies that for every $h > 0$ small enough, the rank of $\Pi_{D(0,\epsilon_1/2h)}$, which is the number of eigenvalues of P_ϕ in $D\big(0, \frac{1}{2}\epsilon_1 h\big)$ counted with algebraic multiplicity, is precisely n_0.

In order to achieve the proof of Theorem 1.3, it just remains to prove the resolvent estimate stated there. On the one hand, it follows easily from (2-14), (2-20), and Lemma 2.1 that

$$E(z)=\mathcal{O}(h^{-1}),\ E_-(z)=\mathcal{O}(1), \quad \text{and} \quad E_+(z)=\mathcal{O}(1),$$

uniformly with respect to $z\in\{\operatorname{Re} z<\epsilon_1 h\}$. On the other hand, taking $\epsilon\in(0,\epsilon_1)$, it follows from (2-19) that $E_{-+}^{-1}(z)=\mathcal{O}(h^{-1})$, uniformly with respect to $z\in\{\operatorname{Re} z<\epsilon_1 h\}\cap\{|z|>\epsilon h\}$. Plugging all these estimates into (2-15), we obtain the announced result.

Finally, since $\sigma(P_\phi^*)=\overline{\sigma(P_\phi)}$ and, for all $z\notin\sigma(P_\phi)$, $\|(P_\phi^*-\bar{z})^{-1}\|=\|(P_\phi-z)^{-1}\|$, it follows easily that the conclusions of Theorem 1.3 also hold true for P_ϕ^*. □

3. Geometric preparation

Let us begin this section by observing that the identity $b\cdot\nabla V=0$ arising from (1-3) implies that $\mathcal{U}\subset\{x\in\mathbb{R}^d, b(x)=0\}$, where we recall that $\mathcal{U}$ denotes the set of critical points of the Morse function V, as can be easily proved using a Taylor expansion. Moreover, we have the following:

Lemma 3.1. *Suppose that Assumptions 1 and 3 hold true and let $\boldsymbol{u}\in\mathcal{U}$ be a critical point of V. Then, there exists a smooth map $J_{\boldsymbol{u}}:\mathbb{R}^d\to\mathcal{M}_d(\mathbb{R})$ such that $J_{\boldsymbol{u}}(\boldsymbol{u})$ is antisymmetric and $b(x)=J_{\boldsymbol{u}}(x)\nabla V(x)$ for all x in some neighborhood of $\boldsymbol{u}$. Moreover,*

$$J_{\boldsymbol{u}}(\boldsymbol{u})=B(\boldsymbol{u})\operatorname{Hess}V(\boldsymbol{u})^{-1},$$

where $B(\boldsymbol{u})=\operatorname{Jac}_{\boldsymbol{u}}b$ is the Jacobian matrix of b at $\boldsymbol{u}$.

Proof. Let $\boldsymbol{u}\in\mathcal{U}$ that we assume to be 0 to lighten the notation. Thanks to the Taylor formula, there exists a smooth map $G:\mathbb{R}^d\to\mathcal{M}_d(\mathbb{R})$ such that $b(x)=G(x)x$ for all $x\in\mathbb{R}^d$ and $G(0)=\operatorname{Jac}_0 b$. The same construction works for ∇V and denoting by $\mathcal{S}_d$ the set of symmetric matrices, there exists a smooth map $A:\mathbb{R}^d\to\mathcal{S}_d$ such that $\nabla V(x)=A(x)x$ for all $x\in\mathbb{R}^d$ and $A(0)=\operatorname{Hess}V(0)$. The equation $\langle b(x),\nabla V(x)\rangle=0$ for all $x\in\mathbb{R}^d$ then yields $\langle G(x)x, A(x)x\rangle=0$ and hence, since $A(x)$ is symmetric, $\langle A(x)G(x)x,x\rangle=0$ for all $x\in\mathbb{R}^d$. Expanding $A(x)G(x)$ in powers of x, this implies that

$$\forall\, x\in\mathbb{R}^d,\quad \langle A(0)G(0)x,x\rangle=0.$$

Hence, the matrix $A(0)G(0)$ is antisymmetric. Since $A(0)$ is symmetric and invertible (since V is a Morse function), this implies that $G(0)A(0)^{-1}$ is antisymmetric. Moreover, $A(x)$ is then also invertible in a neighborhood $\mathcal{V}$ of 0 and we can thus define $J_0(x)=G(x)A(x)^{-1}$ on $\mathcal{V}$. One then has

$$J_0(x)\nabla V(x)=G(x)A(x)^{-1}A(x)x=b(x)$$

for all $x\in\mathcal{V}$ and $J_0(0)=G(0)A(0)^{-1}$ is antisymmetric thanks to the above analysis. □

Remark 3.2. It is not clear from the above proof that the relation $b\cdot\nabla V=0$ implies the existence of a smooth map $J:\mathbb{R}^d\to\mathcal{M}_d(\mathbb{R})$ with antisymmetric matrix values such that $b=J\nabla V$. However, it follows from (1-3) that for such a map J, the vector fields of the form $b_h = J\nabla V + h\nu$ enter our framework as soon as

$$\operatorname{div}\nu = 0 \quad \text{and} \quad \Big(\sum_{i=1}^{d} \partial_i J_{ij}\Big)_{j=1,\dots,d} \cdot \nabla V = \nu \cdot \nabla V. \tag{3-1}$$

This is for instance the case when

$$\nu = \Big(\sum_{i=1}^{d} \partial_i J_{ij}\Big)_{j=1,\dots,d},$$

which is in particular satisfied when J appears to be constant. Moreover, when $\nu = \big(\sum_{i=1}^{d} \partial_i J_{ij}\big)_{j=1,\dots,d}$, $L_{V,b,\nu}$ (or equivalently P_ϕ) admits a supersymmetric structure according (see (1-8)) to

$$L_{V,b,\nu} = -he^{\frac{V}{h}} \operatorname{div} \circ (e^{-\frac{V}{h}}(I_d - J)\nabla) = h\nabla^*(I_d - J)\nabla,$$

where the adjoint is considered with respect to m_h, or equivalently

$$P_\phi = \Delta_\phi + b_h \cdot d_\phi = d_\phi^*(I_d - J)d_\phi,$$

where the adjoint is now considered with respect to the Lebesgue measure. Using this structure, we may follow the general approach of [Hérau et al. 2011] to analyze the spectrum of P_ϕ. Nevertheless, the operator P_ϕ still does not have any PT-symmetry and following this approach would again require us to use Theorem A.4 instead of Fan inequalities in the final part of the analysis. We believe that this approach may yield complete asymptotic expansions of the small eigenvalues of P_ϕ (or $L_{V,b,\nu}$) in this setting.

However, when J has antisymmetric matrix values and (3-1) holds but $\nu \neq \big(\sum_{i=1}^{d} \partial_i J_{ij}\big)_{j=1,\dots,d}$, the operator P_ϕ is not supersymmetric anymore (see [Michel 2016] for related results).

We are now in position to prove Lemma 1.8. Throughout the rest of this section, we denote by

$$-\mu_1 < 0 < \mu_2 \leq \cdots \leq \mu_d$$

the eigenvalues of $\operatorname{Hess} V(\boldsymbol{s})$ counted with multiplicity. For brevity, we will denote

$$B = B(\boldsymbol{s}) = \operatorname{Jac}_{\boldsymbol{s}} b \quad \text{and} \quad J = J(\boldsymbol{s}) = B(\boldsymbol{s})(\operatorname{Hess} V(\boldsymbol{s}))^{-1}.$$

We recall from Lemma 3.1 that J is antisymmetric.

Step 1: Let us first prove that $\det(\operatorname{Hess} V(\boldsymbol{s}) + B^*) < 0$. Since the matrix $\operatorname{Hess} V(\boldsymbol{s}) + B^*$ is real, it thus admits at least one negative eigenvalue.

Since $\operatorname{Hess} V(\boldsymbol{s})$ is real and symmetric, there exists $P \in \mathcal{M}_d(\mathbb{R})$ such that

$$P^* = P^{-1} \quad \text{and} \quad \operatorname{Hess} V(\boldsymbol{s}) = PDP^{-1},$$

where $D := \operatorname{Diag}(-\mu_1, \mu_2, \dots, \mu_d)$. Then:

$$\operatorname{Hess} V(\boldsymbol{s}) + B^* = \operatorname{Hess} V(\boldsymbol{s})(I_d - J) = PD(I_d - P^{-1}JP)P^{-1}. \tag{3-2}$$

Since $(P^{-1}JP)^* = -P^{-1}JP$, there exist moreover $p \in \{0, \dots, \lfloor \frac{d}{2} \rfloor\}$, $\eta_1, \dots, \eta_p > 0$, and $Q \in \mathcal{M}_d(\mathbb{R})$ satisfying $Q^* = Q^{-1}$ such that

$$Q^{-1}P^{-1}JPQ = \begin{bmatrix} A_1 & & (0) & \\ & & \ddots & \\ (0) & & A_p & \\ & & & (0) \end{bmatrix},$$

where, for every $k \in \{1, \dots, p\}$,

$$A_k = \begin{bmatrix} 0 & -\eta_k \\ \eta_k & 0 \end{bmatrix}.$$

Here, the rank of the matrix J is $2p$ and its nonzero eigenvalues are the $\pm i\eta_k,\ k \in \{1, \dots, p\}$. Therefore,

$$Q^{-1}(I_d - P^{-1}JP)Q = \begin{bmatrix} B_1 & & (0) & \\ & & \ddots & \\ (0) & & B_p & \\ & & & I_{d-2p} \end{bmatrix}, \tag{3-3}$$

where, for every $k \in \{1, \dots, p\}$,

$$B_k = \begin{bmatrix} 1 & \eta_k \\ -\eta_k & 1 \end{bmatrix}.$$

We then deduce from (3-2) and (3-3) that

$$\det(\operatorname{Hess} V(\boldsymbol{s}) + B^*) = -(\Pi_{k=1}^{d}\mu_k)(\Pi_{k=1}^{p}(1+\eta_k^2)) < 0,$$

which concludes this first step.

Step 2: Let us denote by μ a negative eigenvalue of $\operatorname{Hess} V(\boldsymbol{s}) + B^*$ and let us show that μ is its only negative eigenvalue and has geometric multiplicity one.

Assume first by contradiction that μ has geometric multiplicity two and denote by ξ_1, ξ_2 two associated unitary eigenvectors such that $\langle \xi_1, \xi_2 \rangle = 0$. Let us also define $\xi_i' := P^{-1}\xi_i$ for $i \in \{1, 2\}$ so that ξ_1' and ξ_2' are orthogonal and unitary. According to (3-2), for $i \in \{1, 2\}$,

$$D(I_d - P^{-1}JP)\xi_i' = \mu\xi_i' \quad \text{and hence,} \quad D^{-1}\xi_i' = \frac{1}{\mu}(I_d - P^{-1}JP)\xi_i'.$$

In particular, since $(P^{-1}JP)^* = -P^{-1}JP$, for every $(a, b) \in \mathbb{R}^2$ satisfying $a^2 + b^2 = 1$,

$$\langle D^{-1}(a\xi_1' + b\xi_2'), a\xi_1' + b\xi_2' \rangle = \frac{1}{\mu}.$$

Applying the max-min principle to the symmetric matrix D^{-1}, this shows that the second eigenvalue $\mu_2(D^{-1})$ of the matrix D^{-1} satisfies $\mu_2(D^{-1}) \le 1/\mu < 0$, contradicting

$$D^{-1} = \operatorname{Diag}\left(-\frac{1}{\mu_1}, \frac{1}{\mu_2}, \dots, \frac{1}{\mu_d}\right).$$

Hence the negative eigenvalue μ has geometric multiplicity one and we have to show that it is the only negative eigenvalue of Hess $V(\boldsymbol{s}) + B^*$. We reason again by contradiction, assuming that Hess $V(\boldsymbol{s}) + B^*$ admits another negative eigenvalue that we denote by η. Note in particular that it follows from the relation (see (3-2))

$$\operatorname{Hess} V(\boldsymbol{s})(I_d + J) = \operatorname{Hess} V(\boldsymbol{s})(\operatorname{Hess} V(\boldsymbol{s}) + B^*)^*(\operatorname{Hess} V(\boldsymbol{s}))^{-1}$$

that η is also an eigenvalue of Hess $V(\boldsymbol{s}) - B^*(\boldsymbol{s}) = \operatorname{Hess} V(\boldsymbol{s})(I_d + J)$. Denote now by ξ_1 a unitary eigenvector of Hess $V(\boldsymbol{s}) + B^*$ associated with μ and by ξ_2 a unitary eigenvector of Hess $V(\boldsymbol{s}) - B^*$ associated with η. Defining again $\xi_i' := P^{-1}\xi_i$ for $i \in \{1, 2\}$, we thus have

$$D^{-1}\xi_1' = \frac{1}{\mu}(I_d - P^{-1}JP)\xi_1' \quad \text{and} \quad D^{-1}\xi_2' = \frac{1}{\eta}(I_d + P^{-1}JP)\xi_2'.$$

It follows that

$$\langle D^{-1}\xi_1', \xi_2'\rangle = 0, \quad \langle D^{-1}\xi_1', \xi_1'\rangle = \frac{1}{\mu} \quad \text{and} \quad \langle D^{-1}\xi_2', \xi_2'\rangle = \frac{1}{\eta}.$$

The vectors ξ_1' and ξ_2' are in particular linearly independent and it holds for some positive constant c and every $(a, b) \in \mathbb{R}^2 \setminus \{(0, 0)\}$,

$$\langle D^{-1}(a\xi_1' + b\xi_2'), a\xi_1' + b\xi_2'\rangle = \frac{a^2}{\mu} + \frac{b^2}{\eta} \leq -c\|a\xi_1' + b\xi_2'\|^2.$$

Applying again the max-min principle to the symmetric matrix D^{-1} leads to $\mu_2(D^{-1}) \leq -c < 0$ and hence to a contradiction. This concludes the proof of the second step.

Step 3: Let us now prove the relation

$$\det(\operatorname{Hess} V(\boldsymbol{s}) + 2|\mu|\xi\xi^*) = -\det \operatorname{Hess} V(\boldsymbol{s}), \tag{3-4}$$

which is equivalent to

$$\det(I_d + 2|\mu|D^{-1}\xi'\xi'^*) = -1, \tag{3-5}$$

where ξ denotes a unitary eigenvector of Hess $V(\boldsymbol{s}) + B^*$ associated with μ and $\xi' := P^{-1}\xi$. To this end, note first that obviously

$$\forall x \in (\xi')^\perp, \quad (I_d + 2|\mu|D^{-1}\xi'\xi'^*)x = x. \tag{3-6}$$

Moreover, since $D^{-1}\xi' = 1/\mu(I_d - P^{-1}JP)\xi'$,

$$\begin{aligned}(I_d + 2|\mu|D^{-1}\xi'\xi'^*)\xi' &= \xi' + 2|\mu|D^{-1}\xi' \\ &= -\xi' + 2P^{-1}JP\xi'.\end{aligned} \tag{3-7}$$

Since $P^{-1}JP\xi'$ belongs to $(\xi')^\perp$, we deduce (3-5) and then (3-4) from (3-6) and (3-7).

Step 4: To conclude the proof of the second item of Lemma 1.8, it only remains to show that the real symmetric matrix $M_V := \operatorname{Hess} V(\boldsymbol{s}) + 2|\mu|\xi\xi^*$ is positive definite, where we recall that ξ denotes a unitary eigenvector of Hess $V(\boldsymbol{s}) + B^*$ associated with μ. This is an easy consequence of the max-min principle and of the relation $\det M_V = -\det D > 0$ obtained in the previous step. Defining again $\xi' := P^{-1}\xi$, we have

$$\forall\, x \in ((1,0,\dots,0)^*)^{\perp}, \quad \begin{aligned}\langle (D + 2|\mu|\xi'\xi'^*)x, x\rangle &= \langle Dx, x\rangle + 2|\mu|\langle \xi, x\rangle^2 \\ &\geq \mu_2\|x\|^2,\end{aligned}$$

which implies that the second eigenvalue of $D + 2|\mu|\xi'\xi'^*$, that is, the second eigenvalue of M_V, is greater than or equal to μ_2, and hence positive. The first eigenvalue of M_V is then positive according to $\det M_V > 0$. This concludes this step of the proof.

Step 5: We now prove the third item of Lemma 1.8. Since $\operatorname{Hess} V(\boldsymbol{s})(I_d - J)\xi = \mu\xi$ and $J^* = -J$, it first holds that

$$(\operatorname{Hess} V(\boldsymbol{s}))^{-1}\xi = \frac{1}{\mu}(I_d - J)\xi \quad \text{and then} \quad \langle (\operatorname{Hess} V(\boldsymbol{s}))^{-1}\xi, \xi\rangle = \frac{1}{\mu}, \tag{3-8}$$

which proves the second part of the third item of Lemma 1.8. Defining again $\xi' := P^{-1}\xi$, this also means

$$-\frac{1}{\mu_1} + \sum_{k=2}^{d}\Big(\frac{1}{\mu_k} + \frac{1}{\mu_1}\Big)\xi_k'^2 = -\frac{1}{\mu_1}\xi_1'^2 + \sum_{k=2}^{d}\frac{1}{\mu_k}\xi_k'^2 = \langle D^{-1}\xi', \xi'\rangle = \frac{1}{\mu}.$$

This implies that $1/\mu \geq -1/\lambda_1$, i.e., that $|\mu| \geq \mu_1$, with equality if and only if $\xi' = \pm(1,0,\dots,0)^*$, that is, if and only if ξ is a unitary eigenvector of $(\operatorname{Hess} V(\boldsymbol{s}))^{-1}$ associated with $-1/\mu_1$, which is equivalent to the relation $J\xi = 0$ by (3-8), and hence to $B^*\xi = 0$ since $J = -(\operatorname{Hess} V(\boldsymbol{s}))^{-1}B^*$.

4. Spectral analysis in the case of Morse functions

4A. ***Construction of accurate quasimodes.*** In the following, we assume that Assumption 4 is satisfied. Let us then consider some arbitrary $\boldsymbol{m} \in \mathcal{U}^{(0)} \setminus \{\underline{\boldsymbol{m}}\}$; that is, according to Assumption 4, a local minimum of V which is not the global minimum $\underline{\boldsymbol{m}}$ of V. According to the labeling procedure of Section 1C leading to the definitions (1-17)–(1-19), it holds in particular that $\boldsymbol{m} = \boldsymbol{m}_{i,j}$ and $\boldsymbol{\sigma}(\boldsymbol{m}) = \sigma_i$ for some $i \in \{2,\dots,N\}$ and $j \in \{1,\dots,N_i\}$. For every $\boldsymbol{s} \in \boldsymbol{j}(\boldsymbol{m})$ and $\rho, \delta > 0$, where we recall that the mapping $\boldsymbol{j} : \mathcal{U}^{(0)} \to \mathcal{P}(\mathcal{V}^{(1)} \cup \{\boldsymbol{s}_1\})$ has been defined in (1-18) and that $V(\boldsymbol{s}) = \boldsymbol{\sigma}(\boldsymbol{m})$, we define the set

$$\mathcal{B}_{\boldsymbol{s},\rho,\delta} := \{V \leq \boldsymbol{\sigma}(\boldsymbol{m}) + \delta\} \cap \{x \in \mathbb{R}^d,\ |\xi(\boldsymbol{s}) \cdot (x - \boldsymbol{s})| \leq \rho\}$$

and the set $\mathcal{C}_{\boldsymbol{s},\rho,\delta}$ by

$$\mathcal{C}_{\boldsymbol{s},\rho,\delta} \quad \text{is the connected component of } \mathcal{B}_{\boldsymbol{s},\rho,\delta} \text{ containing } \boldsymbol{s}, \tag{4-1}$$

where $\xi(\boldsymbol{s})$ has been defined in Lemma 1.8. We recall that $\xi(\boldsymbol{s})$ is an unitary eigenvector of the matrix $\operatorname{Hess} V(\boldsymbol{s}) + B^*(\boldsymbol{s})$ associated with its only negative eigenvalue $\mu(\boldsymbol{s})$ which has geometric multiplicity one. Let us also define

$$E_{\boldsymbol{m},\rho,\delta} := (E_-(\boldsymbol{m}) \cap \{V < \boldsymbol{\sigma}(\boldsymbol{m}) + \delta\}) \setminus \bigcup_{\boldsymbol{s} \in \boldsymbol{j}(\boldsymbol{m})} \mathcal{C}_{\boldsymbol{s},\rho,\delta}, \tag{4-2}$$

where

$$E_-(\boldsymbol{m}) \text{ is the connected component of } \{V < \sigma_{i-1}\} \text{ containing } \boldsymbol{m}. \tag{4-3}$$

According to Assumption 4 and Remark 1.7, we recall that there is precisely one connected component $\widehat{E}(\boldsymbol{m}) \neq E(\boldsymbol{m})$ of $\{V < \boldsymbol{\sigma}(\boldsymbol{m})\}$ such that $\overline{E(\boldsymbol{m})} \cap \overline{\widehat{E}(\boldsymbol{m})} \neq \varnothing$ (see examples in Figure 3). Moreover, $\boldsymbol{j}(\boldsymbol{m}) = \partial \widehat{E}(\boldsymbol{m}) \cap \partial E(\boldsymbol{m})$ and the global minimum $\hat{\boldsymbol{m}}$ of $V|_{\widehat{E}(\boldsymbol{m})}$ satisfies $\boldsymbol{\sigma}(\hat{\boldsymbol{m}}) > \boldsymbol{\sigma}(\boldsymbol{m})$ and $V(\hat{\boldsymbol{m}}) < V(\boldsymbol{m})$ (see [Michel 2019], where the notation $\widehat{E}(\boldsymbol{m})$ is introduced for an arbitrary Morse function).

According to the geometry of the Morse function V around $\partial E(\boldsymbol{m})$ and to Lemma 1.8, we have:

Lemma 4.1. *Assume that Assumption 4 is satisfied and let $\boldsymbol{m} \in \mathcal{U}^{(0)} \setminus \{\underline{\boldsymbol{m}}\}$, $\boldsymbol{s} \in \boldsymbol{j}(\boldsymbol{m})$, and $\xi(\boldsymbol{s})$ be some unitary eigenvector of the matrix* Hess $V(\boldsymbol{s}) + B^*(\boldsymbol{s})$ *associated with its unique negative eigenvalue (see Lemma 1.8). Then, there exists a neigborhood $\mathcal{O}$ of $\boldsymbol{s}$ such that*

$$\forall x \in \mathcal{O} \setminus \{\boldsymbol{s}\}, \quad (x - \boldsymbol{s} \in \xi(\boldsymbol{s})^{\perp} \Rightarrow V(x) > V(\boldsymbol{s})).$$

It follows that there exist $\rho_0, \delta_0 > 0$ sufficiently small such that for all $\rho \in (0, \rho_0]$ and $\delta \in (0, \delta_0]$, the set $E_{\boldsymbol{m},3\rho,3\delta}$ defined in (4-2) has exactly two connected components, $E^{+}_{\boldsymbol{m},3\rho_0,3\delta_0}$ and $E^{-}_{\boldsymbol{m},3\rho,3\delta}$, containing respectively $\boldsymbol{m}$ and $\hat{\boldsymbol{m}}$.

Proof. For brevity, we denote $\xi = \xi(\boldsymbol{s})$. By a continuity argument, note that to prove the first part of Lemma 4.1, it is sufficient to prove that the linear hyperplane $\xi^{\perp}$ does not meet the cone $\{X \in \mathbb{R}^n;\ \langle \text{Hess}\, V(\boldsymbol{s})X, X\rangle \leq 0\}$ outside the origin. The second part of the lemma then simply follows from the observation that the set $\mathcal{C}_{\boldsymbol{s},\rho,\delta}$ defined in (4-1) is thus an arbitrarily small neighborhood of $\boldsymbol{s}$ when $\rho, \delta > 0$ tend to 0.

For $d \geq 3$, it is then enough to show that for any column vector $X \in \mathbb{R}^d \setminus \{0\}$ such that $\langle \text{Hess}\, V(\boldsymbol{s})X, X\rangle = 0$, it holds that $\text{Span} X \oplus \xi^{\perp} = \mathbb{R}^d$, i.e., $\langle X, \xi\rangle \neq 0$. Indeed, when $d \geq 3$, any linear hyperplane meets $\{X \in \mathbb{R}^n : \langle \text{Hess}\, V(\boldsymbol{s})X, X\rangle > 0\}$ and then meets $\{X \in \mathbb{R}^d \setminus \{0\} : \langle \text{Hess}\, V(\boldsymbol{s})X, X\rangle = 0\}$ if and only if it meets $\{X \in \mathbb{R}^d \setminus \{0\} : \langle \text{Hess}\, V(\boldsymbol{s})X, X\rangle \leq 0\}$. Let us then consider $X \in \mathbb{R}^d \setminus \{0\}$ such that $\langle \text{Hess}\, V(\boldsymbol{s})X, X\rangle = 0$ and let us prove that $\langle X, \xi\rangle \neq 0$. To show this, we work in orthonormal coordinates of $\mathbb{R}^d$ where Hess $V(\boldsymbol{s})$ is diagonal, i.e., where Hess $V(\boldsymbol{s}) = \text{Diag}(-\mu_1, \mu_2, \dots, \mu_d)$. It then follows from $\langle \text{Hess}\, V(\boldsymbol{s})X, X\rangle = 0$ and from the third item of Lemma 1.8 that

$$\mu_1 X_1^2 = \sum_{k=2}^{d} \mu_k X_k^2 \quad \text{and} \quad \frac{1}{\mu_1}\xi_1^2 > \sum_{k=2}^{d} \frac{1}{\mu_k}\xi_k^2 \geq 0.$$

In particular, $X_1 \neq 0$ and thus, by multiplying the two above relations,

$$|\xi_1 X_1| > \left(\sum_{k=2}^{d} \frac{1}{\mu_k}\xi_k^2\right)^{\frac{1}{2}} \left(\sum_{k=2}^{d} \mu_k X_k^2\right)^{\frac{1}{2}} \geq \left|\sum_{k=2}^{d} \xi_k X_k\right|,$$

the last inequality resulting from the Cauchy–Schwarz inequality. The relation $\langle X, \xi\rangle \neq 0$ follows.

When $d = 2$, the situation is slightly different because for any hyperplane H, either $H \setminus \{0\} \subset \{X \in \mathbb{R}^2 \setminus \{0\} : \langle \text{Hess}\, V(\boldsymbol{s})X, X\rangle \leq 0\}$ or $H \setminus \{0\} \subset \{X \in \mathbb{R}^2 \setminus \{0\} : \langle \text{Hess}\, V(\boldsymbol{s})X, X\rangle > 0\}$. Take again orthonormal coordinates where Hess $V(\boldsymbol{s}) = \text{Diag}(-\mu_1, \mu_2)$. We have then only to prove that the vector $\xi' := (-\xi_2, \xi_1)^*$, which spans $\xi^{\perp}$, satisfies

$$-\mu_1 \xi_2^2 + \mu_2 \xi_1^2 = \langle \text{Hess}\, V(\boldsymbol{s})\xi', \xi'\rangle > 0.$$

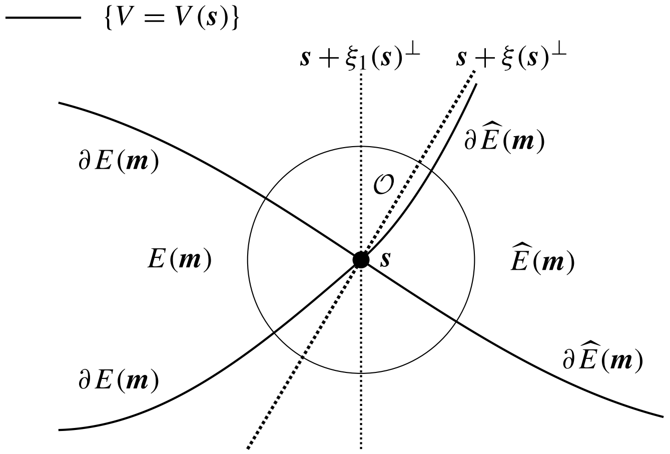


Figure 4. Representation of the Morse function V near $s \in j(m)$. Here, $\xi_1(s)$ denotes an eigenvector of Hess $V(s)$ associated with its negative eigenvalue and $B^*(s)\xi(s) \neq 0$. Note that according to the last item in Lemma 1.8, $s+\xi_1(s)^\perp$ and $s+\xi(s)^\perp$ coincide if and only if $B^*(s)\xi(s) = 0$.

This is obviously satisfied since it is equivalent to

$$0 > \frac{1}{\mu_2}\xi_2^2 - \frac{1}{\mu_2}\xi_1^2 = \langle (\text{Hess}\, V(s))^{-1}\xi, \xi\rangle,$$

which holds true thanks to (iii) of Lemma 1.8. This concludes the proof of Lemma 4.1. □

Let us now define, for every $h \in (0,1]$ and for every $\rho_0, \delta_0 > 0$ small enough, the function $\kappa_{m,h}$ on the sublevel set $E_-(m) \cap \{V < \sigma(m) + 3\delta_0\}$ (see (4-3)) as follows:

(1) On the disjoint open sets $E^+_{m,3\rho_0,3\delta_0}$ and $E^-_{m,3\rho_0,3\delta_0}$ introduced in Lemma 4.1,

$$\kappa_{m,h}(x) := \begin{cases} +1 & \text{for } x \in E^+_{m,3\rho_0,3\delta_0}, \\ -1 & \text{for } x \in E^-_{m,3\rho_0,3\delta_0}. \end{cases} \tag{4-4}$$

(2) For every $s \in j(m)$ and $x \in \mathcal{C}_{s,3\rho_0,3\delta_0} \cap \{V < \sigma(m) + 3\delta_0\}$ (see (4-1)),

$$\kappa_{m,h}(x) := C_{s,h}^{-1} \int_0^{\xi(s)\cdot(x-s)} \chi(\rho_0^{-1}\eta) e^{-\frac{|\mu(s)|\eta^2}{2h}}\, d\eta, \tag{4-5}$$

where the orientation of $\xi(s)$ is chosen in such a way that there exists a neighborhood $\mathcal{O}$ of s such that $E(m)\cap\mathcal{O}$ is included in the half-space $\{\xi(s)\cdot(x-s) > 0\}$ (see Lemma 4.1 and Figures 4 and 5), $\chi \in C^\infty(\mathbb{R}; [0,1])$ is even and satisfies $\chi \equiv 1$ on $[-1,1]$, $\chi(\eta) = 0$ for $|\eta| \geq 2$, and

$$C_{s,h} := \frac{1}{2}\int_{-\infty}^{+\infty} \chi(\rho_0^{-1}\eta) e^{-\frac{|\mu(s)|\eta^2}{2h}}\, d\eta.$$

Note in particular that

$$\text{there exists } \gamma > 0 \text{ s.t. } C_{s,h}^{-1} = \sqrt{\frac{2|\mu(s)|}{\pi h}}(1 + \mathcal{O}(e^{-\frac{\gamma}{h}})). \tag{4-6}$$

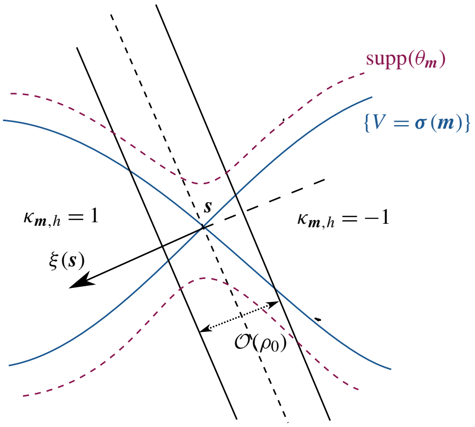

supp($\theta_{\boldsymbol{m}}$)
$\{V=\boldsymbol{\sigma}(\boldsymbol{m})\}$
$\kappa_{\boldsymbol{m},h}=1$
$\boldsymbol{s}$
$\kappa_{\boldsymbol{m},h}=-1$
$\xi(\boldsymbol{s})$
$\mathcal{O}(\rho_0)$

Figure 5. The support of the function $\kappa_{\boldsymbol{m},h}$.

Note also that for every $\rho_0,\delta_0>0$ small enough, thanks to the definitions (4-4) and (4-5), and since the sets $E^+_{\boldsymbol{m},3\rho_0,3\delta_0}$, $E^-_{\boldsymbol{m},3\rho_0,3\delta_0}$, and $\mathcal{C}_{\boldsymbol{s},3\rho_0,3\delta_0}$, $\boldsymbol{s}\in\boldsymbol{j}(\boldsymbol{m})$, are two by two disjoint (see Lemma 4.1), $\kappa_{\boldsymbol{m},h}$ is well defined and is $\mathcal{C}^\infty$ on $E_-(\boldsymbol{m})\cap\{V<\boldsymbol{\sigma}(\boldsymbol{m})+3\delta_0\}$.

Consider now a smooth function $\theta_{\boldsymbol{m}}$ such that

$$\theta_{\boldsymbol{m}}(x):=\begin{cases}1 & \text{for } x\in\{V\le\boldsymbol{\sigma}(\boldsymbol{m})+\frac{3}{2}\delta_0\}\cap E_-(\boldsymbol{m}),\\ 0 & \text{for } x\in\mathbb{R}^d\setminus(\{V<\boldsymbol{\sigma}(\boldsymbol{m})+2\delta_0\}\cap E_-(\boldsymbol{m})).\end{cases}\tag{4-7}$$

The function $\theta_{\boldsymbol{m}}\kappa_{\boldsymbol{m},h}$ then belongs to $\mathcal{C}^\infty_c(\mathbb{R}^d;[-1,1])$ and

$$\operatorname{supp}\theta_{\boldsymbol{m}}\kappa_{\boldsymbol{m},h}\subset E_-(\boldsymbol{m})\cap\{V<\boldsymbol{\sigma}(\boldsymbol{m})+2\delta_0\}.$$

Definition 4.2. For any $\boldsymbol{m}\in\mathcal{U}^{(0)}$ let us define the function $\psi_{\boldsymbol{m},h}$ by

$$\psi_{\boldsymbol{m},h}(x):=\theta_{\boldsymbol{m}}(x)(\kappa_{\boldsymbol{m},h}(x)+1)$$

when $\boldsymbol{m}\neq\underline{\boldsymbol{m}}$ and, when $\boldsymbol{m}=\underline{\boldsymbol{m}}$, $\psi_{\boldsymbol{m},h}(x):=1$. We then define, for any $\boldsymbol{m}\in\mathcal{U}^{(0)}$, the quasimode $\varphi_{\boldsymbol{m},h}$ by

$$\varphi_{\boldsymbol{m},h}(x):=\frac{\psi_{\boldsymbol{m},h}(x)}{\|\psi_{\boldsymbol{m},h}\|_{L^2(m_h)}}.$$

Note that, for every $h\in(0,1]$, it holds that $L_{V,b,\nu}\varphi_{\underline{\boldsymbol{m}},h}=0$ and for every $\boldsymbol{m}\in\mathcal{U}^{(0)}\setminus\{\underline{\boldsymbol{m}}\}$, the quasimodes $\psi_{\boldsymbol{m},h}$ and $\varphi_{\boldsymbol{m},h}$ belong to $\mathcal{C}^\infty_c(\mathbb{R}^d;\mathbb{R}^+)$ with supports included in $E_-(\boldsymbol{m})\cap\{V<\boldsymbol{\sigma}(\boldsymbol{m})+2\delta_0\}$. We have more precisely the following lemma resulting from the previous construction.

Lemma 4.3. *Assume that Assumption 4 is satisfied. For every $\boldsymbol{m}\in\mathcal{U}^{(0)}$ and every small $\epsilon>0$ fixed, there exist $\rho_0,\delta_0>0$ small enough such that for every $h\in(0,1]$:*

(i) $\operatorname{supp}\psi_{\boldsymbol{m},h}\subset\overline{E(\boldsymbol{m})}+B(0,\epsilon)$.

(ii) *When $\boldsymbol{m}\neq\underline{\boldsymbol{m}}$, there exists a neighborhood $\mathcal{O}_{\rho_0,\delta_0}$ of $\overline{E(\boldsymbol{m})}$ such that*

$$\mathcal{O}_{\rho_0,\delta_0}\setminus\bigcup_{\boldsymbol{s}\in\boldsymbol{j}(\boldsymbol{m})}\mathcal{C}_{\boldsymbol{s},3\rho_0,3\delta_0}\subset\{\theta_{\boldsymbol{m}}\kappa_{\boldsymbol{m},h}=1\}.$$

In particular,

$$\operatorname{argmin}_{\operatorname{supp}\psi_{\boldsymbol{m},h}} V = \operatorname{argmin}_{\{\theta_{\boldsymbol{m}}\kappa_{\boldsymbol{m},h}=1\}} V = \operatorname{argmin}_{E(\boldsymbol{m})} V = \{\boldsymbol{m}\}.$$

(iii) *When $\boldsymbol{m} \neq \underline{\boldsymbol{m}}$,*

$$\forall\, x \in \operatorname{supp} \nabla\psi_{\boldsymbol{m},h}, \qquad \Big(V(x) < \boldsymbol{\sigma}(\boldsymbol{m}) + \tfrac{3}{2}\delta_0 \Rightarrow x \in \bigcup_{\boldsymbol{s}\in \boldsymbol{j}(\boldsymbol{m})} \mathcal{C}_{\boldsymbol{s},3\rho_0,3\delta_0} \Big).$$

Let moreover $\boldsymbol{m}'$ belong to $\mathcal{U}^{(0)}$ with $\boldsymbol{m} \neq \boldsymbol{m}'$. The following then hold true for every $\rho_0, \delta_0 > 0$ small enough and every $h \in (0,1]$:

(iv) *If $\boldsymbol{\sigma}(\boldsymbol{m}) = \boldsymbol{\sigma}(\boldsymbol{m}')$, then* $\operatorname{supp}(\psi_{\boldsymbol{m},h}) \cap \operatorname{supp}(\psi_{\boldsymbol{m}',h}) = \varnothing$.

(v) *If $\boldsymbol{\sigma}(\boldsymbol{m}) > \boldsymbol{\sigma}(\boldsymbol{m}')$, then*

- *either* $\operatorname{supp}(\psi_{\boldsymbol{m},h}) \cap \operatorname{supp}(\psi_{\boldsymbol{m}',h}) = \varnothing$,
- *or* $\psi_{\boldsymbol{m},h} = 2$ *on* $\operatorname{supp}(\psi_{\boldsymbol{m}',h})$ *and* $V(\boldsymbol{m}') > V(\boldsymbol{m})$.

Proof. The first part of Lemma 4.3 follows from Assumption 4 and from the construction of the quasimodes $\varphi_{\boldsymbol{m},h}$ defined in Definition 4.2 for $\boldsymbol{m} \in \mathcal{U}^{(0)}$; see (4-4), (4-5), and (4-7). We now then prove the second part of Lemma 4.3.

When $\boldsymbol{\sigma}(\boldsymbol{m}) = \boldsymbol{\sigma}(\boldsymbol{m}')$ and $\boldsymbol{m} \neq \boldsymbol{m}'$, note first that $\boldsymbol{m}$ and $\boldsymbol{m}'$ differ from $\underline{\boldsymbol{m}}$ since $\boldsymbol{\sigma}(\boldsymbol{m}) = +\infty$ if and only if $\boldsymbol{m} = \underline{\boldsymbol{m}}$. When moreover $\boldsymbol{m}' \notin E_-(\boldsymbol{m})$, we have $E_-(\boldsymbol{m}) \neq E_-(\boldsymbol{m}')$ and hence $E_-(\boldsymbol{m}) \cap E_-(\boldsymbol{m}') = \varnothing$, implying $\operatorname{supp}(\psi_{\boldsymbol{m},h}) \cap \operatorname{supp}(\psi_{\boldsymbol{m}',h}) = \varnothing$. In the case when $\boldsymbol{m}' \in E_-(\boldsymbol{m})$, the statement of Lemma 4.3 follows from (ii) of Assumption 4 and Remark 1.7, which imply that $\overline{E(\boldsymbol{m})} \cap \overline{E(\boldsymbol{m}')} = \varnothing$ (see the first item of Lemma 4.3).

When $\boldsymbol{\sigma}(\boldsymbol{m}) > \boldsymbol{\sigma}(\boldsymbol{m}')$ and $\boldsymbol{m}' \notin E(\boldsymbol{m})$, we have $\overline{E(\boldsymbol{m})} \cap \overline{E(\boldsymbol{m}')} = \varnothing$, and again, according to the first item of Lemma 4.3, $\operatorname{supp}(\psi_{\boldsymbol{m},h}) \cap \operatorname{supp}(\psi_{\boldsymbol{m}',h}) = \varnothing$ for every $\rho_0, \delta_0 > 0$ small enough. Lastly, when $\boldsymbol{\sigma}(\boldsymbol{m}) > \boldsymbol{\sigma}(\boldsymbol{m}')$ and $\boldsymbol{m}' \in E(\boldsymbol{m})$, we have $\overline{E(\boldsymbol{m}')} \subset E_-(\boldsymbol{m}') \subset E(\boldsymbol{m})$ and then, according to the second item of Lemma 4.3, $\psi_{\boldsymbol{m},h} = 2$ on $\operatorname{supp}(\psi_{\boldsymbol{m}',h})$ for every $\rho_0, \delta_0 > 0$ small enough. Besides, the relation $V(\boldsymbol{m}') > V(\boldsymbol{m})$ follows from $\boldsymbol{m}' \in E(\boldsymbol{m})$ and from the first item of Assumption 4. □

4B. *Quasimodal estimates.* We write in the sequel $a \approx b$ and $a \lesssim b$ to mean, in the limit $h \to 0$, equality/inequality up to a multiplicative factor $1 + \mathcal{O}(h)$. Moreover, for brevity, we define, for any critical point $\boldsymbol{u}$ of V,

$$D_{\boldsymbol{u}} := \sqrt{|\det \operatorname{Hess} V(\boldsymbol{u})|} > 0.$$

Proposition 4.4. *Assume that Assumption 4 is satisfied and consider the families $(\psi_{\boldsymbol{m},h}, \boldsymbol{m} \in \mathcal{U}^{(0)})$ and $(\varphi_{\boldsymbol{m},h}, \boldsymbol{m} \in \mathcal{U}^{(0)})$ of Definition 4.2. Then, for every $\boldsymbol{m} \in \mathcal{U}^{(0)} \setminus \{\underline{\boldsymbol{m}}\}$ and $\rho_0, \delta_0 > 0$ small enough, in the limit $h \to 0$,*

$$\|\psi_{\boldsymbol{m},h}\|^2_{L^2(m_h)} \approx 4\frac{D_{\underline{\boldsymbol{m}}}}{D_{\boldsymbol{m}}} e^{-\frac{V(\boldsymbol{m})-V(\underline{\boldsymbol{m}})}{h}}. \tag{4-8}$$

Moreover, there exists $C > 0$ such that for every $\boldsymbol{m}, \boldsymbol{m}' \in \mathcal{U}^{(0)}$, in the limit $h \to 0$,

$$\langle \varphi_{\boldsymbol{m},h}, \varphi_{\boldsymbol{m}',h} \rangle = \delta_{\boldsymbol{m},\boldsymbol{m}'} + \mathcal{O}(e^{-\frac{c}{h}}). \tag{4-9}$$

Proof. To prove the relation (4-8), write, according to Definition 4.2,

$$\|\psi_{\boldsymbol{m},h}\|^2_{L^2(m_h)} = Z_h^{-1}\int(\theta_{\boldsymbol{m}}(\kappa_{\boldsymbol{m},h}+1))^2 e^{-\frac{V(x)}{h}}\,dx,$$

where Z_h is the normalizing constant defined by (1-10). Hence, according to Lemma 4.3 and standard tail estimates and Laplace asymptotics, we get, in the limit $h\to 0$,

$$Z_h \eqsim (2\pi h)^{\frac{d}{2}} D_{\underline{\boldsymbol{m}}}^{-1} e^{-\frac{V(\boldsymbol{m})}{h}}$$

as well as

$$\int(\theta_{\boldsymbol{m}}(\kappa_{\boldsymbol{m},h}+1))^2 e^{-\frac{V(x)}{h}}\,dx \eqsim 4(2\pi h)^{\frac{d}{2}} D_{\boldsymbol{m}}^{-1} e^{-\frac{V(\boldsymbol{m})}{h}}.$$

The estimate (4-8) then follows easily.

Let us now prove the relation (4-9). According to Definition 4.2, note first that $\langle\varphi_{\boldsymbol{m},h},\varphi_{\boldsymbol{m},h}\rangle = 1$ for every $\boldsymbol{m}\in\mathcal{U}^{(0)}$. Moreover, when $\boldsymbol{m},\boldsymbol{m}'\in\mathcal{U}^{(0)}$ and $\boldsymbol{m}\neq\boldsymbol{m}'$, it follows from Lemma 4.3 that, up to switching $\boldsymbol{m}$ and $\boldsymbol{m}'$, we are in one of the two following cases:

- either $\operatorname{supp}(\varphi_{\boldsymbol{m},h})\cap\operatorname{supp}(\varphi_{\boldsymbol{m}',h})=\varnothing$, and then
$$\langle\varphi_{\boldsymbol{m},h},\varphi_{\boldsymbol{m}',h}\rangle = 0,$$
- or $\psi_{\boldsymbol{m},h}=2$ on $\operatorname{supp}(\psi_{\boldsymbol{m}',h})$ and $V(\boldsymbol{m}')>V(\boldsymbol{m})$, and then, using the preceding estimates,
$$\begin{aligned}\langle\varphi_{\boldsymbol{m},h},\varphi_{\boldsymbol{m}',h}\rangle &= \frac{2}{\|\psi_{\boldsymbol{m},h}\|_{L^2(m_h)}}\int_{\operatorname{supp}\psi_{\boldsymbol{m}',h}}\frac{\psi_{\boldsymbol{m}',h}}{\|\psi_{\boldsymbol{m}',h}\|_{L^2(m_h)}}\frac{e^{-\frac{V(x)}{h}}}{Z_h}\,dx\\ &= \frac{1}{\|\psi_{\boldsymbol{m},h}\|_{L^2(m_h)}\|\psi_{\boldsymbol{m}',h}\|_{L^2(m_h)}}\mathcal{O}(e^{-\frac{V(\boldsymbol{m}')-V(\boldsymbol{m})}{h}}) = \mathcal{O}(e^{-\frac{C}{h}}),\end{aligned}$$
where $C=\frac{V(\boldsymbol{m}')-V(\boldsymbol{m})}{2}>0$.

This leads to (4-9). □

Proposition 4.5. *For every $\boldsymbol{m}\in\underline{\mathcal{U}}^{(0)}$ and $\rho_0,\delta_0>0$ small enough, in the limit $h\to 0$,*

$$\langle L_{V,b,\nu}\psi_{\boldsymbol{m},h},\psi_{\boldsymbol{m},h}\rangle_{L^2(m_h)} \eqsim \sum_{\boldsymbol{s}\in\boldsymbol{j}(\boldsymbol{m})}\frac{2|\mu(\boldsymbol{s})|}{\pi}\frac{D_{\boldsymbol{m}}}{D_{\boldsymbol{s}}}e^{-\frac{V(\boldsymbol{s})-V(\boldsymbol{m})}{h}} \tag{4-10}$$

and then

$$\langle L_{V,b,\nu}\varphi_{\boldsymbol{m},h},\varphi_{\boldsymbol{m},h}\rangle_{L^2(m_h)} \eqsim \sum_{\boldsymbol{s}\in\boldsymbol{j}(\boldsymbol{m})}\frac{|\mu(\boldsymbol{s})|}{2\pi}\frac{D_{\boldsymbol{m}}}{D_{\boldsymbol{s}}}e^{-\frac{V(\boldsymbol{s})-V(\boldsymbol{m})}{h}}. \tag{4-11}$$

Proof. Note first that thanks to (1-3), one has $\operatorname{div}(b_h m_h)=0$ and hence,

$$\forall u\in\mathcal{C}_c^\infty(\mathbb{R}^d;\mathbb{R}),\quad \langle b_h\cdot\nabla u,u\rangle_{L^2(m_h)} = -\frac{1}{2}\int u^2\operatorname{div}(b_h m_h)\,dx = 0.$$

Using this relation together with (1-4), (4-4)–(4-7), Definition 4.2, and Lemma 4.3, we get, in the limit $h \to 0$,

$$
\begin{aligned}
&\langle L_{V,b,\nu}\psi_{\boldsymbol{m},h}, \psi_{\boldsymbol{m},h}\rangle_{L^2(m_h)} \\
&\quad = \langle (-h\Delta + \nabla V \cdot \nabla)\psi_{\boldsymbol{m},h}, \psi_{\boldsymbol{m},h}\rangle_{L^2(m_h)} = Z_h^{-1} h \int |\nabla(\theta_{\boldsymbol{m}}(\kappa_{\boldsymbol{m},h}+1))|^2 e^{-\frac{V}{h}}\,dx \\
&\quad = Z_h^{-1} h \int \theta_{\boldsymbol{m}}^2 |\nabla \kappa_{\boldsymbol{m},h}|^2 e^{-\frac{V}{h}}\,dx + Z_h^{-1}\mathcal{O}(e^{-\frac{\sigma(\boldsymbol{m})+\delta_0}{h}}) \\
&\quad = Z_h^{-1}\mathcal{O}(e^{-\frac{\sigma(\boldsymbol{m})+\delta_0}{h}}) + Z_h^{-1} \sum_{\boldsymbol{s}\in \boldsymbol{j}(\boldsymbol{m})} C_{\boldsymbol{s},h}^{-2} h \int_{\mathcal{C}_{\boldsymbol{s},3\rho_0,3\delta_0}} \theta_{\boldsymbol{m}}^2(x)\chi^2(\rho_0^{-1}\xi\cdot(x-\boldsymbol{s}))e^{-\frac{|\mu|(\xi\cdot(x-\boldsymbol{s}))^2}{h}} e^{-\frac{V}{h}}\,dx,
\end{aligned}
\tag{4-12}
$$

where for short we denote $\xi = \xi(\boldsymbol{s})$ and $\mu = \mu(\boldsymbol{s})$. From the second item in Lemma 1.8 and the Taylor expansion of $V + |\mu|\langle \xi, \cdot - \boldsymbol{s}\rangle^2$ around $\boldsymbol{s} \in \boldsymbol{j}(\boldsymbol{m})$,

$$
V(x) + |\mu|(\xi\cdot(x-\boldsymbol{s}))^2 = V(\boldsymbol{s}) + \tfrac{1}{2}\langle \operatorname{Hess} V(\boldsymbol{s})(x-\boldsymbol{s}), x-\boldsymbol{s}\rangle + |\mu|\langle \xi\xi^*(x-\boldsymbol{s}), x-\boldsymbol{s}\rangle + \mathcal{O}(|x-\boldsymbol{s}|^3),
$$

so it is clear that for ρ_0 and δ_0 small enough, $V + |\mu|\langle \xi, \cdot - \boldsymbol{s}\rangle^2$ uniquely attains its minimal value in $\mathcal{C}_{\boldsymbol{s},3\rho_0,3\delta_0}$ at $\boldsymbol{s}$ since

$$
\nabla(V + |\mu|\langle \xi, \cdot - \boldsymbol{s}\rangle^2)(\boldsymbol{s}) = 0 \quad \text{and} \quad \operatorname{Hess}(V + |\mu|\langle \xi, \cdot - \boldsymbol{s}\rangle^2)(\boldsymbol{s}) = M_V.
$$

Moreover, using again the second item in Lemma 1.8 and a standard Laplace method, in the limit $h \to 0$, for every $\boldsymbol{s} \in \boldsymbol{j}(\boldsymbol{m})$,

$$
C_{\boldsymbol{s},h}^{-2} \int_{\mathcal{C}_{\boldsymbol{s},3\rho_0,3\delta_0}} \theta_{\boldsymbol{m}}^2 \chi^2(\rho_0^{-1}\langle \xi, \cdot - \boldsymbol{s}\rangle) e^{-\frac{|\mu|\langle \xi,\cdot - \boldsymbol{s}\rangle^2}{h}} e^{-\frac{V}{h}}\,dx \approx \frac{(2\pi h)^{\frac{d}{2}}}{C_{\boldsymbol{s},h}^2 D_{\boldsymbol{s}}} e^{-\frac{V(\boldsymbol{s})}{h}} \approx \frac{2(2\pi h)^{\frac{d}{2}}|\mu|}{\pi h D_{\boldsymbol{s}}} e^{-\frac{V(\boldsymbol{s})}{h}},
\tag{4-13}
$$

where we used (4-6) for the last part. The statement of Proposition 4.5 then follows from (4-12) and (4-13), using also $Z_h \approx (2\pi h)^{d/2} D_{\underline{\boldsymbol{m}}}^{-1} e^{-V(\underline{\boldsymbol{m}})/h}$. □

Proposition 4.6. *Let $\boldsymbol{m} \in \underline{\mathcal{U}}^{(0)}$. For ρ_0 and δ_0 sufficiently small, in the limit $h \to 0$,*

$$
\|L_{V,b,\nu}\psi_{\boldsymbol{m},h}\|_{L^2(m_h)}^2 = \langle L_{V,b,\nu}\psi_{\boldsymbol{m},h}, \psi_{\boldsymbol{m},h}\rangle_{L^2(m_h)}\mathcal{O}(h)
\tag{4-14}
$$

and

$$
\|L_{V,b,\nu}^*\psi_{\boldsymbol{m},h}\|_{L^2(m_h)}^2 = \langle L_{V,b,\nu}\psi_{\boldsymbol{m},h}, \psi_{\boldsymbol{m},h}\rangle_{L^2(m_h)}\mathcal{O}(1).
\tag{4-15}
$$

Proof. Let $\boldsymbol{s} \in \boldsymbol{j}(\boldsymbol{m})$ and denote for short $\xi = \xi(\boldsymbol{s})$ and $\mu = \mu(\boldsymbol{s})$. We first recall the Taylor expansion of $V + |\mu|\langle \xi, \cdot - \boldsymbol{s}\rangle^2$ around $\boldsymbol{s}$,

$$
V(x) + |\mu|(\xi\cdot(x-\boldsymbol{s}))^2 = V(\boldsymbol{s}) + \tfrac{1}{2}\langle M_V(x-\boldsymbol{s}), x-\boldsymbol{s}\rangle + \mathcal{O}(|x-\boldsymbol{s}|^3),
$$

which implies, according to the second item of Lemma 1.8, that for ρ_0 and δ_0 small enough,

- $\nabla(V + |\mu|\langle \xi, \cdot - \boldsymbol{s}\rangle^2)(\boldsymbol{s}) = 0$,
- $V + |\mu|\langle \xi, \cdot - \boldsymbol{s}\rangle^2$ uniquely attains its minimal value in $\mathcal{C}_{\boldsymbol{s},3\rho_0,3\delta_0}$ at $\boldsymbol{s}$.

Note now that according to (1-4),

$$L_{V,b,\nu}\psi_{\boldsymbol{m},h} = \theta_{\boldsymbol{m}} L_{V,h}\kappa_{\boldsymbol{m},h} + (1+\kappa_{\boldsymbol{m},h})L_{V,h}\theta_{\boldsymbol{m}} - 2h\nabla\kappa_{\boldsymbol{m},h}\cdot\nabla\theta_{\boldsymbol{m}},$$

and on $\mathcal{C}_{\boldsymbol{s},3\rho_0,3\delta_0}$, for every $\boldsymbol{s}\in\boldsymbol{j}(\boldsymbol{m})$, according to (4-5),

$$\begin{aligned} L_{V,b,\nu}\kappa_{\boldsymbol{m},h} &= -h\Delta\kappa_{\boldsymbol{m},h} + \nabla V\cdot\nabla\kappa_{\boldsymbol{m},h} + b_h\cdot\nabla\kappa_{\boldsymbol{m},h} \\ &= C_{\boldsymbol{s},h}^{-1}\chi(\rho_0^{-1}\langle\xi,\cdot-\boldsymbol{s}\rangle)e^{-\frac{|\mu|\langle\xi,\cdot-\boldsymbol{s}\rangle^2}{2h}}(\nabla V\cdot\xi + b_h\cdot\xi + |\mu|\langle\xi,\cdot-\boldsymbol{s}\rangle) \\ &\qquad -hC_{\boldsymbol{s},h}^{-1}\operatorname{div}(\chi(\rho_0^{-1}\langle\xi,\cdot-\boldsymbol{s}\rangle)\xi)e^{-\frac{|\mu|\langle\xi,\cdot-\boldsymbol{s}\rangle^2}{2h}}, \end{aligned}$$

where we recall that $b_h = b + h\nu$. It then follows from (4-4)–(4-7) that in the limit $h\to 0$,

$$\begin{aligned} \|L_{V,b,\nu}\psi_{\boldsymbol{m},h}\|^2_{L^2(m_h)} &= \sum_{\boldsymbol{s}\in\boldsymbol{j}(\boldsymbol{m})} \|\mathbf{1}_{\mathcal{C}_{\boldsymbol{s},3\rho_0,3\delta_0}} L_{V,b,\nu}\psi_{\boldsymbol{m},h}\|^2_{L^2(m_h)} + \frac{\mathcal{O}(e^{-\frac{\sigma(\boldsymbol{m})+\delta_0}{h}})}{Z_h} \\ &= \sum_{\boldsymbol{s}\in\boldsymbol{j}(\boldsymbol{m})} \frac{C_{\boldsymbol{s},h}^{-2}}{Z_h}\int_{\mathcal{C}_{\boldsymbol{s},3\rho_0,3\delta_0}} \chi^2(\rho_0^{-1}\xi\cdot(x-\boldsymbol{s}))e^{-\frac{V+|\mu|(\xi\cdot(x-\boldsymbol{s}))^2}{h}} \\ &\qquad\qquad \times(\nabla V\cdot\xi + b\cdot\xi + |\mu|\xi\cdot(x-\boldsymbol{s}) + h\nu\cdot\xi)^2\,dx + \frac{\mathcal{O}(e^{-\frac{\sigma(\boldsymbol{m})+c}{h}})}{Z_h} \end{aligned}$$

for some real constant $c\in(0,\delta_0)$. Moreover, using $b(\boldsymbol{s}) = 0$ and the first item of Lemma 1.8, the Taylor expansion of $\nabla V + b$ around $\boldsymbol{s}$ satisfies

$$\begin{aligned} (\nabla V + b)\cdot\xi + |\mu|\xi\cdot(x-\boldsymbol{s}) &= \langle(\operatorname{Hess} V(\boldsymbol{s}) + B)(x-\boldsymbol{s}),\xi\rangle + |\mu|\xi\cdot(x-\boldsymbol{s}) + \mathcal{O}((x-\boldsymbol{s})^2) \\ &= \mu\xi\cdot(x-\boldsymbol{s}) + |\mu|\xi\cdot(x-\boldsymbol{s}) + \mathcal{O}((x-\boldsymbol{s})^2) = \mathcal{O}((x-\boldsymbol{s})^2). \end{aligned}$$

Then from Proposition 4.5, standard tail estimates, and Laplace asymptotics, in the limit $h\to 0$,

$$\begin{aligned} \|L_{V,b,\nu}\psi_{\boldsymbol{m},h}\|^2_{L^2(m_h)} &= \sum_{\boldsymbol{s}\in\boldsymbol{j}(\boldsymbol{m})} \frac{C_{\boldsymbol{s},h}^{-2}}{Z_h}\int_{\mathcal{C}_{\boldsymbol{s},3\rho_0,3\delta_0}} \mathcal{O}((x-\boldsymbol{s})^4 + h^2)e^{-\frac{V+|\mu|(\xi\cdot(x-\boldsymbol{s}))^2}{h}}\,dx + \frac{\mathcal{O}(e^{-\frac{\sigma(\boldsymbol{m})+c}{h}})}{Z_h} \\ &= \langle L_{V,b,\nu}\psi_{\boldsymbol{m},h},\psi_{\boldsymbol{m},h}\rangle_{L^2(m_h)}\mathcal{O}(h), \end{aligned}$$

which proves (4-14).

To prove (4-15), we observe that since $L^*_{V,b,\nu} = L_{V,-b,-\nu}$, the same computation as above shows that in the limit $h\to 0$,

$$\begin{aligned} \|L^*_{V,b,\nu}\psi_{\boldsymbol{m},h}\|^2_{L^2(m_h)} &= \sum_{\boldsymbol{s}\in\boldsymbol{j}(\boldsymbol{m})} \frac{C_{\boldsymbol{s},h}^{-2}}{Z_h}\int_{\mathcal{C}_{\boldsymbol{s},3\rho_0,3\delta_0}} \chi^2(\rho_0^{-1}\xi\cdot(x-\boldsymbol{s}))e^{-\frac{V+|\mu|(\xi\cdot(x-\boldsymbol{s}))^2}{h}} \\ &\qquad\qquad \times(\nabla V\cdot\xi - b\cdot\xi + |\mu|\xi\cdot(x-\boldsymbol{s}) - h\nu\cdot\xi)^2\,dx + \frac{\mathcal{O}(e^{-\frac{\sigma(\boldsymbol{m})+c}{h}})}{Z_h}. \end{aligned}$$

However, contrary to the preceding case, one has here only

$$\nabla V\cdot\xi - b\cdot\xi + |\mu|\xi\cdot(x-\boldsymbol{s}) = \mathcal{O}(x-\boldsymbol{s}),$$

which implies, in the limit $h \to 0$,

$$\begin{aligned}\|L^*_{V,b,\nu}\psi_{\boldsymbol{m},h}\|^2_{L^2(m_h)} &= \sum_{\boldsymbol{s}\in\boldsymbol{j}(\boldsymbol{m})} \frac{C^{-2}_{\boldsymbol{s},h}}{Z_h} \int_{\mathcal{C}_{\boldsymbol{s},3\rho_0,3\delta_0}} \mathcal{O}((x-\boldsymbol{s})^2+h^2)e^{-\frac{V+|\mu|(\xi\cdot(x-\boldsymbol{s}))^2}{h}}\,dx + \frac{\mathcal{O}(e^{-\frac{V(\boldsymbol{s})+c}{h}})}{Z_h} \\ &= \langle L_{V,b,\nu}\psi_{\boldsymbol{m},h}, \psi_{\boldsymbol{m},h}\rangle_{L^2(m_h)}\mathcal{O}(1),\end{aligned}$$

which is exactly (4-15). □

4C. *Proof of Theorem 1.9.* Throughout this section, for ease of notation, we denote

$$\langle\,\cdot\,,\cdot\,\rangle = \langle\,\cdot\,,\cdot\,\rangle_{L^2(m_h)}, \qquad \|\cdot\| = \|\cdot\|_{L^2(m_h)}, \qquad L_{V,b,\nu} = L_V,$$

and we label the local minima $\boldsymbol{m}_1, \dots, \boldsymbol{m}_{n_0}$ of V in so that $(S(\boldsymbol{m}_j))_{j\in\{1,\dots,n_0\}}$ is nonincreasing (see (1-19)):

$$S(\boldsymbol{m}_1) = +\infty \qquad \text{and, for all } j \in \{2,\dots,n_0\}, \qquad S(\boldsymbol{m}_{j+1}) \le S(\boldsymbol{m}_j) < +\infty.$$

For all $j \in \{1,\dots,n_0\}$, we will also denote, for ease of notation,

$$S_j := S(\boldsymbol{m}_j), \quad \varphi_j := \varphi_{\boldsymbol{m}_j,h}, \quad \text{and} \quad \tilde{\lambda}_j(h) := \langle L_V\varphi_j, \varphi_j\rangle.$$

From Proposition 4.5, one knows that for all $j \in \{2,\dots,n_0\}$, one has

$$\tilde{\lambda}_j(h) = \sum_{\boldsymbol{s}\in\boldsymbol{j}(\boldsymbol{m}_j)} \frac{|\mu(\boldsymbol{s})|}{2\pi}\frac{D_{\boldsymbol{m}_j}}{D_{\boldsymbol{s}}} e^{-\frac{S_j}{h}}(1+\mathcal{O}(h)). \tag{4-16}$$

Moreover, since $(S_j)_{j\in\{1,\dots,n_0\}}$ is nonincreasing, we deduce from this estimate that there exist $h_0 > 0$ and $C > 0$ such that for all $h \in (0,h_0]$ and all $i, j \in \{1,\dots,n_0\}$, one has

$$i \le j \Rightarrow \lambda_i(h) \le C\lambda_j(h). \tag{4-17}$$

The two following lemmata are straightforward consequences of the previous analysis.

Lemma 4.7. *For every* $j, k \in \{1,\dots,n_0\}$ *and* $h \in (0,1]$, *one has*

$$\langle L_V\varphi_j, \varphi_k\rangle = \delta_{jk}\tilde{\lambda}_j(h).$$

Proof. When $j = k$, the statement is obvious. When $j \ne k$, it follows from Lemma 4.3 that we are in one of the three following cases:

- $\operatorname{supp}(\varphi_j) \cap \operatorname{supp}(\varphi_k) = \varnothing$ and the conclusion is obvious,
- There exists $c_h > 0$ such that $\varphi_j = c_h$ on $\operatorname{supp}(\varphi_k)$ and

$$\langle L_V\varphi_j, \varphi_k\rangle = \langle L_V(c_h), \varphi_k\rangle = 0,$$

- There exists $c_h > 0$ such that $\varphi_k = c_h$ on $\operatorname{supp}(\varphi_j)$ and

$$\langle L_V\varphi_j, \varphi_k\rangle = \langle \varphi_j, L^*_V\varphi_k\rangle = \langle \varphi_j, L^*_V(c_h)\rangle = 0.$$

□

Lemma 4.8. *For ρ_0, δ_0 sufficiently small and every $j \in \{1, \dots, n_0\}$, in the limit $h \to 0$,*

$$\|L_V \varphi_j\| = \mathcal{O}\big(\sqrt{h\tilde{\lambda}_j(h)}\big). \tag{4-18}$$

and

$$\|L_V^* \varphi_j\| = \mathcal{O}\big(\sqrt{\tilde{\lambda}_j(h)}\big). \tag{4-19}$$

Proof. This is a simple rewriting of Proposition 4.6, using the fact that for every $\boldsymbol{m} \in \mathcal{U}^{(0)}$ and $h \in (0, 1]$, $\varphi_{\boldsymbol{m},h} = \psi_{\boldsymbol{m},h}/\|\psi_{\boldsymbol{m},h}\|$. □

We now introduce, for every $h > 0$ small enough, the spectral projector Π_h associated with the n_0 smallest eigenvalues of L_V as described in Theorem 1.3. Let ϵ_0 be given by Theorem 1.3. According to Theorem 1.3, for every $h > 0$ small enough, Π_h satisfies

$$\Pi_h := \frac{1}{2i\pi} \int_{z \in \partial D(0, \frac{\epsilon_0}{2})} (z - L_V)^{-1}\, dz \tag{4-20}$$

and in particular,

$$\Pi_h = \mathcal{O}(1). \tag{4-21}$$

Lemma 4.9. *For all $j \in \{1, \dots, n_0\}$, we have, in the limit $h \to 0$,*

$$\|(1 - \Pi_h)\varphi_j\| = \mathcal{O}\big(\sqrt{h\tilde{\lambda}_j(h)}\big) \tag{4-22}$$

and

$$\|(1 - \Pi_h^*)\varphi_j\| = \mathcal{O}\big(\sqrt{\tilde{\lambda}_j(h)}\big) \tag{4-23}$$

Proof. Thanks to the resolvent identity, one has

$$\begin{aligned}(1 - \Pi_h)\varphi_j &= \frac{1}{2i\pi} \int_{z \in \partial D(0, \frac{\epsilon_0}{2})} (z^{-1} - (z - L_V)^{-1})\varphi_j\, dz \\ &= \frac{-1}{2i\pi} \int_{z \in \partial D(0, \frac{\epsilon_0}{2})} z^{-1}(z - L_V)^{-1} L_V \varphi_j\, dz.\end{aligned}$$

Moreover, it follows from Theorem 1.3 and from (1-14) that for any $z \in \partial D(0, \epsilon_0/2)$,

$$\|(z - L_V)^{-1}\|_{L^2(m_h) \to L^2(m_h)} = \mathcal{O}(1).$$

Combined with (4-18), this proves (4-22). On the other hand, one has similarly

$$(1 - \Pi_h^*)\varphi_j = \frac{-1}{2i\pi} \int_{z \in \partial D(0, \frac{\epsilon_0}{2})} z^{-1}(z - L_V^*)^{-1} L_V^* \varphi_j\, dz$$

and $\|(z - L_V^*)^{-1}\|_{L^2(m_h) \to L^2(m_h)} = \mathcal{O}(1)$. Then, (4-23) follows immediately from (4-19). □

Proposition 4.10. *For every $j \in \{1, \dots, n_0\}$ and $h > 0$ small enough, define $v_j := \Pi_h \varphi_j$. Then, there exists $c > 0$ such that for all $j, k \in \{1, \dots, n_0\}$, in the limit $h \to 0$,*

$$\langle v_j, v_k \rangle = \delta_{jk} + \mathcal{O}(e^{-\frac{c}{h}}) \tag{4-24}$$

and

$$\langle L_V v_j, v_k\rangle = \delta_{jk}\tilde{\lambda}_j(h) + \mathcal{O}\big(\sqrt{h\tilde{\lambda}_j(h)\tilde{\lambda}_k(h)}\,\big). \tag{4-25}$$

In particular, it follows from (4-24) *that for every* $h > 0$ *small enough, the family* $(v_1, \dots, v_{n_0})$ *is a basis of* $\operatorname{Ran}\Pi_h$.

Proof. Since, for some $c > 0$, every $j \in \{1, \dots, n_0\}$, and every $h > 0$ small enough, $\tilde{\lambda}_j(h) = \mathcal{O}(e^{-c/h})$, the first identity follows directly from (4-9), (4-22), and from the relation

$$\langle v_j, v_k\rangle = \langle \varphi_j, \varphi_k\rangle + \langle \varphi_j, v_k - \varphi_k\rangle + \langle v_j - \varphi_j, v_k\rangle.$$

To prove the second estimate, observe that

$$\begin{aligned}\langle L_V v_j, v_k\rangle &= \langle L_V \Pi_h\varphi_j, \Pi_h\varphi_k\rangle\\ &= \langle L_V\varphi_j, \varphi_k\rangle + \langle L_V(\Pi_h - 1)\varphi_j, \varphi_k\rangle + \langle \Pi_h L_V\varphi_j, (\Pi_h - 1)\varphi_k\rangle\\ &= \langle L_V\varphi_j, \varphi_k\rangle + \langle (\Pi_h - 1)\varphi_j, L_V^*\varphi_k\rangle + \langle \Pi_h L_V\varphi_j, (\Pi_h - 1)\varphi_k\rangle.\end{aligned}$$

Moreover, thanks to Lemma 4.8, (4-21), and Lemma 4.9, one has

$$|\langle (\Pi_h - 1)\varphi_j, L_V^*\varphi_k\rangle| \le \|(\Pi_h - 1)\varphi_j\|\|L_V^*\varphi_k\| = \mathcal{O}\big(\sqrt{h\tilde{\lambda}_j(h)\tilde{\lambda}_k(h)}\,\big)$$

and

$$|\langle \Pi_h L_V\varphi_j, (\Pi_h - 1)\varphi_k\rangle| \le \|\Pi_h\|\|L_V\varphi_j\|\|(\Pi_h - 1)\varphi_k\| = \mathcal{O}\big(\sqrt{h^2\tilde{\lambda}_j(h)\tilde{\lambda}_k(h)}\,\big).$$

Gathering these two estimates and using Lemma 4.7, we obtain (4-25). □

We now orthonormalize the basis $(v_1, \dots, v_{n_0})$ of $\operatorname{Ran}\Pi_h$ by a Gram–Schmidt procedure: for all $j \in \{1, \dots, n_0\}$, let us define by induction

$$\tilde{e}_j = v_j - \sum_{k=1}^{j-1} \frac{\langle v_j, \tilde{e}_k\rangle}{\|\tilde{e}_k\|^2}\tilde{e}_k \quad \text{and then} \quad e_j = \frac{\tilde{e}_j}{\|\tilde{e}_j\|}. \tag{4-26}$$

Lemma 4.11. *There exists* $c > 0$ *such that for all* $j \in \{1, \dots, n_0\}$, *in the limit* $h \to 0$,

$$\tilde{e}_j = v_j + \sum_{k=1}^{j-1} \alpha_{j,k} v_k$$

with $\alpha_{jk} = \mathcal{O}(e^{-c/h})$. *In particular,*

$$\forall\, j \in \{1, \dots, n_0\}, \quad \|\tilde{e}_j\| = 1 + \mathcal{O}(e^{-\frac{c}{h}}).$$

Proof. One proceeds by induction on j. For $j = 1$, one has $\tilde{e}_1 = v_1 = \varphi_1 = 1$ and there is nothing to prove. Suppose now that the above formula is true for all $\tilde{e}_l$ with $1 \le l \le j < n_0$. Then $\tilde{e}_{j+1} = v_{j+1} - r_{j+1}$ with

$$r_{j+1} = \sum_{k=1}^{j} \frac{\langle v_{j+1}, \tilde{e}_k\rangle}{\|\tilde{e}_k\|^2}\tilde{e}_k.$$

Since by induction, $\|\tilde{e}_k\| = 1 + \mathcal{O}(e^{-c/h})$ for all $k \in \{1, \dots, j\}$, it follows that

$$r_{j+1} = (1 + \mathcal{O}(e^{-\frac{c}{h}})) \sum_{k=1}^{j} \langle v_{j+1}, \tilde{e}_k \rangle \tilde{e}_k.$$

Moreover, for all $k \in \{1, \dots, j\}$, one also has by induction,

$$\tilde{e}_k = v_k + \sum_{l=1}^{k-1} \alpha_{k,l} v_l = \sum_{l=1}^{k} \beta_{k,l} v_k$$

with $\beta_{k,l} = \mathcal{O}(1)$ for any $l \in \{1, \dots, k\}$ (and actually $\beta_{k,l} = \mathcal{O}(e^{-c/h})$ when $l < k$), which implies

$$r_{j+1} = (1 + \mathcal{O}(e^{-\frac{c}{h}})) \sum_{k=1}^{j} \sum_{l,m=1}^{k} \beta_{k,l} \beta_{k,m} \langle v_{j+1}, v_l \rangle v_m.$$

Since, thanks to Proposition 4.10, $\langle v_{j+1}, v_l \rangle = \mathcal{O}(e^{-c/h})$ for all $l, m \leq k < j+1$,

$$r_{j+1} = \sum_{m=1}^{j} \gamma_{j,m} v_m,$$

where $\gamma_{j,m} = \mathcal{O}(e^{-c/h})$ for all $m \in \{1, \dots, j\}$. This proves the first part of the lemma. The second part is obvious. □

Proposition 4.12. *For all $j, k \in \{1, \dots, n_0\}$, in the limit $h \to 0$,*

$$\langle L_V e_j, e_k \rangle = \delta_{jk} \tilde{\lambda}_j(h) + \mathcal{O}\big(\sqrt{h \tilde{\lambda}_j(h) \tilde{\lambda}_k(h)}\big).$$

Proof. Thanks to Lemma 4.11, for all $j, k \in \{1, \dots, n_0\}$,

$$\langle L_V \tilde{e}_j, \tilde{e}_k \rangle = \langle L_V v_j, v_k \rangle + \sum_{p=1}^{j-1} \sum_{q=1}^{k-1} \alpha'_{p,q} \langle L_V v_p, v_q \rangle,$$

where $\alpha'_{p,q} = \alpha_{j,p} \alpha_{k,q} = \mathcal{O}(e^{-c/h})$, for all p, q. Combined with Proposition 4.10, this implies

$$\langle L_V \tilde{e}_j, \tilde{e}_k \rangle = \delta_{jk} \tilde{\lambda}_j(h) + \mathcal{O}\big(\sqrt{h \tilde{\lambda}_j(h) \tilde{\lambda}_k(h)}\big) + \sum_{p=1}^{j-1} \sum_{q=1}^{k-1} \alpha'_{p,q} \langle L_V v_p, v_q \rangle. \tag{4-27}$$

On the other hand, thanks to Proposition 4.10 and (4-17), one has in the limit $h \to 0$, for all $1 \leq p < j$ and $1 \leq q < k$,

$$\begin{aligned} \langle L_V v_p, v_q \rangle = \delta_{pq} \tilde{\lambda}_p(h) + \mathcal{O}\big(\sqrt{h \tilde{\lambda}_p(h) \tilde{\lambda}_q(h)}\big) &= \mathcal{O}\big(\sqrt{\tilde{\lambda}_p(h) \tilde{\lambda}_q(h)}\big) \\ &= \mathcal{O}\big(\sqrt{\tilde{\lambda}_j(h) \tilde{\lambda}_k(h)}\big). \end{aligned}$$

Combined with (4-27) and using the fact that $\alpha'_{p,q} = \mathcal{O}(e^{-c/h}) = \mathcal{O}(\sqrt{h})$, this shows that

$$\langle L_V \tilde{e}_j, \tilde{e}_k \rangle = \delta_{jk} \tilde{\lambda}_j(h) + \mathcal{O}\big(\sqrt{h \tilde{\lambda}_j(h) \tilde{\lambda}_k(h)}\big).$$

Finally, since $e_k = (1+\mathcal{O}(e^{-c/h}))\tilde{e}_k$ according to Lemma 4.11, we obtain

$$\langle L_V e_j, e_k\rangle = (1+\mathcal{O}(e^{-\frac{c}{h}}))\langle L_V \tilde{e}_j, \tilde{e}_k\rangle = \delta_{jk}\tilde{\lambda}_j(h) + \mathcal{O}\big(\sqrt{h\tilde{\lambda}_j(h)\tilde{\lambda}_k(h)}\,\big),$$

which completes the proof. □

We are now in position to prove Theorem 1.9. We recall that $(e_1, \dots, e_{n_0})$ is an orthonormal basis of $\operatorname{Ran}\Pi_h$ and that $L_V|_{\operatorname{Ran}\Pi_h} : \operatorname{Ran}\Pi_h \to \operatorname{Ran}\Pi_h$ has exactly n_0 eigenvalues $\lambda_1, \dots, \lambda_{n_0}$, with $\lambda_j = 0$ if and only if $j = 1$, counted with algebraic multiplicity. Let us denote $\hat{e}_j = e_{n_0+1-j}$ and let $\mathcal{M}$ denote the matrix of L_V in the basis $(\hat{e}_1, \dots, \hat{e}_{n_0})$. Since this basis is orthonormal,

$$\mathcal{M} = (\langle L_V \hat{e}_k, \hat{e}_j\rangle)_{j,k\in\{1,\dots,n_0\}}.$$

Moreover, since

$$L_V(\hat{e}_{n_0}) = L_V(e_1) = 0 \quad \text{and} \quad L_V^*(\hat{e}_{n_0}) = 0,$$

$\mathcal{M}$ has the form

$$\mathcal{M} = \begin{pmatrix} \mathcal{M}' & 0 \\ 0 & 0 \end{pmatrix} \quad \text{with } \mathcal{M}' := (\langle L_V \hat{e}_k, \hat{e}_j\rangle)_{j,k\in\{1,\dots n_0-1\}}.$$

On the other hand, denoting $\hat{\lambda}_j(h) := \tilde{\lambda}_{n_0+1-j}(h)$ for $j \in \{1, \dots, n_0-1\}$, one deduces from Proposition 4.12 that for every $j, k \in \{1, \dots n_0 - 1\}$, in the limit $h \to 0$,

$$\langle L_V \hat{e}_k, \hat{e}_j\rangle = \langle L_V e_{n_0-k}, e_{n_0-j}\rangle = \delta_{jk}\hat{\lambda}_j(h) + \mathcal{O}\big(\sqrt{h\hat{\lambda}_j(h)\hat{\lambda}_k(h)}\,\big);$$

that is,

$$\langle L_V \hat{e}_k, \hat{e}_j\rangle = \sqrt{\hat{\lambda}_j(h)\hat{\lambda}_k(h)}\big(\delta_{jk} + \mathcal{O}\big(\sqrt{h}\,\big)\big). \tag{4-28}$$

For all $j \in \{1, \dots, n_0 - 1\}$, let us now define

$$\hat{S}_j := S_{n_0+1-j} \quad \text{and} \quad \nu_j := \zeta(\boldsymbol{m}_{n_0+1-j}) = \sum_{\boldsymbol{s}\in \boldsymbol{j}(\boldsymbol{m}_{n_0+1-j})} \frac{|\mu(\boldsymbol{s})|}{2\pi}\frac{D_{\boldsymbol{m}_{n_0+1-j}}}{D_{\boldsymbol{s}}}$$
$$= e^{\frac{\hat{S}_j}{h}}\hat{\lambda}_j(h)(1+\mathcal{O}(h)),$$

where $\zeta(\boldsymbol{m})$, $\boldsymbol{m} \in \mathcal{U}^{(0)} \setminus \{\underline{\boldsymbol{m}}\}$, is defined in (1-21), and the last estimate follows from (4-16). Since the sequence $(S_j)_{j\in\{2,\dots,n_0\}}$ is nonincreasing, there exists a partition $J_1 \sqcup \ldots \sqcup J_p$ of $\{1, \dots, n_0 - 1\}$ such that for all $k \in \{1, \dots, p\}$, there exists $\iota(k) \in \{1, \dots, n_0 - 1\}$ such that

$$\forall\, j \in J_k, \quad \hat{S}_j = \hat{S}_{\iota(k)} \quad \text{and} \quad \forall\, 1 \le k < k' \le p, \quad \hat{S}_{\iota(k)} < \hat{S}_{\iota(k')}. \tag{4-29}$$

Hence, we deduce from (4-28) that

$$\mathcal{M}' = \widehat{\Omega}\big(\mathcal{J} + \big(\sqrt{h}\,\big)\big)\widehat{\Omega}$$

with

$$\mathcal{J} = \operatorname{diag}(\nu_j,\ j = 1, \dots, n_0 - 1)$$

and

$$\widehat{\Omega} = \operatorname{diag}(e^{-\frac{\hat{S}_j}{2h}},\ j = 1, \dots, n_0 - 1) = \operatorname{diag}(e^{-\frac{\hat{S}_{\iota(k)}}{2h}} I_{r_k},\ k = 1, \dots, p),$$

where, for every $k \in \{1, \dots, p\}$, $r_k = \operatorname{card}(J_k)$. Factorizing by $e^{-\hat{S}_{\iota(1)}/h}$, we get

$$\mathcal{M}' = e^{-\frac{\hat{S}_{\iota(1)}}{h}} \Omega\big(\mathcal{J} + \mathcal{O}(\sqrt{h}\,)\big)\Omega$$

with

$$\Omega = \operatorname{diag}(e^{\frac{\hat{S}_{\iota(1)} - \hat{S}_{\iota(k)}}{2h}} I_{r_k}, k = 1, \dots, p).$$

Denoting $\tau_1 = 1$ and, for $k \in \{2, \dots, p\}$, $\tau_k = e^{(\hat{S}_{\iota(k-1)} - \hat{S}_{\iota(k)})/(2h)}$, we observe that, thanks to (4-29), τ_k is exponentially small when $h \to 0$. Moreover, with this notation, one has

$$\Omega = \operatorname{diag}(\tau_1 I_{r_1}, \tau_1\tau_2 I_{r_2}, \dots, (\Pi_{j=1}^{p} \tau_j) I_{r_p}).$$

This shows that $e^{-\hat{S}_{\iota(1)}/h}\mathcal{M}'$ is a graded matrix in the sense of Definition A.1. Hence, we can apply Theorem A.4 and we get that in the limit $h \to 0$,

$$\sigma(\mathcal{M}') \subset \bigsqcup_{k=1}^{p} e^{-\frac{\hat{S}_{\iota(1)}}{h}} \varepsilon_k^2 \big(\sigma(M_k) + \mathcal{O}(\sqrt{h}\,)\big),$$

where for every $k \in \{1, \dots, p\}$, $\varepsilon_k = \prod_{l=1}^{k} \tau_l$ and $M_k = \operatorname{diag}(\nu_j, j \in J_k)$. Moreover, still according to Theorem A.4, $\mathcal{M}'$ admits in the limit $h \to 0$, for every $k \in \{1, \dots, p\}$ and every eigenvalue λ of M_k with multiplicity r'_k, exactly r'_k eigenvalues counted with multiplicity of order $e^{-\hat{S}_{\iota(1)}/h}\varepsilon_k^2\big(\lambda + \mathcal{O}(\sqrt{h}\,)\big)$.

Going back to the initial parameters, one has, for every $k \in \{1, \dots, p\}$,

$$e^{-\frac{\hat{S}_{\iota(1)}}{h}} \varepsilon_k^2 = e^{-\frac{\hat{S}_{\iota(k)}}{h}} \quad \text{and} \quad \sigma(M_k) = \{\nu_j, j \in J_k\}.$$

Hence, the eigenvalues of $\mathcal{M}'$ satisfy

$$\forall\, j \in \{1, \dots, n_0 - 1\}, \quad \lambda_{n_0+1-j}(h) = e^{-\frac{\hat{S}_j}{h}} \big(\nu_j + \mathcal{O}(\sqrt{h}\,)\big),$$

which is exactly the announced result.

4D. ***Proof of Theorem 1.11.*** As in the preceding subsection, for brevity we denote

$$\langle\,\cdot\,,\cdot\,\rangle = \langle\,\cdot\,,\cdot\,\rangle_{L^2(m_h)}, \qquad \|\cdot\| = \|\cdot\|_{L^2(m_h)}, \qquad L_{V,b,\nu} = L_V,$$

and we label the local minima $\boldsymbol{m}_1, \dots, \boldsymbol{m}_{n_0}$ of V so that $(S(\boldsymbol{m}_j))_{j \in \{1,\dots,n_0\}}$ is nonincreasing (see (1-19)):

$$S(\boldsymbol{m}_1) = +\infty \quad \text{and, for all } j \in \{2, \dots, n_0\}, \quad S(\boldsymbol{m}_{j+1}) \leq S(\boldsymbol{m}_j) < +\infty.$$

Let moreover $\boldsymbol{m}^* \in \mathcal{U}^{(0)} \setminus \{\underline{\boldsymbol{m}}\}$ be such that

$$S(\boldsymbol{m}^*) = S(\boldsymbol{m}_2) \quad \text{and} \quad \zeta(\boldsymbol{m}^*) = \min_{\boldsymbol{m} \in S^{-1}(S(\boldsymbol{m}_2))} \zeta(\boldsymbol{m}), \tag{4-30}$$

where the prefactors $\zeta(\boldsymbol{m})$, $\boldsymbol{m} \in \mathcal{U}^{(0)} \setminus \{\underline{\boldsymbol{m}}\}$, are defined in (1-21), and let us define, for any $h > 0$,

$$\lambda(h) := \zeta(\boldsymbol{m}^*) e^{-\frac{S(\boldsymbol{m}^*)}{h}}.$$

According to the unitary equivalence (see (1-14))

$$L_V = \frac{1}{h} \mathsf{U} P_\phi \mathsf{U}^*,$$

and to the localization of the spectrum of P_ϕ stated in Proposition 1.1 and in Theorem 1.3, for every $h > 0$ small enough, taking ϵ_0 as in the statement of Theorem 1.3,

$$\|e^{-tL_V} - \Pi_0\| \le \|e^{-tL_V}\Pi_h - \Pi_0\| + \|e^{-tL_V}(\mathrm{Id} - \Pi_h)\|, \tag{4-31}$$

where, as in the preceding subsection,

$$\Pi_h := \frac{1}{2i\pi}\int_{z\in\partial D(0,\frac{\epsilon_0}{2})} (z - L_V)^{-1}\,dz.$$

Moreover, it follows from Proposition 1.1 that $\sigma(P_\phi) \subset \Gamma_{\Lambda_0} \subset \tilde{\Gamma}_{\Lambda_0}$ with

$$\tilde{\Gamma}_{\Lambda_0} = \{z \in \mathbb{C},\ |\mathrm{Im}(z)| \le \Lambda_0(\mathrm{Re}(z) + 1)\}.$$

Hence, for every $t > 0$, the operator $e^{-tL_V}(I - \Pi_h)$ can be written as the complex integral

$$e^{-tL_V}(\mathrm{Id} - \Pi_h) = -\int_{\Gamma_0\cup\Gamma_\pm} e^{-tz}(z - L_V)^{-1}\,dz,$$

where

$$\Gamma_0 = \left\{\frac{\epsilon_0}{2} + i\Lambda_0 x,\ x \in \left[-\frac{\epsilon_0}{2} - \frac{1}{h}, \frac{\epsilon_0}{2} + \frac{1}{h}\right]\right\}$$

and

$$\Gamma_\pm = \left\{x \pm i\Lambda_0\left(x + \frac{1}{h}\right),\ x \in \left[\frac{\epsilon_0}{2}, +\infty\right)\right\}.$$

From the resolvent estimates proven in Theorem 1.3, $(z - L_V)^{-1} = \mathcal{O}(1)$ uniformly on Γ_0, and then, for every $t > 0$,

$$\int_{\Gamma_0} e^{-tz}(z - L_V)^{-1}\,dz = e^{-t\frac{\epsilon_0}{2}}\mathcal{O}\left(\frac{1}{h}\right).$$

Using in addition the resolvent estimates proven in Proposition 1.1, $\|(z - L_V)^{-1}\| \le 1/\mathrm{Re}\,z \le 2/\epsilon_0$ on $\Gamma_\pm$, and then

$$\int_{\Gamma_\pm} e^{-tz}(z - L_V)^{-1}\,dz = \mathcal{O}(1)\int_{\frac{\epsilon_0}{2}}^{+\infty} e^{-tx}\,dx = \frac{e^{-t\frac{\epsilon_0}{2}}}{t}\mathcal{O}(1).$$

It follows that for every $t > 0$,

$$\|e^{-tL_V}(\mathrm{Id} - \Pi_h)\| = e^{-t\frac{\epsilon_0}{2}}\mathcal{O}\left(\frac{1}{t} + \frac{1}{h}\right).$$

Moreover, $e^{-tL_V}(\mathrm{Id} - \Pi_h) = \mathcal{O}(1)$ since $\Pi_h = \mathcal{O}(1)$ (see (4-21)) and $e^{-tL_V} = \mathcal{O}(1)$ (by maximal accretivity of L_V). Hence, there exists $C > 0$ such that for every $t \ge 0$ and $h > 0$ small enough,

$$\|e^{-tL_V}(I - \Pi_h)\| \le C\min\left\{1, \frac{e^{-t\frac{\epsilon_0}{2}}}{h}\right\} \le 2Ce^{-\lambda(h)t}.$$

Thus, according to (4-31), it just remains to show that

$$\text{there exists } C > 0 \text{ such that } \|e^{-tL_V}\Pi_h - \Pi_0\| \le Ce^{-(\lambda(h) - C\sqrt{h})t}. \tag{4-32}$$

To this end, let us first recall from Proposition 1.1 that the spectral projector $\Pi_{\{0\}}$ associated with the eigenvalue 0 of L_V has rank 1 and is actually the orthogonal projector Π_0 on Span$\{1\}$ according to the relations

$$\mathrm{Span}\{1\} = \mathrm{Im}\,\Pi_{\{0\}} = \mathrm{Im}\,\Pi^*_{\{0\}} = (\mathrm{Ker}\,\Pi_{\{0\}})^{\perp}.$$

It follows that

$$e^{-tL_V}\Pi_h - \Pi_0 = e^{-tL_V}(\Pi_h - \Pi_{\{0\}}).$$

Since moreover $\Pi_h - \Pi_{\{0\}} = \mathcal{O}(1)$ (thanks to the resolvent estimate of Theorem 1.3), it suffices to show that

$$\text{there exists } C > 0 \text{ such that } \|e^{-tL_V}(\Pi_h - \Pi_{\{0\}})|_{\mathrm{Ran}(\Pi_h - \Pi_{\{0\}})}\| \le Ce^{-(\lambda(h) - C\sqrt{h})t}.$$

Using the notation of the preceding subsection, this means proving that the matrix $\mathcal{M}'$ of L_V in the orthonormal basis $(\hat{e}_1, \dots, \hat{e}_{n_0-1})$ of $\mathrm{Ran}(\Pi_h - \Pi_0)$ is such that

$$\text{there exists } C > 0 \text{ such that } \|e^{-t\mathcal{M}'}\| \le Ce^{-(\lambda(h) - C\sqrt{h})t}.$$

Let us now consider a subset $\mathcal{V}^{(0)}$ (in general nonunique) of $\mathcal{U}^{(0)} \setminus \{0\}$ such that

$$\boldsymbol{m} \in \mathcal{V}^{(0)} \mapsto (\zeta(\boldsymbol{m}), S(\boldsymbol{m})) \in \{(\zeta(\boldsymbol{m}), S(\boldsymbol{m})), \boldsymbol{m} \in \mathcal{U}^{(0)} \setminus \{0\}\} \quad \text{is a bijection.}$$

Then, for any $K > 0$ and for every $h > 0$ small enough, the closed disks of the complex plane

$$D_{\boldsymbol{m},K} := D(\zeta(\boldsymbol{m})e^{-\frac{S(\boldsymbol{m})}{h}}, K\sqrt{h}e^{-\frac{S(\boldsymbol{m})}{h}}), \quad \boldsymbol{m} \in \mathcal{V}^{(0)},$$

are included in $\{\mathrm{Re}\, z > 0\}$ and two by two disjoint. Moreover, according to Theorem 1.9, $K > 0$ can be chosen large enough so that when $h > 0$ is small enough, the $n_0 - 1$ nonzero small eigenvalues of L_V are included in

$$\bigcup_{\boldsymbol{m} \in \mathcal{V}^{(0)}} D\Big(\zeta(\boldsymbol{m})e^{-\frac{S(\boldsymbol{m})}{h}}, \frac{K}{2}\sqrt{h}e^{-\frac{S(\boldsymbol{m})}{h}}\Big).$$

In particular, for every $t \ge 0$ and for every $h > 0$ small enough,

$$e^{-t\mathcal{M}'} = \sum_{\boldsymbol{m} \in \mathcal{V}^{(0)}} \frac{1}{2i\pi} \int_{z \in \partial D_{\boldsymbol{m},K}} e^{-tz}(z - \mathcal{M}')^{-1}\, dz.$$

Using now the specific form of $\mathcal{M}'$ exhibited in the preceding section and Theorem A.4, for every $\boldsymbol{m} \in \mathcal{V}^{(0)}$, in the limit $h \to 0$,

$$\frac{1}{2i\pi} \int_{z \in \partial D_{\boldsymbol{m},K}} e^{-tz}(z - \mathcal{M}')^{-1}\, dz = \mathcal{O}(e^{-t\zeta(\boldsymbol{m})e^{-\frac{S(\boldsymbol{m})}{h}}(1 - K\sqrt{h})}).$$

Indeed, the resolvent estimate of Theorem A.4 implies

$$\forall z \in \partial D_{\boldsymbol{m},K} \quad , \|(\mathcal{M}' - z)^{-1}\| = \mathcal{O}(\mathrm{dist}(z, \sigma(\mathcal{M}'))^{-1}) = \mathcal{O}\Big(\frac{1}{\sqrt{h}}e^{\frac{S(\boldsymbol{m})}{h}}\Big). \tag{4-33}$$

The relation (4-32) follows easily, which concludes the first part of Theorem 1.11.

Finally, let us assume that the element $\boldsymbol{m}^*$ satisfying (4-30) is unique. In this case, $\boldsymbol{m}^*$ necessarily belongs to $\mathcal{V}^{(0)}$ and the associated eigenvalue $\lambda(\boldsymbol{m}^*, h)$ (see (1-20)) is then real and simple for every $h > 0$ small enough. In particular,

$$\frac{1}{2i\pi}\int_{z\in\partial D_{\boldsymbol{m}^*,K}} e^{-tz}(z-\mathcal{M}')^{-1}\,dz = e^{-t\lambda(\boldsymbol{m}^*,h)}\Pi_{\{\lambda(\boldsymbol{m}^*,h)\}},$$

where $\Pi_{\{\lambda(\boldsymbol{m}^*,h)\}}$ is the spectral projector (whose rank is one)

$$\Pi_{\{\lambda(\boldsymbol{m}^*,h)\}} = \frac{1}{2i\pi}\int_{z\in\partial D_{\boldsymbol{m}^*,K}} (z-\mathcal{M}')^{-1}\,dz.$$

Further, the resolvent estimate (4-33) shows $\Pi_{\{\lambda(\boldsymbol{m}^*,h)\}} = \mathcal{O}(1)$. Since in addition, in this case (see (1-20)),

$$\forall\, \boldsymbol{m} \in \mathcal{V}^{(0)}\setminus\{\boldsymbol{m}^*\},\qquad \lambda(\boldsymbol{m}^*,h) = \zeta(\boldsymbol{m}^*)e^{-\frac{S(\boldsymbol{m}^*)}{h}}(1+\mathcal{O}(\sqrt{h})) \geq \zeta(\boldsymbol{m})e^{-\frac{S(\boldsymbol{m})}{h}}(1-K\sqrt{h})$$

for every $K > 0$ and for every $h > 0$ small enough, we obtain that in the limit $h \to 0$,

$$e^{-t\mathcal{M}'} = \mathcal{O}(e^{-t\lambda(\boldsymbol{m}^*,h)}),$$

and thus the relation (4-32) remains valid if one replaces $\lambda(h) - C\sqrt{h}$ there by $\lambda(\boldsymbol{m}^*, h)$. This concludes the proof of Theorem 1.11.

Appendix: Some results in linear algebra

The aim of this appendix is to give some handy tools of linear algebra adapted to the setting of nonreversible metastable problems considered in this paper. Let us start with some notation.

Given any matrix $M \in \mathcal{M}_d(\mathbb{C})$ and $\lambda \in \sigma(M)$, we denote by $m(\lambda)$ the multiplicity of λ, $m(\lambda) = \dim\operatorname{Ker}(M-\lambda)^d$. We recall that for every $r > 0$ small enough,

$$m(\lambda) = \operatorname{rank}(\Pi_{D(\lambda,r)}(M)) =: n(D(\lambda,r);\, M), \tag{A-1}$$

where

$$\Pi_{D(\lambda,r)}(M) = \frac{1}{2i\pi}\int_{\partial D(\lambda,r)} (z-M)^{-1}\,dz.$$

We denote by $\mathscr{D}_0(E)$ the set of complex matrices on a vector space E which are diagonalizable and invertible.

Given two subsets $A, B \subset \mathbb{C}$, we say that $A \subset B + \mathcal{O}(h)$ if there exists $C > 0$ such that $A \subset B + B(0, Ch)$.

Definition A.1. Let $\mathscr{E} = (E_j)_{j=1,\dots,p}$ be a sequence of finite-dimensional vector spaces E_j of dimension $r_j > 0$, let $E = \oplus_{j=1,\dots,p} E_j$ and let $\tau = (\tau_2, \dots, \tau_p) \in (\mathbb{R}_+^*)^{p-1}$. Suppose that $(h,\tau) \mapsto \mathcal{M}_h(\tau)$ is a map from $(0,1] \times (\mathbb{R}_+^*)^{p-1}$ into the set of complex matrices on E.

We say that $\mathcal{M}_h(\tau)$ is an $(\mathscr{E}, \tau, h)$-graded matrix if there exists $\mathcal{M}' \in \mathscr{D}_0(E)$ independent of (h,τ) such that $\mathcal{M}_h(\tau) = \Omega(\tau)(\mathcal{M}' + \mathcal{O}(h))\Omega(\tau)$ with $\Omega(\tau)$ and $\mathcal{M}'$ such that

- $\mathcal{M}' = \operatorname{diag}(M_j, j = 1, \dots, p)$ with $M_j \in \mathscr{D}_0(E_j)$,
- $\Omega(\tau) = \operatorname{diag}(\varepsilon_j(\tau)I_{r_j}, j = 1, \dots, p)$ with $\varepsilon_1(\tau) = 1$ and $\varepsilon_j(\tau) = \left(\prod_{k=2}^{j}\tau_k\right)$ for all $j \geq 2$.

Throughout, we denote by $\mathscr{G}(\mathscr{E},\tau,h)$ the set of $(\mathscr{E},\tau,h)$-graded matrices.

Lemma A.2. *Suppose that $\mathcal{M}_h(\tau)$ is a family of $(\mathscr{E},\tau,h)$-graded matrices and that $p\geq 2$. Then, one has*

$$\mathcal{M}_h(\tau)=\begin{pmatrix} J(h) & \tau_2 B_h^+(\tau')^* \\ \tau_2 B_h^-(\tau') & \tau_2^2\mathcal{N}_h(\tau')\end{pmatrix}, \tag{A-2}$$

where

- *$J(h)=M_1+\mathcal{O}(h)$ with $M_1\in\mathscr{D}_0(E_1)$,*
- *$\mathcal{N}_h(\tau')\in\mathscr{G}(\mathscr{E}',\tau',h)$ with $\tau'=(\tau_3,\dots,\tau_p)$ and $\mathscr{E}'=(E_j)_{j=2,\dots,p}$,*
- *$B_h^\pm(\tau')\in\mathscr{M}(E_1,\oplus_{j=2}^p E_j)$ satisfies*

  $$B_h^\pm(\tau')^*=(b_2^\pm(h)^*,\tau_3 b_3^\pm(h)^*,\tau_3\tau_4 b_4^\pm(h)^*,\dots,\tau_3\cdots\tau_p b_p^\pm(h)^*)$$

  with $b_j^\pm(h):E_1\to E_j$ independent of τ and $b_j(h)=\mathcal{O}(h)$.

Moreover, the matrix $\mathcal{N}_h(\tau')-B_h^-(\tau')J(h)^{-1}B_h^+(\tau')^*$ belongs to $\mathscr{G}(\mathscr{E}',\tau',h)$.

Proof. Assume that $\mathcal{M}_h(\tau)=\Omega(\tau)(\mathcal{M}'+\mathcal{O}(h))\Omega(\tau)$ with $\Omega(\tau)$ and $\mathcal{M}'$ as in Definition A.1. First observe that

$$\Omega(\tau)=\begin{pmatrix} I_{r_1} & 0\\ 0 & \tau_2\Omega'(\tau')\end{pmatrix}$$

with

$$\Omega'(\tau')=\begin{pmatrix} I_{r_2} & 0 & \dots & \dots & 0\\ 0 & \tau_3 I_{r_3} & 0 & \dots & 0\\ \vdots & 0 & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & 0\\ 0 & \dots & \dots & 0 & \tau_3\tau_4\dots\tau_p I_{r_p}\end{pmatrix}.$$

On the other hand, we can write

$$\mathcal{M}'+\mathcal{O}(h)=\begin{pmatrix} J(h) & B'^+(h)^*\\ B'^-(h) & \mathcal{N}'(h)\end{pmatrix}$$

where $J(h)=M_1+\mathcal{O}(h)$ with $M_1\in\mathscr{D}_0(E_1)$, $B'^\pm(h)=\mathcal{O}(h)$, and $\mathcal{N}'(h)=\mathcal{N}_0'+\mathcal{O}(h)$ with $\mathcal{N}_0'=\operatorname{diag}(M_j,\ j=2,\dots,p)$. Therefore,

$$\Omega(\tau)(\mathcal{M}'+\mathcal{O}(h))\Omega(\tau)=\begin{pmatrix} J(h) & \tau_2 B'^+(h)^*\Omega'(\tau')\\ \tau_2\Omega'(\tau')B'^-(h) & \tau_2^2\Omega'(\tau')\mathcal{N}'(h)\Omega'(\tau')\end{pmatrix}$$

has exactly the form (A-2) with $B_h^\pm(\tau')=\Omega'(\tau')B'^\pm(h)$ and $\mathcal{N}_h(\tau')=\Omega'(\tau')\mathcal{N}'(h)\Omega'(\tau')$. By construction, $\mathcal{N}_h(\tau')$ belongs to $\mathscr{G}(\mathscr{E}',\tau',h)$ and $B_h^\pm(\tau')$ has the required form. □

Lemma A.3. *Let M be a complex diagonalizable matrix. Then there exists $C>0$ such that*

$$\forall\lambda\notin\sigma(M),\quad \|(M-\lambda)^{-1}\|\leq C\operatorname{dist}(\lambda,\sigma(M))^{-1}.$$

Proof. Let P be an invertible matrix such that $PMP^{-1} = D$ is diagonal. Then

$$\|(M-\lambda)^{-1}\| = \|P(D-\lambda)^{-1}P^{-1}\| \leq C\|(D-\lambda)^{-1}\| = C\,\mathrm{dist}(\lambda, \sigma(M))^{-1}. \qquad \square$$

The following theorem gives precise information on the spectrum of graded matrices as introduced above. The proof is based on standard arguments, namely on the Schur complement method and complex analysis. The use of these two tools permits us to work by induction and to decompose the base vector space in order to isolate eigenspaces corresponding to eigenvalues of the same order and to see the remainder of the matrix as a perturbation. Similar arguments were used in [Michel 2019] in a self-adjoint framework. We believe that this result could be useful in other contexts where the computation of clouds of eigenvalues cannot be carried out by standard self-adjoint arguments.

Theorem A.4. *Suppose that $\mathcal{M}_h(\tau)$ is $(\mathscr{E}, \tau, h)$-graded. Then, there exists $\tilde{\tau}_0, h_0 > 0$ such that for all $0 < \tau_j \leq \tilde{\tau}_0$ and all $h \in (0, h_0]$, one has*

$$\sigma(\mathcal{M}_h(\tau)) \subset \bigsqcup_{j=1}^{p} \varepsilon_j(\tau)^2(\sigma(M_j) + \mathcal{O}(h)).$$

Moreover, for any eigenvalue λ of M_j with multiplicity $m_j(\lambda)$, there exists $K > 0$ such that, denoting $D_j := \{z \in \mathbb{C},\ |z - \varepsilon_j(\tau)^2\lambda_j| < K\varepsilon_j(\tau)^2 h\}$, one has

$$n(D_j; \mathcal{M}_h(\tau)) = m_j(\lambda), \tag{A-3}$$

where $n(D_j; \mathcal{M}_h(\tau))$ is defined by (A-1). Moreover, there exists $C > 0$ such that

$$\|(\mathcal{M}_h(\tau) - z)^{-1}\| \leq C\,\mathrm{dist}(z, \sigma(\mathcal{M}_h(\tau)))^{-1}$$

for all $z \in \mathbb{C} \setminus \bigcup_{j=1}^{p} \bigcup_{\lambda \in \sigma(M_j)} B(\varepsilon_j(\tau)^2\lambda, \varepsilon_j(\tau)^2 K h)$.

Proof. We prove the theorem by induction on p. Throughout the proof the notation $\mathcal{O}(\cdot)$ is uniform with respect to the parameters h and τ. For $p = 1$, one has $\mathcal{M}_h(\tau) = M_1 + \mathcal{O}(h)$ with $M_1 \in \mathcal{M}_{r_1}(\mathbb{R})$ independent of h, diagonalizable and invertible. Let us denote λ^1_j, $j = 1, \dots, n_1$, its eigenvalues and $m_j = m(\lambda^1_j)$ the corresponding multiplicities. The function $z \mapsto (\mathcal{M}_h - z)^{-1}$ is meromorphic on $\mathbb{C}$ with poles in $\sigma(\mathcal{M}_h)$. Moreover, Lemma A.3 and the identity

$$\mathcal{M}_h - z = (M_1 - z)(\mathrm{Id} + (M_1 - z)^{-1}\mathcal{O}(h)), \quad \forall z \notin \sigma(M_1)$$

show that for any $C > 0$ large enough, $(\mathcal{M}_h - z)$ is invertible on $\mathbb{C} \setminus \bigcup_{j=1}^{n_1} D(\lambda^1_j, Ch)$ with $\|(M_1 - z)^{-1}\| = \mathcal{O}(1/(Ch))$ and

$$(\mathcal{M}_h - z)^{-1} = (\mathrm{Id} + (M_1 - z)^{-1}\mathcal{O}(h))^{-1}(M_1 - z)^{-1}. \tag{A-4}$$

Hence, for every $C > 0$ large enough, the associated spectral projector writes

$$\Pi_{D(\lambda^1_j, Ch)}(\mathcal{M}_h) = \frac{1}{2i\pi} \int_{\partial D(\lambda^1_j, Ch)} \Big(\mathrm{Id} + \mathcal{O}\Big(\frac{1}{C}\Big)\Big)^{-1} (z - M_1)^{-1}\, dz.$$

This implies that for $C > 0$ large enough,

$$\operatorname{rank}(\Pi_{D(\lambda_j^1, Ch)}(\mathcal{M}_h)) = \operatorname{rank}(\Pi_{D(\lambda_j^1, Ch)}(M_1)) = m_j,$$

which is exactly (A-3). As a consequence

$$\sum_{j=1}^{n_1} \operatorname{rank}(\Pi_{D(\lambda_j^1, Ch)}(\mathcal{M}_h)) = \sum_{j=1}^{n_1} m_j = r_1$$

is maximal and hence $\sigma(\mathcal{M}_h) \subset \bigcup_{j=1}^{n_1} D(\lambda_j^1, Ch)$. Finally, (A-4) shows that one has, for any $z \in \mathbb{C} \setminus \bigcup_{j=1}^{n_1} D(\lambda_j^1, Ch)$,

$$\|(\mathcal{M}_h - z)^{-1}\| \leq C' \|(M_1 - z)^{-1}\|$$

for some constant $C' > 0$. Using Lemma A.3 we get

$$\|(\mathcal{M}_h - z)^{-1}\| \leq C' \operatorname{dist}(z, \sigma(M_1))^{-1} \leq C'' \operatorname{dist}(z, \sigma(\mathcal{M}_h))^{-1}$$

for all $z \in \mathbb{C} \setminus \bigcup_{j=1}^{n_1} D(\lambda_j^1, 2Ch)$. This completes the initialization step.

Suppose now that $p \geq 2$ and let $\mathcal{M}_h(\tau) \in \mathscr{G}(\mathscr{E}, \tau, h)$. We have

$$\mathcal{M}_h(\tau) = \begin{pmatrix} J(h) & \tau_2 B_h^+(\tau')^* \\ \tau_2 B_h^-(\tau') & \tau_2^2 \mathcal{N}_h(\tau') \end{pmatrix}$$

with $J(h)$, $B_h^\pm(\tau')$ and $\mathcal{N}_h(\tau')$ as in Lemma A.2. In order to lighten the notation, we will drop the variables τ, τ' in the proof below. For $\lambda \in \mathbb{C}$, let

$$\mathcal{P}(\lambda) := \mathcal{M}_h(\tau) - \lambda = \begin{pmatrix} J(h) - \lambda & \tau_2 B_h^{+,*} \\ \tau_2 B_h^- & \tau_2^2 \mathcal{N}_h - \lambda \end{pmatrix}. \tag{A-5}$$

This is a holomorphic function, and since it is nontrivial, its inverse is well defined excepted for a finite number of values of λ which are exactly the spectral values of $\mathcal{M}_h$.

We first study the part of the spectrum of $\mathcal{M}_h$ which is of largest modulus. Let λ_n^1, $n = 1, \dots, n_1$, denote the eigenvalues of the matrix M_1. Since $J(h) = M_1 + \mathcal{O}(h)$ and $M_1 \in \mathscr{D}_0(E_1)$, the initialization step shows that there exist $C > 0$ such that $\sigma(J(h)) \subset \bigcup_{n=1}^{n_1} D(\lambda_n^1, Ch)$. Moreover, since M_1 is invertible, there exists $c_1, d_1 > 0$ and $h_0 > 0$ such that for all $n = 1, \dots, n_1$, one has $\lambda_n^1 \in K(c_1, d_1)$ where $K(c_1, d_1) = \{z \in \mathbb{C},\ c_1 \leq |z| \leq d_1\}$. Let $n \in \{1, \dots, n_1\}$ be fixed and consider $D_n = D_n(h) = \{z \in \mathbb{C},\ |z - \lambda_n^1| \leq Mh\}$ for some $M > C > 0$ and $\tilde{D}_n = \{z \in \mathbb{C},\ |z - \lambda_n^1| \leq 2Mh\}$. Observe that for $h > 0$ small enough, the disks $\tilde{D}_n$ are disjoint. By definition, one has $\mathcal{N}_h(\tau') = \mathcal{O}(1)$ and since $|\lambda| \geq c_1 - \mathcal{O}(h) \geq c_1/2$, this implies that for $\tau_2 > 0$ small enough with respect to c_1 and $\lambda \in \tilde{D}_n$, the matrix $\tau_2^2 \mathcal{N}_h(\tau') - \lambda$ is invertible, and $(\tau_2^2 \mathcal{N}_h(\tau') - \lambda)^{-1} = \mathcal{O}(1)$. Moreover, it follows from the initialization step that for $\lambda \in \tilde{D}_n \setminus D_n$, $J(h) - \lambda$ is invertible and

$$\|(J(h) - \lambda)^{-1}\| = \mathcal{O}(\operatorname{dist}(\lambda, \sigma(J(h))^{-1}) = \mathcal{O}(h^{-1}).$$

Combined with the fact that $B_h^{\pm} = \mathcal{O}(h)$, this implies that for $h > 0$ small enough and $\lambda \in \tilde{D}_n \setminus D_n$, $J(h) - \lambda - \tau_2^2 B_h^{+,*}(\tau_2^2 \mathcal{N}_h - \lambda)^{-1} B_h^-$ is invertible with

$$\begin{aligned}\big(J(h) - \lambda - \tau_2^2 B_h^{+,*}(\tau_2^2 \mathcal{N}_h - \lambda)^{-1} B_h^-\big)^{-1}&\\ &= (J(h)-\lambda)^{-1}\big(I - \tau_2^2 B_h^{+,*}(\tau_2^2 \mathcal{N}_h - \lambda)^{-1} B_h^- (J(h)-\lambda)^{-1}\big)^{-1}\\ &= (J(h)-\lambda)^{-1}(I + \mathcal{O}(h)). \qquad \text{(A-6)}\end{aligned}$$

Hence, the standard Schur complement procedure shows that for $\lambda \in \tilde{D}_n \setminus D_n$, $\mathcal{P}(\lambda)$ is invertible with inverse $\mathcal{E}(\lambda)$ given by

$$\mathcal{E}(\lambda) = \begin{pmatrix} E(\lambda) & -\tau_2 E(\lambda) B_h^{+,*}(\tau_2^2 \mathcal{N}_h - \lambda)^{-1} \\ -\tau_2(\tau_2^2 \mathcal{N}_h - \lambda)^{-1} B_h^- E(\lambda) & E_0(\lambda) \end{pmatrix} \qquad \text{(A-7)}$$

with

$$E(\lambda) = \big(J(h) - \lambda - \tau_2^2 B_h^{+,*}(\tau_2^2 \mathcal{N}_h - \lambda)^{-1} B_h^-\big)^{-1}$$

and

$$E_0(\lambda) = (\tau_2^2 \mathcal{N}_h - \lambda)^{-1} + \tau_2^2(\tau_2^2 \mathcal{N}_h - \lambda)^{-1} B_h^- E(\lambda) B_h^{+,*}(\tau_2^2 \mathcal{N}_h - \lambda)^{-1}.$$

Let us now consider the spectral projector $\Pi_{D_n}(\mathcal{M}_h)$. Then,

$$\operatorname{rank}(\Pi_{D_n}(\mathcal{M}_h)) \geq \operatorname{rank}(\tilde{\Pi}_n),$$

where we defined

$$\tilde{\Pi}_n = \begin{pmatrix} \mathrm{Id} & 0 \\ 0 & 0 \end{pmatrix} \Pi_{D_n}(\mathcal{M}_h) \begin{pmatrix} \mathrm{Id} & 0 \\ 0 & 0 \end{pmatrix}.$$

On the other hand, an elementary computation shows that

$$\tilde{\Pi}_n = \frac{1}{2i\pi} \int_{\partial D_n} \begin{pmatrix} E(\lambda) & 0 \\ 0 & 0 \end{pmatrix} d\lambda = \begin{pmatrix} E_n & 0 \\ 0 & 0 \end{pmatrix}$$

with

$$\begin{aligned} E_n &= \frac{1}{2i\pi} \int_{\partial D_n} \big(J(h) - \lambda - \tau_2^2 B_h^{+,*}(\tau_2^2 \mathcal{N}_h - \lambda)^{-1} B_h^-\big)^{-1} d\lambda \\ &= \frac{1}{2i\pi} \int_{\partial D_n} (J(h)-\lambda)^{-1}(I + \mathcal{O}(h))\, d\lambda, \end{aligned}$$

where the last equality follows from (A-6). It follows that for $h > 0$ small enough, the rank of E_n is bounded from below by the multiplicity $m(\lambda_n^1)$ of λ_n^1 and hence

$$\operatorname{rank}(\Pi_{D_n}(\mathcal{M}_h)) \geq m(\lambda_n^1) \qquad \text{(A-8)}$$

for all $n = 1, \dots, n_1$.

Let us now study the part of the spectrum of order smaller than τ_2^2. Thanks to the last part of Lemma A.2, the matrix $\mathcal{Z}_h(\tau') := \mathcal{N}_h - B_h^- J(h)^{-1} B_h^{+,*}$ is classical $(\mathscr{E}', \tau')$-graded. Hence, it follows from the induction hypothesis that uniformly with respect to h, one has

$$\sigma(\mathcal{Z}_h(\tau')) \subset \bigsqcup_{j=2}^{p} \tilde{\varepsilon}_j^2(\sigma(M_j) + \mathcal{O}(h)) \tag{A-9}$$

with $\tilde{\varepsilon}_j = \tau_2^{-1}\varepsilon_j = \prod_{l=3}^{j} \tau_l$ for $j \geq 3$ and $\tilde{\varepsilon}_2 = 1$. One also knows that for all $j = 2, \dots, p$ and all $\lambda \in \sigma(M_j)$,

$$\operatorname{rank} \Pi_{D_j}(\mathcal{Z}_h) = m_j(\lambda)$$

where $D_j = D(\lambda\tilde{\varepsilon}_j^2, Kh\tilde{\varepsilon}_j^2)$ for some $K > 0$. Moreover, for all $z \notin \bigcup_{j=2}^{p} \bigcup_{\lambda\in\sigma(M_j)} D(\lambda\tilde{\varepsilon}_j^2, Kh\tilde{\varepsilon}_j^2)$, one has the resolvent estimate

$$(\mathcal{Z}_h(\tau') - z)^{-1} = \mathcal{O}(\operatorname{dist}(z, \sigma(\mathcal{Z}_h(\tau'))^{-1}). \tag{A-10}$$

For $j = 2, \dots, p$, let $\lambda_1^j, \dots, \lambda_{n_j}^j$ denote the eigenvalues of the matrix $M_j \in \mathcal{D}_0$. As above, there exist $c_j, d_j > 0$ such that $\lambda_n^j \in K(c_j, d_j)$ for all $n = 1, \dots, n_j$. Suppose now that $j \in \{2, \dots, p\}$ and $n \in \{1, \dots, n_j\}$ are fixed and consider, for $M > K$,

$$D'_{j,n} = \{z \in \mathbb{C},\ |z - \varepsilon_j^2\lambda_n^j| \leq Mh\varepsilon_j^2\} = \tau_2^{-2}\{z' \in \mathbb{C},\ |z' - \tilde{\varepsilon}_j^2\lambda_n^j| \leq Mh\tilde{\varepsilon}_j^2\}.$$

Since M_1 is invertible, $J(h) - \lambda$ is invertible and $(J(h) - \lambda)^{-1} = \mathcal{O}(1)$ for λ in $D'_{j,n}$ and h, τ_2 small enough. Moreover, for any $\lambda \in \partial D'_{j,n}$, noting $\lambda' = \tau_2^{-2}\lambda$,

$$\begin{aligned} \tau_2^2\mathcal{N}_h - \lambda - \tau_2^2 B_h^-(J(h) - \lambda)^{-1}B_h^{+,*} &= \tau_2^2(\mathcal{N}_h - \lambda' - B_h^-(J(h) - \lambda)^{-1}B_h^{+,*}) \\ &= \tau_2^2(\mathcal{Z}_h - \lambda' - B_h^-((J(h) - \lambda)^{-1} - J(h)^{-1})B_h^{+,*}) \\ &= \tau_2^2(\mathcal{Z}_h - \lambda')(I + \mathcal{O}(h^2|\lambda|\|(\mathcal{Z}_h - \lambda')^{-1})\|). \end{aligned}$$

Hence, according to the relations (A-9), (A-10), and to $\varepsilon_j = \tau_2\tilde{\varepsilon}_j$,

$$\begin{aligned} \tau_2^2\mathcal{N}_h - \lambda - \tau_2^2 B_h^-(J(h) - \lambda)^{-1}B_h^{+,*} &= \tau_2^2(\mathcal{Z}_h - \lambda')(I + \mathcal{O}(h^2\varepsilon_j^2\|(\mathcal{Z}_h - \lambda')^{-1})\|) \\ &= \tau_2^2(\mathcal{Z}_h - \lambda')\left(I + \mathcal{O}\left(h\frac{\varepsilon_j^2}{\tilde{\varepsilon}_j^2}\right)\right) \\ &= \tau_2^2(\mathcal{Z}_h - \lambda')(I + \mathcal{O}(h\tau_2^2)). \end{aligned} \tag{A-11}$$

The latter operator is then invertible around $\partial D'_{j,n}$ for h, τ_2 small enough, and the Schur complement formula then permits us to write the inverse of $\mathcal{P}(\lambda)$ as

$$\mathcal{E}(\lambda) = \begin{pmatrix} E_0(\lambda) & -\tau_2(J(h) - \lambda)^{-1}B_h^{+,*}E(\lambda) \\ -\tau_2 E(\lambda)B_h^-(J(h) - \lambda)^{-1} & E(\lambda) \end{pmatrix} \tag{A-12}$$

with

$$E(\lambda) = \left(\tau_2^2\mathcal{N}_h - \lambda - \tau_2^2 B_h^-(J(h) - \lambda)^{-1}B_h^{+,*}\right)^{-1}$$

and

$$E_0(\lambda) = (J(h) - \lambda)^{-1} + \tau_2^2(J(h) - \lambda)^{-1}B_h^{+,*}E(\lambda)B_h^-(J(h) - \lambda)^{-1}.$$

As above, let us consider the corresponding projector $\Pi_{D'_{j,n}}(\mathcal{M}_h)$. From $\lambda = \tau_2^2\lambda'$, we get

$$\Pi_{D'_{j,n}}(\mathcal{M}_h) = \frac{\tau_2^2}{2i\pi}\int_{\partial\hat{D}'_{j,n}} \mathcal{E}(\tau_2^2\lambda')\, d\lambda'$$

with $\hat{D}'_{j,n} = \{z' \in \mathbb{C},\ |z' - \tilde{\varepsilon}_j^2\lambda_n^j| \leq Mh\tilde{\varepsilon}_j^2\}$. It follows moreover from (A-11) that for every $\lambda' \in \partial\hat{D}'_{j,n}$ and h, τ_2 small enough,

$$E(\tau_2^2\lambda') = \tau_2^{-2}(\mathcal{Z}_h - \lambda')^{-1}(I + \mathcal{O}(h)), \tag{A-13}$$

and the same argument as above shows that $\operatorname{rank}(\Pi_{D'_{j,n}}(\mathcal{M}_h)) \geq \operatorname{rank}(E'_n)$ with

$$E'_n = \frac{\tau_2^2}{2i\pi}\int_{\partial\hat{D}'_{j,n}} E(\tau_2^2 z)\, dz = \frac{1}{2i\pi}\int_{\partial\hat{D}'_{j,n}} (\mathcal{Z}_h - z)^{-1}(I + \mathcal{O}(h))^{-1}\, dz.$$

By the induction hypothesis, this shows that for h small enough, the rank of E'_n is exactly the multiplicity of λ_n^j and hence

$$\operatorname{rank}(\Pi_{D'_{j,n}}(\mathcal{M}_h)) \geq m(\lambda_n^j)$$

for all $j = 2, \dots, p$ and $n = 1, \dots, n_j$. Combined with (A-8), this shows that for all $j = 1, \dots, p$ and $n = 1, \dots, n_j$,

$$\operatorname{rank}(\Pi_{D_{j,n}}(\mathcal{M}_h)) \geq m(\lambda_n^j)$$

with $D_{j,n} = \varepsilon_j^2 D(\lambda_n^j, Mh)$. Since $\sum_{j,n} m(\lambda_n^j)$ is equal to the total dimension of the space, this implies

$$\operatorname{rank}(\Pi_{D_{j,n}}(\mathcal{M}_h)) = m(\lambda_n^j) \tag{A-14}$$

which proves the localization of the spectrum and (A-3).

It remains to prove the resolvent estimate. Suppose that $\lambda \in \mathbb{C}$ is such that

$$\lambda \notin \bigcup_{j=1}^{p} \bigcup_{\mu\in\sigma(M_j)} D(\varepsilon_j^2(\tau)\mu, \varepsilon_j^2(\tau)Kh).$$

We suppose first that $|\lambda| \geq c_0$ for $c_0 > 0$ such that $|\lambda_n^1| \geq 2c_0$ for all $n = 1, \dots, n_1$. Then $\mathcal{P}(\lambda) = \mathcal{M}_h(\tau) - \lambda$ is invertible with inverse $\mathcal{E}(\lambda)$ given by (A-7). Using (A-6) it is clear that $E(\lambda) = \mathcal{O}(h^{-1}) = \mathcal{O}(\operatorname{dist}(\lambda, \sigma(\mathcal{M}_h(\tau))^{-1})$. On the other hand, since $(\tau_2^2\mathcal{N}_h - \lambda)^{-1} = \mathcal{O}(1)$ and $B_h^{\pm} = \mathcal{O}(h)$ we have also $E_0(\lambda) = \mathcal{O}(1)$ and then $\mathcal{E}(\lambda) = \mathcal{O}(\operatorname{dist}(\lambda, \sigma(\mathcal{M}_h(\tau)))^{-1})$.

Suppose now that $|\lambda| \leq c_0$. Then $\mathcal{P}(\lambda) = \mathcal{M}_h(\tau) - \lambda$ is invertible with inverse $\mathcal{E}(\lambda)$ given by (A-12). Setting $\lambda' = \tau_2^{-2}\lambda$ one deduces from (A-13) and from (A-9) and (A-10) that

$$E(\lambda) = \mathcal{O}(\tau_2^{-2}\operatorname{dist}(\lambda', \sigma(\mathcal{Z}_h))^{-1}) = \mathcal{O}(\operatorname{dist}(\lambda, \sigma(\mathcal{M}_h(\tau))^{-1}).$$

This completes the proof. □

Acknowledgements.

The authors thank the anonymous referees for their remarks which improved the quality of the paper. Both authors are members of the ANR project QuAMProcs 19-CE40-0010-01.

References

[Bouchet and Reygner 2016] F. Bouchet and J. Reygner, "Generalisation of the Eyring–Kramers transition rate formula to irreversible diffusion processes", *Ann. Henri Poincaré* **17**:12 (2016), 3499–3532. MR

[Bovier et al. 2004] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein, "Metastability in reversible diffusion processes, I: Sharp asymptotics for capacities and exit times", *J. Eur. Math. Soc.* (*JEMS*) **6**:4 (2004), 399–424. MR

[Bovier et al. 2005] A. Bovier, V. Gayrard, and M. Klein, "Metastability in reversible diffusion processes, II: Precise asymptotics for small eigenvalues", *J. Eur. Math. Soc.* (*JEMS*) **7**:1 (2005), 69–99. MR

[Di Gesù and Le Peutrec 2017] G. Di Gesù and D. Le Peutrec, "Small noise spectral gap asymptotics for a large system of nonlinear diffusions", *J. Spectr. Theory* **7**:4 (2017), 939–984. MR

[Di Gesù et al. 2020] G. Di Gesù, T. Lelièvre, D. Le Peutrec, and B. Nectoux, "The exit from a metastable state: Concentration of the exit point distribution on the low energy saddle points, part 1", *J. Math. Pures Appl.* (9) **138** (2020), 242–306. MR

[Helffer 1988] B. Helffer, *Semi-classical analysis for the Schrödinger operator and applications*, Lecture Notes in Mathematics **1336**, Springer, 1988. MR

[Helffer 2013] B. Helffer, *Spectral Theory and its Applications*, Cambridge Studies in Advanced Mathematics **139**, Cambridge University Press, 2013.

[Helffer and Nier 2005] B. Helffer and F. Nier, *Hypoelliptic estimates and spectral theory for Fokker–Planck operators and Witten Laplacians*, Lecture Notes in Mathematics **1862**, Springer, 2005. MR

[Helffer and Nier 2006] B. Helffer and F. Nier, *Quantitative analysis of metastability in reversible diffusion processes via a Witten complex approach: the case with boundary*, Mém. Soc. Math. Fr. (N.S.) **105**, Société Mathématique de France, Paris, 2006. MR

[Helffer and Sjöstrand 1985] B. Helffer and J. Sjöstrand, "Puits multiples en mécanique semi-classique, IV: Étude du complexe de Witten", *Comm. Partial Differential Equations* **10**:3 (1985), 245–340. MR

[Helffer et al. 2004] B. Helffer, M. Klein, and F. Nier, "Quantitative analysis of metastability in reversible diffusion processes via a Witten complex approach", *Mat. Contemp.* **26** (2004), 41–85. MR

[Hérau et al. 2011] F. Hérau, M. Hitrik, and J. Sjöstrand, "Tunnel effect and symmetries for Kramers–Fokker–Planck type operators", *J. Inst. Math. Jussieu* **10**:3 (2011), 567–634. MR

[Kato 1995] T. Kato, *Perturbation theory for linear operators*, Springer, Berlin, 1995. MR

[Landim and Seo 2018] C. Landim and I. Seo, "Metastability of nonreversible random walks in a potential field and the Eyring–Kramers transition rate formula", *Comm. Pure Appl. Math.* **71**:2 (2018), 203–266. MR

[Landim et al. 2019] C. Landim, M. Mariani, and I. Seo, "Dirichlet's and Thomson's principles for non-selfadjoint elliptic operators with application to non-reversible metastable diffusion processes", *Arch. Ration. Mech. Anal.* **231**:2 (2019), 887–938. MR

[Le Peutrec 2010] D. Le Peutrec, "Small eigenvalues of the Neumann realization of the semiclassical Witten Laplacian", *Ann. Fac. Sci. Toulouse Math.* (6) **19**:3-4 (2010), 735–809. MR

[Le Peutrec and Nectoux 2019] D. Le Peutrec and B. Nectoux, "Small eigenvalues of the Witten Laplacian with Dirichlet boundary conditions: the case with critical points on the boundary", preprint, 2019. To appear in *Ann. PDE*. arXiv

[Lelièvre et al. 2013] T. Lelièvre, F. Nier, and G. A. Pavliotis, "Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion", *J. Stat. Phys.* **152**:2 (2013), 237–274. MR

[Menz and Schlichting 2014] G. Menz and A. Schlichting, "Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape", *Ann. Probab.* **42**:5 (2014), 1809–1884. MR

[Michel 2016] L. Michel, "Around supersymmetry for semiclassical second order differential operators", *Proc. Amer. Math. Soc.* **144**:10 (2016), 4487–4500. MR

[Michel 2019] L. Michel, "About small eigenvalues of the Witten Laplacian", *Pure Appl. Anal.* **1**:2 (2019), 149–206. MR

[Michel and Zworski 2018] L. Michel and M. Zworski, "A semiclassical approach to the Kramers–Smoluchowski equation", *SIAM J. Math. Anal.* **50**:5 (2018), 5362–5379. MR

[Sjöstrand and Zworski 2007] J. Sjöstrand and M. Zworski, "Elementary linear algebra for advanced spectral problems", *Ann. Inst. Fourier (Grenoble)* **57**:7 (2007), 2095–2141. MR

[Witten 1982] E. Witten, "Supersymmetry and Morse theory", *J. Differential Geometry* **17**:4 (1982), 661–692. MR

Received 26 Sep 2019. Revised 24 Aug 2020. Accepted 15 Sep 2020.

DORIAN LE PEUTREC: dorian.le-peutrec@univ-orleans.fr
Institut Denis Poisson, Université d'Orléans, Orléans, France

LAURENT MICHEL: laurent.michel@math.u-bordeaux.fr
Institut Mathématiques de Bordeaux, Université de Bordeaux, Talence, France

PROBABILITY and
MATHEMATICAL
PHYSICS

Vol. 1, No. 1, 2020
https://doi.org/10.2140/pmp.2020.1.55

JOINT DISTRIBUTION OF BUSEMANN FUNCTIONS IN THE EXACTLY SOLVABLE CORNER GROWTH MODEL

WAI-TONG LOUIS FAN AND TIMO SEPPÄLÄINEN

The 1+1-dimensional corner growth model with exponential weights is a centrally important exactly solvable model in the Kardar–Parisi–Zhang class of statistical mechanical models. While significant progress has been made on the fluctuations of the growing random shape, understanding of the optimal paths, or geodesics, is less developed. The Busemann function is a useful analytical tool for studying geodesics. We describe the joint distribution of the Busemann functions, simultaneously in all directions of growth. As applications of this description we derive a marked point process representation for the Busemann function across a single lattice edge and calculate some marginal distributions of Busemann functions and semi-infinite geodesics.

1. Introduction

The corner growth model in the Kardar–Parisi–Zhang class. The planar corner growth model (CGM) is a directed last-passage percolation (LPP) model on the planar integer lattice $\mathbb{Z}^2$ whose paths are allowed to take nearest-neighbor steps $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$. In the exactly solvable case the random weights attached to the vertices of $\mathbb{Z}^2$ are i.i.d. exponentially or geometrically distributed random variables.

The exact solvability of the exponential and geometric CGM has been fundamental to the 20-year progress in the study of the 1+1-dimensional Kardar–Parisi–Zhang (KPZ) universality class. After the initial breakthrough by Baik, Deift and Johansson [Baik et al. 1999] on planar LPP on Poisson points,

Part of this research was done while Fan was a Van Vleck visiting assistant professor at UW-Madison. Fan was partially supported by National Science Foundation grant DMS-1804492. Seppäläinen was partially supported by National Science Foundation grants DMS-1602486 and DMS-1854619 and by the Wisconsin Alumni Research Foundation.

MSC2010: 60K35, 60K37.

Keywords: Busemann function, Catalan number, Catalan triangle, cocycle, competition interface, corner growth model, directed percolation, geodesic, last-passage percolation, M/M/1 queue, multiclass fixed point, rhobar distance, TASEP.

the Tracy–Widom limit of the geometric and exponential CGM followed in [Johansson 2000]. This work relied on techniques that today would be called *integrable probability*, a subject that applies ideas from representation theory and integrable systems to study stochastic models. A large literature has followed. Recent reviews appear in [Corwin 2018; 2016; 2012]. A different line of work was initiated in [Balázs et al. 2006] that gave a probabilistic proof of the KPZ exponents of the exponential CGM, following the seminal work [Cator and Groeneboom 2006] on the planar Poisson LPP. The proof utilized the tractable stationary version of the CGM and developed estimates by coupling perturbed versions of the CGM process. This opening led to the first proofs of KPZ exponents for the asymmetric simple exclusion process (ASEP) [Balázs and Seppäläinen 2010] and the KPZ equation [Balázs et al. 2011], to the discovery of the first exactly solvable positive-temperature lattice polymer model [Seppäläinen 2012], to a proof of KPZ exponents for a class of zero-range processes outside known exactly solvable models [Balázs et al. 2012], and most recently to Doob transforms and martingales in random walks in random environments (RWRE) that manifest KPZ behavior [Balázs et al. 2019b]. The estimates from [Balázs et al. 2006] have also been applied to coalescence times of geodesics [Pimentel 2016] and to the local behavior of Airy processes [Pimentel 2018].

Joint distribution of Busemann functions. The present article places the stationary CGM into a larger context by describing the natural coupling of all the stationary CGMs. This coupling arises from the joint distribution of the Busemann functions in all directions of growth.

Let $G_{x,y}$ denote the last-passage value between points x and y on the lattice $\mathbb{Z}^2$ (the precise definition follows in (2-1) in Section 2). The Busemann function $B^{\rho}_{x,y}$ is the limit of increments $G_{v_n,y} - G_{v_n,x}$ as v_n is taken to infinity in the direction parametrized by ρ. In a given direction this limit exists almost surely. These limits are extended to a process $B^{\bullet}$ by taking limits in the parameter ρ. Finite-dimensional distributions of $B^{\bullet}$ are identified as the unique invariant distributions of multiclass LPP processes. These distributions are conveniently described in terms of mappings that represent FIFO (first-in-first-out) queues. Key points of the development are (i) an intertwining between two types of multiclass processes, called the *multiline process* and the *coupled process*, and (ii) a triangular array representation of the intertwining mapping.

The results of this paper will have various applications in the study of the CGM, and they can be extended to other 1+1-dimensional growth and polymer models that have a tractable stationary version. A forthcoming work of the authors develops the joint distribution of Busemann functions for the positive-temperature log-gamma polymer model.

Two applications have been completed recently. Properties of the joint Busemann process $B^{\bullet}$ discovered here are applied in [Janjigian et al. 2019] to describe (i) the overall structure of the geodesics of the exponential corner growth model and (ii) the statistics of a new object termed the “instability graph” that captures the geometry of the jumps of the Busemann functions on the lattice. The joint Busemann distribution is necessary for a full picture of the geodesics because in a fixed direction semi-infinite geodesics are almost surely unique and coalesce (these facts are reviewed in Section 2B below) but there are random directions of nonuniqueness. The joint distribution captures the jumps of the Busemann function as the direction varies. These correspond to jumps in coalescence points and nonuniqueness of geodesics.

The article [Balázs et al. 2019a] gives a proof of the nonexistence of bi-infinite geodesics in the exponential CGM, based on couplings with the stationary version of the LPP process. The joint distribution described here is a critical ingredient of the proof.

An analogue of the Busemann function on the Airy sheet. An interesting similarity appears between our paper and recent work on the universal objects that arise from LPP. Basu, Ganguly and Hammond [Basu et al. 2019] study an analogue of the Busemann function in the Brownian last-passage model. Instead of the lattice scale and all spatial directions, they look at a difference of last-passage values on the scale $n^{2/3}$ into a fixed macroscopic direction, where universal objects such as Airy processes arise. Translated to the CGM, their object of interest is the weak limit $Z(z)$ of the scaled difference $n^{-1/3}[G_{(n^{2/3},0),(n+zn^{2/3},n)} - G_{(-n^{2/3},0),(n+zn^{2/3},n)} + 4n^{2/3}]$ that they call the *difference weight profile*. In terms of the Airy sheet $\{W(x, y) : x, y \in \mathbb{R}\}$ constructed recently by Dauvergne, Ortmann and Virág [Dauvergne et al. 2018], the limit $Z(z) = W(1, z) - W(-1, z)$.

The limit $Z(\cdot)$ is a continuous process, while the Busemann process we construct is a jump process. But like the Busemann process, the limit $Z(\cdot)$ is constant in a neighborhood of each point, except for a small set of exceptional points. In both settings this constancy reflects the same underlying phenomenon, namely the coalescence of geodesics. In our lattice setting, Theorem 3.4 gives a precise description of these exceptional directions in terms of an inhomogeneous Poisson process.

Past work. We mention related past work on queues, particle systems, and the CGM.

Queueing fixed points. We formulate a queueing operator as a mapping of bi-infinite sequences of interarrival times and service times into a bi-infinite sequence of interdeparture times (details in Section 2C). When the service times are i.i.d. exponential (memoryless, or $\cdot/M/1$ queue), it is classical that i.i.d. exponential times are preserved by the mapping from the interarrival process to the interdeparture process, subject to the stability condition that the mean interarrival time exceed the mean service time. Anantharam [1993] proved the uniqueness of this fixed point and Chang [1994] gave a shorter argument. (An unpublished manuscript of Liggett and Shiga is also cited in [Mountford and Prabhakar 1995].) Convergence to the fixed point was proved in [Mountford and Prabhakar 1995]. These results were partially extended to general $\cdot/G/1$ queues in [Mairesse and Prabhakar 2003; Prabhakar 2003].

We look at LPP processes with multiple classes of input, but this is *not* the same as a multiclass queue that serves customers in different priority classes. In queueing terms, the present paper describes the unique invariant distribution in a situation where a single memoryless queueing operator transforms a vector of interarrival processes into a vector of interdeparture processes. It is fairly evident a priori that this operation cannot preserve an independent collection of interarrival processes because they are correlated after passing through the same queueing operator. (For example, this operation preserves monotonicity.) It turns out that the queueing mappings themselves provide a way to describe the structure of the invariant distribution.

Multiclass measures for particle systems. In a series of remarkable papers, P. A. Ferrari and J. B. Martin [2006; 2007; 2009] developed queueing descriptions of the stationary distributions of the multiclass totally asymmetric simple exclusion process (TASEP) and the Aldous–Diaconis–Hammersley process.

The intertwining that establishes our Theorem 5.5 became possible after the discovery of a way to apply the ideas of Ferrari and Martin to the CGM. We use the terms multiline process and coupled process to highlight the analogy with their work.

Busemann functions and semi-infinite geodesics. Existence and properties of Busemann functions and semi-infinite geodesics are reviewed in Sections 2A and 2B. Two strategies exist for proving the existence of Busemann functions for the exponential CGM.

(i) Proofs by Ferrari and Pimentel [2005] and Coupier [2011] relied on C. Newman's approach to geodesics [Howard and Newman 2001; Licea and Newman 1996; Newman 1995]. This strategy is feasible because the exact solvability shows that the shape function (2-4) satisfies the required curvature hypotheses.

(ii) A direct argument from the stationary growth model to the Busemann limit was introduced in [Georgiou et al. 2015] for the log-gamma polymer, and applied to the exponential CGM in the lecture notes [Seppäläinen 2018]. An application of this strategy to the CGM with general i.i.d. weights appears in [Georgiou et al. 2017a; 2017b], where the role of the regularity of the shape function becomes explicit.

A sampling of other significant work on Busemann functions and geodesics can be found in [Cator and Pimentel 2012; 2013; Ferrari et al. 2009; Bakhtin et al. 2014; Hoffman 2005; 2008].

Organization of the paper. Section 2 collects preliminaries on the CGM and queues. The main results for Busemann functions and semi-infinite geodesics are stated in Section 3. Section 4 proves a key lemma for the queueing operator. Section 5 introduces the multiline process, the coupled process, and the multiclass LPP process, and then states and proves results on their invariant distributions. The key intertwining between the multiline process and the coupled process appears in (5-8) in the proof of Theorem 5.5 in Section 5D. Section 6 proves the results of Section 3. For the proof of Theorem 3.4, Section 6B introduces a triangular array representation for the intertwining mapping. Auxiliary matters on queues and exponential distributions are relegated to Appendices A and B.

Notation and conventions. Points $x = (x_1, x_2)$, $y = (y_1, y_2) \in \mathbb{R}^2$ are ordered coordinatewise: $x \le y$ if and only if $x_1 \le y_1$ and $x_2 \le y_2$. The ℓ^1 norm is $|x|_1 = |x_1| + |x_2|$. Subscripts indicate restricted subsets of the reals and integers: for example $\mathbb{Z}_{>0} = \{1, 2, 3, \dots\}$. Boldface notation for vectors: $\boldsymbol{e}_1 = (1, 0)$, $\boldsymbol{e}_2 = (0, 1)$, and members of the simplex $[\boldsymbol{e}_2, \boldsymbol{e}_1] = \{t\boldsymbol{e}_1 + (1-t)\boldsymbol{e}_2 : 0 \le t \le 1\}$ are denoted by $\boldsymbol{u}$.

For $n \in \mathbb{Z}_{>0}$, $[n] = \{1, 2, \dots, n\}$, with the convention that $[n] = \varnothing$ for $n \in \mathbb{Z}_{\le 0}$. A finite integer interval is denoted by $[\![m, n]\!] = \{m, m+1, \dots, n\}$, and $[\![m, \infty[\![= \{m, m+1, m+2, \dots\}$.

For $0 < \alpha < \infty$, $X \sim \mathrm{Exp}(\alpha)$ means that random variable X has exponential distribution with rate α; in other words $P(X > t) = e^{-\alpha t}$ for $t > 0$ and $E(X) = \alpha^{-1}$. In the discussion we parametrize exponential variables with their mean. For $0 < \rho < \infty$, ν^ρ is the probability distribution on the space $\mathbb{R}^{\mathbb{Z}}_{\ge 0}$ of bi-infinite sequences under which the coordinates are i.i.d. exponential variables with common mean ρ. Higher-dimensional product measures are denoted by $\nu^{(\rho_1, \rho_2, \dots, \rho_n)} = \nu^{\rho_1} \otimes \nu^{\rho_2} \otimes \cdots \otimes \nu^{\rho_n}$.

For $0 \le p \le 1$, $X \sim \mathrm{Ber}(p)$ means that random variable X has Bernoulli distribution with parameter p; in other words $P(X = 1) = p = 1 - P(X = 0)$.

In general, E^μ represents expectation under a measure μ.

2. Preliminaries

Section 2A introduces the main objects of discussion: the planar corner growth model (CGM), which is a special case of last-passage percolation (LPP), and Busemann functions. Section 2B explains the significance of Busemann functions in the description of directed semi-infinite geodesics and the asymptotic direction of the competition interface. The somewhat technical Section 2C defines FIFO (first-in-first-out) queueing mappings that are used in Section 3 to describe the joint distribution of the Busemann functions. To be sure, the distribution of the Busemann functions could be described by plain mathematical formulas without their queueing content. But the queueing context gives the mathematics meaning that can help comprehend the results.

2A. ***Busemann functions in the corner growth model.*** The setting for the exponential CGM is the following: $(\Omega, \mathfrak{S}, \mathbb{P})$ is a probability space with generic sample point ω. A group of measure-preserving measurable bijections $\{\theta_x\}_{x\in\mathbb{Z}^2}$ acts on $(\Omega, \mathfrak{S}, \mathbb{P})$. Measure preservation means that $\mathbb{P}(\theta_x A) = \mathbb{P}(A)$ for all sets $A \in \mathfrak{S}$ and $x \in \mathbb{Z}^2$. $Y = (Y_x)_{x\in\mathbb{Z}^2}$ is a random field of independent and identically distributed $\operatorname{Exp}(1)$ random weights defined on Ω that satisfies $Y_x(\theta_y\omega) = Y_{x+y}(\omega)$ for $x, y \in \mathbb{Z}^2$ and $\omega \in \Omega$.

The canonical choice for the sample space is the product space $\Omega = \mathbb{R}_{\geq 0}^{\mathbb{Z}^2}$ with its Borel σ-algebra $\mathfrak{S}$, generic sample point $\omega = (\omega_x)_{x\in\mathbb{Z}^2}$, translations $(\theta_x\omega)_y = \omega_{x+y}$, and coordinate random variables $Y_x(\omega) = \omega_x$. Then $\mathbb{P}$ is the i.i.d. product measure on Ω under which each Y_x is an $\operatorname{Exp}(1)$ random variable.

For $u \leq v$ on $\mathbb{Z}^2$ (coordinatewise ordering) let $\Pi_{u,v}$ denote the set of up-right paths $x_\bullet = (x_i)_{i=0}^{|v-u|_1}$ from $x_0 = u$ to $x_{|v-u|_1} = v$ with steps $x_i - x_{i-1} \in \{\boldsymbol{e}_1, \boldsymbol{e}_2\}$. (The left diagram of Figure 1 illustrates this.) Define the last-passage percolation (LPP) process

$$G_{u,v} = \max_{x_\bullet\in\Pi_{u,v}} \sum_{i=0}^{|v-u|_1} Y_{x_i} \quad \text{for } u \leq v \text{ on } \mathbb{Z}^2. \tag{2-1}$$

For $v \in u + \mathbb{Z}_{>0}^2$ we have the inductive equation

$$G_{u,v} = G_{u,v-\boldsymbol{e}_1} \vee G_{u,v-\boldsymbol{e}_2} + Y_v. \tag{2-2}$$

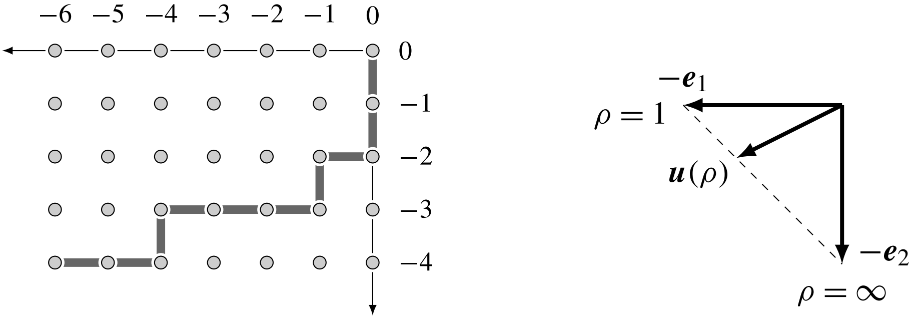


Figure 1. Left: The thickset line segments define an element of $\Pi_{(-6,-4),(0,0)}$. Right: As the parameter ρ increases from 1 to ∞, vector $\boldsymbol{u}(\rho)$ of (2-5) sweeps the directions from $-\boldsymbol{e}_1$ to $-\boldsymbol{e}_2$ in the third quadrant.

The convention of this paper is that growth proceeds in the *south-west* direction (into the *third quadrant* of the plane). Thus the well-known shape theorem (Theorem 5.1 in [Martin 2004], Theorem 3.5 in [Seppäläinen 2018]) of the CGM takes the following form. With probability one,

$$\lim_{r\to\infty} \sup_{x\in(\mathbb{Z}_{\le 0})^2:\,|x|_1\ge r} \frac{|G_{x,0}-g(x)|}{|x|_1}=0 \tag{2-3}$$

with the concave, continuous and one-homogeneous shape function (known since [Rost 1981])

$$g(x)=\bigl(\sqrt{|x_1|}+\sqrt{|x_2|}\,\bigr)^2 \quad \text{for } \ x=(x_1,x_2)\in\mathbb{R}^2_{\le 0}. \tag{2-4}$$

Busemann functions are limits of differences $G_{v,x}-G_{v,y}$ of last-passage values from two fixed points x and y to a common point v that is taken to infinity in a particular direction. These limits are described by relating the direction $\boldsymbol{u}$ that v takes to a real parameter ρ that specifies the distribution of the limits: a bijective mapping between directions $\boldsymbol{u}=(u_1,u_2)\in\,]-\boldsymbol{e}_1,-\boldsymbol{e}_2[$ in the open third quadrant of the plane and parameters $\rho\in(1,\infty)$ is defined by the equations

$$\boldsymbol{u}=\boldsymbol{u}(\rho)=-\left(\frac{1}{1+(\rho-1)^2},\,\frac{(\rho-1)^2}{1+(\rho-1)^2}\right) \iff \rho=\rho(\boldsymbol{u})=\frac{\sqrt{-u_1}+\sqrt{-u_2}}{\sqrt{-u_1}}. \tag{2-5}$$

(See the right diagram of Figure 1 for an illustration.)

The existence and properties of Busemann functions are summarized in the following theorem. By definition, a *down-right lattice path* $\{y_k\}$ satisfies $y_k-y_{k-1}\in\{\boldsymbol{e}_1,-\boldsymbol{e}_2\}$ for all k.

Theorem 2.1. *On the probability space* $(\Omega,\mathfrak{S},\mathbb{P})$ *there exists a cadlag process* $B^\rho=(B^\rho_{x,y})_{x,y\in\mathbb{Z}^2}$ *with state space* $\mathbb{R}^{\mathbb{Z}^2\times\mathbb{Z}^2}$, *indexed by* $\rho\in(1,\infty)$, *with the following properties.*

(i) Path properties. *There is a single event* Ω_0 *such that* $\mathbb{P}(\Omega_0)=1$ *and the following properties hold for all* $\omega\in\Omega_0$, *for all* $\lambda,\rho\in(1,\infty)$ *and* $x,y,z\in\mathbb{Z}^2$:

$$\text{If } \lambda<\rho \text{ then } B^\lambda_{x,x+\boldsymbol{e}_1}\le B^\rho_{x,x+\boldsymbol{e}_1} \text{ and } B^\lambda_{x,x+\boldsymbol{e}_2}\ge B^\rho_{x,x+\boldsymbol{e}_2}. \tag{2-6}$$

$$B^\rho_{x,y}+B^\rho_{y,z}=B^\rho_{x,z}. \tag{2-7}$$

$$Y_x=B^\rho_{x-\boldsymbol{e}_1,x}\wedge B^\rho_{x-\boldsymbol{e}_2,x}. \tag{2-8}$$

Cadlag property: *the path* $\rho\mapsto B^\rho_{x,y}$ *is right continuous and has left limits.*

(ii) Distributional properties. *Each process* B^ρ *is stationary under lattice shifts. The marginal distributions of nearest-neighbor increments are*

$$B^\rho_{x-\boldsymbol{e}_1,x}\sim\mathrm{Exp}(\rho^{-1}) \quad \text{and} \quad B^\rho_{x-\boldsymbol{e}_2,x}\sim\mathrm{Exp}(1-\rho^{-1}). \tag{2-9}$$

Along any down-right path $\{y_k\}_{k\in\mathbb{Z}}$ *on* $\mathbb{Z}^2$, *for fixed* $\rho\in(1,\infty)$ *the increments* $\{B^\rho_{y_k,y_{k+1}}\}_{k\in\mathbb{Z}}$ *are independent.*

(iii) Limits. *Fix* $\rho\in(1,\infty)$ *and let* $\boldsymbol{u}=\boldsymbol{u}(\rho)$ *be the vector determined by* (2-5). *Then there exists an event* $\Omega_0^{(\rho)}$ *such that* $\mathbb{P}(\Omega_0^{(\rho)})=1$ *and the following holds*: *for any sequence* $\{u_n\}$ *in* $\mathbb{Z}^2$ *such that* $|u_n|_1\to\infty$

and $u_n/n \to \boldsymbol{u}$ and for any $\omega \in \Omega_0^{(\rho)}$,

$$B^\rho_{x,y} = \lim_{n\to\infty} [G_{u_n,y} - G_{u_n,x}]. \tag{2-10}$$

Continuity from the left at a fixed $\rho \in (1,\infty)$ holds with probability one: $\lim_{\lambda\nearrow\rho} B^\lambda_{x,y} = B^\rho_{x,y}$ almost surely.

The theorem above is proved as Theorem 4.2 in lecture notes [Seppäläinen 2018]. The central point of the theorem is the limit (2-10), on account of which we call B^ρ the *Busemann function* in direction $\boldsymbol{u}$. We record some observations.

Additivity (2-7) implies that $B^\rho_{x,x} = 0$ and $B^\rho_{x,y} = -B^\rho_{y,x}$. The *weights recovery* property (2-8) can be seen from (2-2) and limits (2-10):

$$\begin{aligned} B^\rho_{x-\boldsymbol{e}_1,x} \wedge B^\rho_{x-\boldsymbol{e}_2,x} &= \lim_{n\to\infty} [G_{u_n,x} - G_{u_n,x-\boldsymbol{e}_1}] \wedge [G_{u_n,x} - G_{u_n,x-\boldsymbol{e}_2}] \\ &= \lim_{n\to\infty} [G_{u_n,x} - G_{u_n,x-\boldsymbol{e}_1} \vee G_{u_n,x-\boldsymbol{e}_2}] = Y_x. \end{aligned}$$

Lemma 2.2. *With probability one, for all $x \in \mathbb{Z}^2$ there exists a random parameter $\rho^*(x) \in (1,\infty)$ such that*

$$\begin{aligned} B^\rho_{x-\boldsymbol{e}_1,x} = Y_x < B^\rho_{x-\boldsymbol{e}_2,x} &\quad \textit{for } \rho \in (1, \rho^*(x)), \\ B^\rho_{x-\boldsymbol{e}_2,x} = Y_x < B^\rho_{x-\boldsymbol{e}_1,x} &\quad \textit{for } \rho \in (\rho^*(x), \infty). \end{aligned} \tag{2-11}$$

The distribution function of $\rho^*(x)$ is $\mathbb{P}\{\rho^*(x) \le \lambda\} = 1 - \lambda^{-1}$ for $1 \le \lambda < \infty$.

Proof. Monotonicity (2-6) and the exponential rates (2-9) force $B^\rho_{x-\boldsymbol{e}_2,x} \nearrow \infty$ almost surely as $\rho \searrow 1$ and $B^\rho_{x-\boldsymbol{e}_1,x} \nearrow \infty$ almost surely as $\rho \nearrow \infty$. Edges $\{x-\boldsymbol{e}_1, x\}$ and $\{x-\boldsymbol{e}_2, x\}$ are part of a down-right path, and hence $B^\rho_{x-\boldsymbol{e}_1,x}$ and $B^\rho_{x-\boldsymbol{e}_2,x}$ are independent exponential random variables for each fixed ρ. Consequently, with probability one, they are distinct for each rational $\rho > 1$. By monotonicity again there is a unique real $\rho^*(x) \in (1,\infty)$ such that for rational $\lambda \in (1,\infty)$,

$$\begin{aligned} \lambda < \rho^*(x) \text{ implies } B^\lambda_{x-\boldsymbol{e}_1,x} < B^\lambda_{x-\boldsymbol{e}_2,x}, \\ \lambda > \rho^*(x) \text{ implies } B^\lambda_{x-\boldsymbol{e}_1,x} > B^\lambda_{x-\boldsymbol{e}_2,x}. \end{aligned} \tag{2-12}$$

By monotonicity the same holds for real λ. Conditions (2-11) follow from weights recovery (2-8). The distribution function comes from (2-11), independence of $B^\rho_{x-\boldsymbol{e}_1,x}$ and $B^\rho_{x-\boldsymbol{e}_2,x}$, and (2-9). □

In particular, for a fixed x, the processes $\{B^\rho_{x-\boldsymbol{e}_1,x}\}_{1<\rho<\infty}$ and $\{B^\rho_{x-\boldsymbol{e}_2,x}\}_{1<\rho<\infty}$ are not independent of each other, even though for a fixed ρ, the random variables $B^\rho_{x-\boldsymbol{e}_1,x}$ and $B^\rho_{x-\boldsymbol{e}_2,x}$ are independent. Vector $\boldsymbol{u}(\rho^*(x))$ is the asymptotic direction of the competition interface emanating from x (see Remark 2.5).

The process $B = \{B^\rho_{x,y}\}$ is a Borel function of the weight configuration Y. Limits (2-10) define B^ρ as a function of Y for a countable dense set of ρ in $(1,\infty)$. The remaining ρ-values $B^\rho_{x,y}$ can then be defined as right limits. Shifts θ_u act on the weights by $(\theta_u Y)_x = Y_{x+u}$. The limits (2-10) give stationarity and ergodicity of B as stated in this lemma.

Lemma 2.3. *Fix $\rho_1, \dots, \rho_n \in (1,\infty)$ and $y_1, \dots, y_n \in \mathbb{Z}^2$. Let $A_x = (B^{\rho_1}_{x,x+y_1}, \dots, B^{\rho_n}_{x,x+y_n})$ and let $0 \ne u \in \mathbb{Z}^2$. Then the $\mathbb{R}^n$-valued process $A = \{A_x\}_{x\in\mathbb{Z}^2}$ is stationary and ergodic under the shift θ_u.*

Proof. Since the i.i.d. process Y is stationary and ergodic under every shift, it suffices to show that $A_x = A_0 \circ \theta_x$ as functions of Y. Let $\boldsymbol{u}^i \in]-\boldsymbol{e}_1, -\boldsymbol{e}_2[$ be associated to ρ_i via (2-5) and fix sequences $\{u^1_m\}, \dots, \{u^n_m\}$ in $\mathbb{Z}^2$ such that, as $m \to \infty$, $|u^i_m|_1 \to \infty$ and $u^i_m/m \to \boldsymbol{u}^i$ for each $i \in [n]$. Then almost surely,

$$\begin{aligned} A_x &= (B^{\rho_1}_{x,x+y_1}, \dots, B^{\rho_n}_{x,x+y_n}) = \lim_{m\to\infty}\big([G_{u^1_m,x+y_1} - G_{u^1_m,x}], \dots, [G_{u^n_m,x+y_n} - G_{u^n_m,x}]\big) \\ &= \lim_{m\to\infty}\big([G_{u^1_m-x,y_1} - G_{u^1_m-x,0}], \dots, [G_{u^n_m-x,y_n} - G_{u^n_m-x,0}]\big) \circ \theta_x \\ &= (B^{\rho_1}_{0,y_1}, \dots, B^{\rho_n}_{0,y_n}) \circ \theta_x = A_0 \circ \theta_x. \end{aligned}$$ □

2B. *Semi-infinite geodesics in the corner growth model.* Let $x_\bullet = \{x_k\}$ be a finite or infinite south-west directed nearest-neighbor path on $\mathbb{Z}^2$ ($x_{k+1} \in \{x_k - \boldsymbol{e}_1, x_k - \boldsymbol{e}_2\}$). Then $x_\bullet$ is a *geodesic* if it gives a maximizing path between any two of its points: for any $k < \ell$ in the index set of $x_\bullet$,

$$G_{x_\ell,x_k} = \sum_{i=k}^{\ell} Y_{x_i}.$$

Given $\rho \in (1,\infty)$, define from each $x \in \mathbb{Z}^2$ the semi-infinite, south-west directed path $\boldsymbol{b}^{\rho,x} = \{\boldsymbol{b}^{\rho,x}_k\}_{k\in\mathbb{Z}_{\ge 0}}$ that starts at $x = \boldsymbol{b}^{\rho,x}_0$ and chooses a step from $\{-\boldsymbol{e}_1, -\boldsymbol{e}_2\}$ by following the minimal increment of B^ρ: for $k \ge 0$,

$$\boldsymbol{b}^{\rho,x}_{k+1} = \begin{cases} \boldsymbol{b}^{\rho,x}_k - \boldsymbol{e}_1, & \text{if } B^\rho_{\boldsymbol{b}^{\rho,x}_k - \boldsymbol{e}_1,\, \boldsymbol{b}^{\rho,x}_k} < B^\rho_{\boldsymbol{b}^{\rho,x}_k - \boldsymbol{e}_2,\, \boldsymbol{b}^{\rho,x}_k}, \\ \boldsymbol{b}^{\rho,x}_k - \boldsymbol{e}_2, & \text{if } B^\rho_{\boldsymbol{b}^{\rho,x}_k - \boldsymbol{e}_2,\, \boldsymbol{b}^{\rho,x}_k} \le B^\rho_{\boldsymbol{b}^{\rho,x}_k - \boldsymbol{e}_1,\, \boldsymbol{b}^{\rho,x}_k}. \end{cases} \tag{2-13}$$

The tie-breaking rule in favor of $-\boldsymbol{e}_2$ is chosen simply to make $\boldsymbol{b}^{\rho,x}$ a cadlag function of ρ. For a given ρ, equality on the right-hand side happens with probability zero. Pictorially, to each point z attach the arrow that points from z to $\boldsymbol{b}^{\rho,z}_1$. For each x the path $\boldsymbol{b}^{\rho,x}$ is constructed by starting at x and following the arrows.

The additivity (2-7) and weights recovery (2-8) imply that $\boldsymbol{b}^{\rho,x}$ is a (semi-infinite) geodesic: let $\ell > k \ge 0$ and suppose $\{y_i\}_{i=k}^{\ell}$ is a south-west directed path from $y_k = \boldsymbol{b}^{\rho,x}_k$ to $y_\ell = \boldsymbol{b}^{\rho,x}_\ell$. Then

$$\sum_{i=k}^{\ell} Y_{y_i} \le \sum_{i=k}^{\ell-1} B^\rho_{y_{i+1},y_i} + Y_{y_\ell} = B^\rho_{y_\ell,y_k} + Y_{y_\ell} = B^\rho_{\boldsymbol{b}^{\rho,x}_\ell,\boldsymbol{b}^{\rho,x}_k} + Y_{\boldsymbol{b}^{\rho,x}_\ell} = \sum_{i=k}^{\ell-1} B^\rho_{\boldsymbol{b}^{\rho,x}_{i+1},\boldsymbol{b}^{\rho,x}_i} + Y_{\boldsymbol{b}^{\rho,x}_\ell} = \sum_{i=k}^{\ell} Y_{\boldsymbol{b}^{\rho,x}_i}.$$

Thus,

$$G_{\boldsymbol{b}^{\rho,x}_\ell,\boldsymbol{b}^{\rho,x}_k} = \sum_{i=k}^{\ell} Y_{\boldsymbol{b}^{\rho,x}_i} = B^\rho_{\boldsymbol{b}^{\rho,x}_\ell,\boldsymbol{b}^{\rho,x}_k} + Y_{\boldsymbol{b}^{\rho,x}_\ell}. \tag{2-14}$$

We call $\boldsymbol{b}^{\rho,x}$ a *Busemann geodesic*.

We state the key properties of semi-infinite geodesics in the next theorem.

Theorem 2.4. *Fix $\rho \in (1,\infty)$ and let $\boldsymbol{u} = \boldsymbol{u}(\rho)$ be the direction associated to ρ by (2-5). The following properties hold with probability one.*

(i) Directedness. *For all $x \in \mathbb{Z}^2$, $\lim_{k\to\infty} \boldsymbol{b}^{\rho,x}_k / k = \boldsymbol{u}$.*

(ii) Uniqueness. *Let $x_\bullet = \{x_k\}_{k\in\mathbb{Z}_{\geq 0}}$ be any semi-infinite geodesic that satisfies $x_k/k \to \boldsymbol{u}$ as $k\to\infty$. Then $x_\bullet = \boldsymbol{b}^{\rho,x_0}$.*

(iii) Coalescence. *For all $x, y \in \mathbb{Z}^2$, the paths $\boldsymbol{b}^{\rho,x}$ and $\boldsymbol{b}^{\rho,y}$ coalesce: there exists $z = z^\rho(x, y) \in \mathbb{Z}^2$ such that $\boldsymbol{b}^{\rho,x} \cap \boldsymbol{b}^{\rho,y} = \boldsymbol{b}^{\rho,z}$.*

It is clear from the construction (2-13) that once $\boldsymbol{b}^{\rho,x}$ and $\boldsymbol{b}^{\rho,y}$ come together, they stay together. We call $z^{\rho(\boldsymbol{u})}(x, y)$ the *coalescence point* of the unique $\boldsymbol{u}$-directed semi-infinite geodesics from x and y. The Busemann function satisfies

$$B^\rho_{x,y} = G_{z^\rho(x,y),y} - G_{z^\rho(x,y),x} \quad \text{a.s.} \tag{2-15}$$

It is important to note that parts (ii) and (iii) of Theorem 2.4 are true with probability one only for a given $\boldsymbol{u}$ and not simultaneously for all directions.

Theorem 2.4(i) follows from an ergodic theorem for Busemann functions and the shape equation (2-3) (see for example Theorem 4.3 in [Georgiou et al. 2017a]). Theorem 2.4(ii)–(iii) were established for the exponential CGM in [Coupier 2011; Ferrari and Pimentel 2005]. The article [Seppäläinen 2020] gives an alternative derivation of Theorem 2.4 based on the properties of the stationary exponential CGM. Versions of Theorem 2.4 for the CGM with general weights appear in [Georgiou et al. 2017a].

Remark 2.5 (competition interface). The *geodesic tree* emanating from x consists of all the geodesics between x and points $y \in x + \mathbb{Z}^2_{\leq 0}$ south and west of x. The semi-infinite geodesics $\boldsymbol{b}^{\rho,x}$ are infinite rays in this tree. Every geodesic to x comes through either $x - \boldsymbol{e}_1$ or $x - \boldsymbol{e}_2$. This dichotomy splits the tree into two subtrees. Between the two subtrees lies a unique path $\{\varphi^x_n\}_{n\in\mathbb{Z}_{\geq 0}}$ on the dual lattice $\left(\frac12, \frac12\right) + \mathbb{Z}^2$ that starts at $\varphi^x_0 = x - \left(\frac12, \frac12\right)$. φ^x_n is a.s. uniquely defined as the point in $x - \left(\frac12, \frac12\right) + \mathbb{Z}^2_{\leq 0}$ that satisfies $|x - \varphi^x_n|_1 = n + 1$ and

$$G_{\varphi^x_n + (-\frac12, \frac12),\, x-\boldsymbol{e}_1} - G_{\varphi^x_n + (-\frac12, \frac12),\, x-\boldsymbol{e}_2} > 0 > G_{\varphi^x_n + (\frac12, -\frac12),\, x-\boldsymbol{e}_1} - G_{\varphi^x_n + (\frac12, -\frac12),\, x-\boldsymbol{e}_2}.$$

(Use the convention $G_{x,y} = -\infty$ if $x \leq y$ fails.) The competition interface has a random asymptotic direction,

$$\lim_{n\to\infty} \frac{\varphi^x_n}{n} = \boldsymbol{u}(\rho^*(x)) \qquad \text{almost surely,} \tag{2-16}$$

where the limit is described in (2-5) and Lemma 2.2. This was first proved in [Ferrari and Pimentel 2005]. The limit came from the study of geodesics with Newman's approach. Identification of the limit came via a mapping of φ^x to a second class particle in the rarefaction fan of TASEP whose limit had been identified in [Ferrari and Kipnis 1995]. An alternative proof that relies on the stationary LPP processes was given in [Georgiou et al. 2017a].

2C. ***Queues.*** We begin with a standard formulation of a queue that obeys FIFO (first-in-first-out) discipline. This treatment goes back to classic works of Lindley [1952] and Loynes [1962]. Modern references that connect queues with LPP include [Glynn and Whitt 1991; Baccelli et al. 2000; Draief et al. 2005].

The inputs are two bi-infinite sequences: the *arrival process* $I = (I_k)_{k\in\mathbb{Z}}$ and the *service process* $\omega = (\omega_j)_{j\in\mathbb{Z}}$ in $\mathbb{R}^{\mathbb{Z}}_{\geq 0}$. They are assumed to satisfy

$$\lim_{m\to-\infty} \sum_{i=m}^{0} (\omega_i - I_{i+1}) = -\infty. \tag{2-17}$$

The interpretation is that I_j is the time between the arrivals of customers $j-1$ and j and ω_j is the service time of customer j. From these inputs three outputs $\tilde{I} = (\tilde{I}_k)_{k\in\mathbb{Z}}$, $J = (J_k)_{k\in\mathbb{Z}}$ and $\widetilde{\omega} = (\widetilde{\omega}_k)_{k\in\mathbb{Z}}$, also elements of $\mathbb{R}^{\mathbb{Z}}_{\geq 0}$, are constructed as follows.

Let $G = (G_k)_{k\in\mathbb{Z}}$ be any function on $\mathbb{Z}$ that satisfies $I_k = G_k - G_{k-1}$. Define the sequence $\widetilde{G} = (\widetilde{G}_\ell)_{\ell\in\mathbb{Z}}$ by

$$\widetilde{G}_\ell = \sup_{k:\,k\leq\ell} \Big\{ G_k + \sum_{i=k}^{\ell} \omega_i \Big\}, \quad \ell \in \mathbb{Z}. \tag{2-18}$$

Under assumption (2-17) the supremum in (2-18) is assumed at some finite k. The *interdeparture time* between customers $\ell-1$ and ℓ is defined by

$$\tilde{I}_\ell = \widetilde{G}_\ell - \widetilde{G}_{\ell-1} \tag{2-19}$$

and the sequence $\tilde{I} = (\tilde{I}_k)_{k\in\mathbb{Z}}$ is the *departure process*. The *sojourn time* J_k of customer k is defined by

$$J_k = \widetilde{G}_k - G_k, \quad k \in \mathbb{Z}. \tag{2-20}$$

The third output,

$$\widetilde{\omega}_k = I_k \wedge J_{k-1}, \quad k \in \mathbb{Z}, \tag{2-21}$$

is the amount of time customer $k-1$ spends as the last customer in the queue.

$\tilde{I}$, J and $\widetilde{\omega}$ are well-defined nonnegative real sequences, and they do not depend on the choice of the function G as long as G has increments $I_k = G_k - G_{k-1}$. The three mappings are denoted by

$$\tilde{I} = D(I,\omega), \quad J = S(I,\omega), \quad \text{and} \quad \widetilde{\omega} = R(I,\omega). \tag{2-22}$$

The queueing story is good for imbuing the mathematics with meaning, but is not necessary for the sequel.

From

$$\widetilde{G}_k = \omega_k + G_k \vee \widetilde{G}_{k-1}, \tag{2-23}$$

follow the useful iterative equations

$$\tilde{I}_k = \omega_k + (I_k - J_{k-1})^+ \quad \text{and} \quad J_k = \omega_k + (J_{k-1} - I_k)^+. \tag{2-24}$$

The difference of the two equations above gives a “conservation law”,

$$I_k + J_k = J_{k-1} + \tilde{I}_k. \tag{2-25}$$

We extend the queueing operator D to mappings

$$D^{(n)} : (\mathbb{R}^{\mathbb{Z}}_{\geq 0})^n \to \mathbb{R}^{\mathbb{Z}}_{\geq 0}$$

of multiple sequences into a single sequence. Let $\zeta, \zeta^1, \zeta^2, \ldots$ denote elements of $\mathbb{R}^{\mathbb{Z}}_{\geq 0}$. Then, as long as the actions below are well-defined, let

$$\begin{aligned} D^{(1)}(\zeta) &= D(\zeta, 0) = \zeta, \\ D^{(2)}(\zeta^1, \zeta^2) &= D\big(D^{(1)}(\zeta^1), \zeta^2\big) = D(\zeta^1, \zeta^2), \\ D^{(3)}(\zeta^1, \zeta^2, \zeta^3) &= D\big(D^{(2)}(\zeta^1, \zeta^2), \zeta^3\big) = D\big(D(\zeta^1, \zeta^2), \zeta^3\big), \\ \text{and, in general,} \quad D^{(n)}(\zeta^1, \zeta^2, \dots, \zeta^n) &= D\big(D^{(n-1)}(\zeta^1, \dots, \zeta^{n-1}), \zeta^n\big) \quad \text{for } n \geq 2. \end{aligned} \tag{2-26}$$

In queueing terms, $D^{(n)}(\zeta^1, \zeta^2, \dots, \zeta^n)$ is the departure process that results from feeding arrival process ζ^1 through a series of $n-1$ service stations labeled $i = 2, 3, \dots, n$. For $i = 2, 3, \dots, n$, ζ^i is the service process at station i. Departures from station $i-1$ are the arrivals at station i. The final output is the departure process from the last station whose service process is ζ^n.

We record some inequalities which are to be understood coordinatewise: for example, $I' \geq I$ means that $I'_k \geq I_k$ for all $k \in \mathbb{Z}$.

Lemma 2.6. *Assuming that the mappings below are well-defined, we have the following inequalities*:

$$D(I, \omega) \geq \omega. \tag{2-27}$$

$$\textit{If } I' \geq I \textit{ then } D(I', \omega) \geq D(I, \omega). \tag{2-28}$$

$$\textit{For } n \geq 2,\ D^{(n)}(\zeta^1, \zeta^2, \zeta^3, \dots, \zeta^n) \geq D^{(n-1)}(\zeta^2, \zeta^3, \dots, \zeta^n). \tag{2-29}$$

Proof. The first part of (2-24) implies (2-27). For (2-28) observe that

$$J_k = \sup_{j:\, j \leq k} \Big\{ G_j - G_k + \sum_{i=j}^{k} \omega_i \Big\} \geq \sup_{j:\, j \leq k} \Big\{ G'_j - G'_k + \sum_{i=j}^{k} \omega_i \Big\} = J'_k.$$

Now (2-24) gives $\tilde{I}'_k \geq \tilde{I}_k$.

Inequality (2-29) comes by induction on n. The case $n = 2$ is (2-27). Then, by induction and (2-28),

$$\begin{aligned} D^{(n)}(\zeta^1, \dots, \zeta^n) &= D\big(D^{(n-1)}(\zeta^1, \dots, \zeta^{n-1}), \zeta^n\big) \geq D\big(D^{(n-2)}(\zeta^2, \dots, \zeta^{n-1}), \zeta^n\big) \\ &= D^{(n-1)}(\zeta^2, \dots, \zeta^n). \end{aligned}$$ □

We record the most basic fact about M/M/1 queues. The following notation will be used in the sequel. Let

$$\lambda = (\lambda_1, \dots, \lambda_n) \in (0, \infty)^n$$

be an n-tuple of positive reals. Let $\zeta = (\zeta^1, \dots, \zeta^n) \in (\mathbb{R}^{\mathbb{Z}}_{\geq 0})^n$ with $\zeta^i = (\zeta^i_k)_{k \in \mathbb{Z}}$ denote an n-tuple of nonnegative bi-infinite random sequences. Then ζ has distribution ν^λ if all the coordinates ζ^i_k are mutually independent with marginal distributions $\zeta^i_k \sim \text{Exp}(\lambda_i^{-1})$. In other words, ζ^i is a sequence of i.i.d. mean λ_i exponential variables, and the sequences are independent.

Lemma 2.7. *Let $n \geq 2$ and let $\lambda = (\lambda_1, \dots, \lambda_n)$ satisfy $\lambda_1 > \dots > \lambda_n > 0$. Let ζ have distribution ν^λ. Then $D^{(n)}(\zeta^1, \dots, \zeta^n)$ has distribution ν^{λ_1}, in other words, $D^{(n)}(\zeta^1, \dots, \zeta^n)$ is a sequence of i.i.d. mean λ_1 exponential random variables.*

Proof. The case $n = 2$ is in Lemma B.2. The general case follows by induction on n. □

3. Joint distribution of the Busemann functions

This section contains the main results on the joint distribution of the Busemann process

$$B^{\bullet} = \{B^{\rho} : 1 < \rho < \infty\}$$

defined in Theorem 2.1. Proofs are in Section 6. The distribution of the n-tuple

$$\{(B^{\rho_1}_{x-\boldsymbol{e}_1,x}, \dots, B^{\rho_n}_{x-\boldsymbol{e}_1,x})\}_{x\cdot \boldsymbol{e}_2=t}$$

on a given lattice level $t \in \mathbb{Z}$ comes through a mapping of a product of exponential distributions. This mapping is developed next.

3A. ***Coupled exponential distributions.*** Fix $n \in \mathbb{Z}_{>0}$ for the moment and define the following two spaces of n-tuples of nonnegative real sequences. The sequences themselves are denoted by $I^i = (I^i_k)_{k\in\mathbb{Z}}$ and $\eta^i = (\eta^i_k)_{k\in\mathbb{Z}}$ for $i \in [n]$:

$$\mathcal{Y}_n = \biggl\{ I = (I^1, I^2, \dots, I^n) \in (\mathbb{R}^{\mathbb{Z}}_{\ge 0})^n : \forall\, i \in [\![2, n]\!],\ \lim_{m\to-\infty} \frac{1}{|m|} \sum_{k=-m}^{0} I^i_k > \lim_{m\to-\infty} \frac{1}{|m|} \sum_{k=-m}^{0} I^{i-1}_k > 0 \biggr\}. \tag{3-1}$$

$$\mathcal{X}_n = \biggl\{ \eta = (\eta^1, \eta^2, \dots, \eta^n) \in (\mathbb{R}^{\mathbb{Z}}_{\ge 0})^n : \eta^i \ge \eta^{i-1}\ \forall\, i \in [\![2, n]\!] \text{ and } \varliminf_{m\to-\infty} \frac{1}{|m|} \sum_{k=-m}^{0} \eta^1_k > 0 \biggr\}. \tag{3-2}$$

The existence of the Cesàro limits as $m \to -\infty$ is part of the definitions. $\mathcal{Y}_n$ and $\mathcal{X}_n$ are Borel subsets of $(\mathbb{R}^{\mathbb{Z}}_{\ge 0})^n$ and thereby separable metric spaces in the product topology. We endow them with their Borel σ-algebras.

Define a mapping $\mathcal{D}^{(n)} : \mathcal{Y}_n \to \mathcal{X}_n$ in terms of the multiqueue mappings $D^{(k)}$ of (2-26) as follows: for $I = (I^1, I^2, \dots, I^n) \in \mathcal{Y}_n$, the image $\eta = (\eta^1, \eta^2, \dots, \eta^n) = \mathcal{D}^{(n)}(I)$ is defined by

$$\eta^i = D^{(i)}(I^i, I^{i-1}, \dots, I^1) \quad \text{for } i = 1, \dots, n. \tag{3-3}$$

In particular, the first sequence is just copied over: $\eta^1 = I^1$. Then $\eta^2 = D(I^2, I^1)$, $\eta^3 = D^{(3)}(I^3, I^2, I^1) = D(D(I^3, I^2), I^1)$, and so on. Iterated application of Lemma A.3 from Appendix A together with the assumption $I \in \mathcal{Y}_n$ ensures that the mappings $D^{(i)}(I^i, I^{i-1}, \dots, I^1)$ are well-defined. Furthermore, $\eta \in \mathcal{X}_n$ follows from inequalities (2-27) and (2-29). Lemma A.3 implies also that $\mathcal{D}^{(n)}$ maps $\mathcal{Y}_n$ into itself. We do not need this feature in the sequel, which is why we did not define $\mathcal{X}_n$ as a subspace of $\mathcal{Y}_n$.

Recall:

For $\rho = (\rho_1, \dots, \rho_n) \in (0, \infty)^n$, $I = (I^1, I^2, \dots, I^n)$ has distribution ν^{ρ} if
all coordinates I^i_k are independent and $I^i_k \sim \text{Exp}(\rho_i^{-1})$ for each $k \in \mathbb{Z}$ and $i \in [n]$. (3-4)

If ρ satisfies $0 < \rho_1 < \rho_2 < \cdots < \rho_n$ then ν^{ρ} is supported on $\mathcal{Y}_n$. For these ρ define the probability measure μ^{ρ} on $\mathcal{X}_n$ as the image of ν^{ρ} under $\mathcal{D}^{(n)}$:

$$\mu^{\rho} = \nu^{\rho} \circ (\mathcal{D}^{(n)})^{-1} \quad \text{for } \rho = (\rho_1, \rho_2, \dots, \rho_n) \text{ such that } 0 < \rho_1 < \rho_2 < \cdots < \rho_n. \tag{3-5}$$

By Lemma 2.7, if η has distribution μ^{ρ} with $0<\rho_1<\rho_2<\cdots<\rho_n$, then for each $i\in[n]$, $\eta^i=(\eta^i_k)_{k\in\mathbb{Z}}$ is a sequence of i.i.d. mean ρ_i exponential variables. The mapping $\mathcal{D}^{(n)}$ couples the variables η^i_k together so that $\eta^{i-1}_k\le\eta^i_k$ for all $i\in[\![2,n]\!]$ and $k\in\mathbb{Z}$.

Translations $\{\theta_\ell\}_{\ell\in\mathbb{Z}}$ act on n-tuples of sequences by $(\theta_\ell\eta)^i_k=\eta^i_{k+\ell}$ for $i\in[n]$ and $k,\ell\in\mathbb{Z}$. A *translation-ergodic* probability measure Q on $\mathcal{X}_n$ is invariant under $\{\theta_\ell\}$ and satisfies $Q(A)\in\{0,1\}$ for any Borel set $A\subset\mathcal{X}_n$ that is invariant under $\{\theta_\ell\}$ (and similarly for any other sequence space).

Theorem 3.1. *The probability measures μ^{ρ} are translation-ergodic and have the following properties*:

(i) Continuity. *The probability measure μ^{ρ} is weakly continuous as a function of ρ on the set of vectors that satisfy* $0<\rho_1<\rho_2<\cdots<\rho_n$.

(ii) Consistency. *If $(\eta^1,\dots,\eta^n)\sim\mu^{(\rho_1,\dots,\rho_n)}$, then $(\eta^1,\dots,\eta^{j-1},\eta^{j+1},\dots,\eta^n)\sim\mu^{(\rho_1,\dots,\rho_{j-1},\rho_{j+1},\dots,\rho_n)}$ for all $j\in[n]$.*

Continuity of $\rho\mapsto\mu^{\rho}$ is proved in Section 6. Translation-covariance of the queueing mappings $(D(\theta_\ell I,\theta_\ell\omega)=\theta_\ell D(I,\omega))$ implies that μ^{ρ} inherits the translation-ergodicity of ν^{ρ}. We omit the proof of consistency. Consistency will be an indirect consequence of the uniqueness of μ^{ρ} as the translation-ergodic invariant distribution of the so-called coupled process (Theorem 5.3).

3B. ***Distribution of Busemann functions.*** Return to the Busemann functions $B^{\bullet}$ defined in Theorem 2.1. For each level $t\in\mathbb{Z}$ define the level-t sequence of weights $\overline{Y}_t=(Y_{(k,t)})_{k\in\mathbb{Z}}$ and for a given $\rho\in(1,\infty)$, sequences of $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ Busemann variables at level t:

$$\overline{B}^{\rho,\boldsymbol{e}_1}_t=(B^{\rho}_{(k-1,t),(k,t)})_{k\in\mathbb{Z}}\quad\text{and}\quad\overline{B}^{\rho,\boldsymbol{e}_2}_t=(B^{\rho}_{(k,t-1),(k,t)})_{k\in\mathbb{Z}}.$$

The next main result characterizes uniquely the distribution of the joint process $(Y,B^{\bullet})$ of weights and Busemann functions.

Theorem 3.2. *Let $Y=(Y_x)_{x\in\mathbb{Z}^2}$ be i.i.d.* Exp(1) *variables as in Section 2A. Let $1<\rho_1<\cdots<\rho_n$. Then at each level $t\in\mathbb{Z}$, the $(n+1)$-tuple of sequences $(\overline{Y}_t,\overline{B}^{\rho_1,\boldsymbol{e}_1}_t,\dots,\overline{B}^{\rho_n,\boldsymbol{e}_1}_t)$ has distribution $\mu^{(1,\rho_1,\dots,\rho_n)}$.*

Once the process $\{\overline{B}^{\rho,\boldsymbol{e}_1}_{t-1}\}_{1<\rho<\infty}$ on a single level $t-1$ is given, the variables $B^{\rho}_{x-\boldsymbol{e}_i,x}$ at higher levels $t,t+1,t+2,\dots$ can be deduced by drawing independent weights $\overline{Y}_t,\overline{Y}_{t+1},\dots$ and by applying queueing mappings. By stationarity, the full distribution will then have been determined. The next lemma describes the single step of computing the $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ Busemann increments on level t from the process $\{\overline{B}^{\rho,\boldsymbol{e}_1}_{t-1}\}_{1<\rho<\infty}$ and independent level-t weights $\overline{Y}_t$. The mappings D and S were specified in (2-22).

Lemma 3.3. *There exists an event of full probability on which*

$$\overline{B}^{\rho,\boldsymbol{e}_1}_t=D(\overline{B}^{\rho,\boldsymbol{e}_1}_{t-1},\overline{Y}_t)\quad\textit{and}\quad\overline{B}^{\rho,\boldsymbol{e}_2}_t=S(\overline{B}^{\rho,\boldsymbol{e}_1}_{t-1},\overline{Y}_t)\quad\textit{for all }\rho\in(1,\infty)\textit{ and }t\in\mathbb{Z}.$$

The remainder of this section describes some distributional properties of $B^{\bullet}$ restricted to horizontal edges and lines on $\mathbb{Z}^2$. The corresponding statements for vertical edges and lines are obtained by replacing ρ with $\rho/(\rho-1)$. This is due to the distributional equality $\{B^{\rho}_{x-\boldsymbol{e}_2,x}\}_{x\in\mathbb{Z}^2}\overset{d}{=}\{B^{\rho/(\rho-1)}_{Rx-\boldsymbol{e}_1,Rx}\}_{x\in\mathbb{Z}^2}$ where $R(x_1,x_2)=(x_2,x_1)$. This follows from (2-5) and the limits (2-10), by reflecting the lattice across the diagonal.

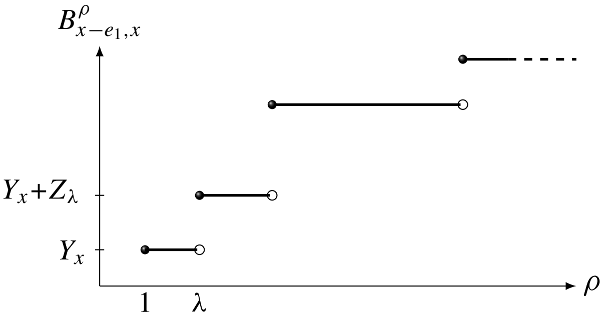


Figure 2. A sample path of the pure jump process $\{B^{\rho}_{x-e_1,x}\}_{\rho\in[1,\infty)}$, with initial value $B^{1}_{x-e_1,x}=Y_x$. The jump times are a Poisson point process on $(1,\infty)$ with intensity $s^{-1}ds$. Given that there is a jump at λ, the jump size is an independent $\text{Exp}(\lambda^{-1})$ variable Z_λ.

3C. ***Marginal distribution on a single edge.*** Lemma 2.2 implies that for a fixed horizontal edge $(x-\boldsymbol{e}_1, x)$ we can extend $\{B^{\rho}_{x-\boldsymbol{e}_1,x} : 1<\rho<\infty\}$ to a cadlag process

$$B^{\bullet}_{x-\boldsymbol{e}_1,x}=\{B^{\rho}_{x-\boldsymbol{e}_1,x} : 1\le\rho<\infty\}$$

by setting $B^{1}_{x-\boldsymbol{e}_1,x}=Y_x$. We describe the distribution of this process in terms of a marked point process. Figure 2 illustrates a sample path of this process.

Let N be the simple point process on the interval $\{s : 1\le s<\infty\}$ that has a point at $s=1$ with probability one, and on the open interval $(1,\infty)$ N is a Poisson point process with parameter measure $s^{-1}\,ds$. (We use N to denote both the random discrete set of locations and the resulting random point measure.) Let N be the ground process of the marked point process $\sum_{t\in N}\delta_{(t,\,Z_t)}$ where the mark Z_t at location $t\in N$ is $\text{Exp}(t^{-1})$-distributed and independent of the other marks. Define the nondecreasing cadlag process $X(\cdot)=\{X(\rho):\rho\in[1,\infty)\}$ as

$$X(\rho)=\sum_{t\,\in\,N\cap[1,\rho]} Z_t; \tag{3-6}$$

namely, $X(\rho)$ is the total weight of the marks in $[1,\rho]$. The Laplace transform of $X(\rho)$ is given in (6-12).

Theorem 3.4. *Fix $x\in\mathbb{Z}^2$. The nondecreasing cadlag processes $B^{\bullet}_{x-\boldsymbol{e}_1,x}$ and $X(\cdot)$ indexed by* $[1,\infty)$ *are equal in distribution.*

A qualitative consequence of Theorem 3.4 is that for any given $\lambda\in(1,\infty)\setminus N$, $\rho\mapsto B^{\rho}_{x-\boldsymbol{e}_1,x}$ is constant in an interval around λ. From identity (2-15), it is evident that this is due to the fact that the coalescence point function $\rho\mapsto z^{\rho}(x-\boldsymbol{e}_1,x)$ is constant in an interval. This is an analogue of the local constancy of the difference weight profile in Theorem 1.1 of [Basu et al. 2019]. The implications of Theorem 3.4 for the coalescence structure of the geodesics of the CGM are explored in [Janjigian et al. 2019].

Theorem 3.4 is proved by establishing that $B^{\bullet}_{x-\boldsymbol{e}_1,x}$ has independent increments and by deducing the distribution of an increment. Independent increments means that for $1=\rho_0<\rho_1<\cdots<\rho_n$, the random variables $Y_x=B^{\rho_0}_{x-\boldsymbol{e}_1,x}, B^{\rho_1}_{x-\boldsymbol{e}_1,x}-B^{\rho_0}_{x-\boldsymbol{e}_1,x},\dots,B^{\rho_n}_{x-\boldsymbol{e}_1,x}-B^{\rho_{n-1}}_{x-\boldsymbol{e}_1,x}$ are independent. For $1\le\lambda<\rho<\infty$, the distribution of the increment is

$$\begin{aligned}&\mathbb{P}\{B^{\rho}_{x-\boldsymbol{e}_1,x}-B^{\lambda}_{x-\boldsymbol{e}_1,x}=0\}=\mathbb{P}\{N(\lambda,\rho]=0\}=\frac{\lambda}{\rho},\\&\mathbb{P}\{B^{\rho}_{x-\boldsymbol{e}_1,x}-B^{\lambda}_{x-\boldsymbol{e}_1,x}>s\}=\Big(1-\frac{\lambda}{\rho}\Big)e^{-s/\rho}\quad\text{for }\ s>0.\end{aligned}\tag{3-7}$$

For the process $B^{\bullet}_{x-\boldsymbol{e}_2,x}$ on a vertical edge, the result of Theorem 3.4 is that

$$\{B^{\rho}_{x-\boldsymbol{e}_2,x}:1<\rho<\infty\}\overset{d}{=}\Big\{X\Big(\Big(\frac{\rho}{\rho-1}\Big)+\Big):1<\rho<\infty\Big\}.\tag{3-8}$$

3D. ***Marginal distribution on a level of the lattice.*** A striking and useful property of the Busemann process $\{B^{\rho}_{(k-1,t),(k,t)}\}_{k\in\mathbb{Z}}$ along a horizontal line in $\mathbb{Z}^2$ for a *fixed* value $\rho\in(1,\infty)$ is that the variables $\{B^{\rho}_{(k-1,t),(k,t)}\}_{k\in\mathbb{Z}}$ are i.i.d. (part (ii) of Theorem 2.1). For example, [Balázs et al. 2006] used this feature heavily to deduce the KPZ fluctuation exponents of the corner growth model. The next theorem shows that this property breaks down totally already for the joint process $\{(B^{\lambda}_{(k-1,t),(k,t)},B^{\rho}_{(k-1,t),(k,t)})\}_{k\in\mathbb{Z}}$ for two parameter values $\lambda<\rho$. Namely, this pair process is not even a Markov chain and not reversible. However, if we restrict attention to the differences $B^{\rho}_{(k-1,t),(k,t)}-B^{\lambda}_{(k-1,t),(k,t)}$, we can recover the reversibility. The differences are of interest because they indicate a jump in the coalescence point $z^{\bullet}((k-1,t),(k,t))$ in (2-15) as a function of the direction.

For the statement of the theorem below, the negative part of a real number is $x^-=(-x)\vee 0$. The Markov chain X_k in part (a) below has a queueing interpretation as the difference between the sojourn time of customer $k-1$ and the waiting time till the arrival of customer k. The details are in the proof in Lemma 6.5.

Theorem 3.5. *Let* $1\le\lambda<\rho<\infty$.

(a) *The sequence of differences* $\{B^{\rho}_{(k-1,t),(k,t)}-B^{\lambda}_{(k-1,t),(k,t)}\}_{k\in\mathbb{Z}}$ *is not a Markov chain, but there exists a stationary reversible Markov chain* $\{X_k\}_{k\in\mathbb{Z}}$ *such that this distributional equality of processes holds*:

$$\{B^{\rho}_{(k-1,t),(k,t)}-B^{\lambda}_{(k-1,t),(k,t)}\}_{k\in\mathbb{Z}}\overset{d}{=}\{X_k^-\}_{k\in\mathbb{Z}}.$$

In particular, the process of differences is reversible:

$$\{B^{\rho}_{(k-1,t),(k,t)}-B^{\lambda}_{(k-1,t),(k,t)}\}_{k\in\mathbb{Z}}\overset{d}{=}\{B^{\rho}_{(-k-1,t),(-k,t)}-B^{\lambda}_{(-k-1,t),(-k,t)}\}_{k\in\mathbb{Z}}.$$

(b) *The sequence of pairs* $\{(B^{\lambda}_{(k-1,t),(k,t)},B^{\rho}_{(k-1,t),(k,t)})\}_{k\in\mathbb{Z}}$ *is not a Markov chain. The joint distribution of two successive pairs*

$$\big((B^{\lambda}_{(k-1,t),(k,t)},B^{\rho}_{(k-1,t),(k,t)}),(B^{\lambda}_{(k,t),(k+1,t)},B^{\rho}_{(k,t),(k+1,t)})\big)$$

*is **not** the same as the joint distribution of its transpose*

$$\big((B^{\lambda}_{(k,t),(k+1,t)},B^{\rho}_{(k,t),(k+1,t)}),(B^{\lambda}_{(k-1,t),(k,t)},B^{\rho}_{(k-1,t),(k,t)})\big).$$

In particular, the process of pairs $\{(B^{\lambda}_{(k-1,t),(k,t)},B^{\rho}_{(k-1,t),(k,t)})\}_{k\in\mathbb{Z}}$ *is **not** reversible.*

3E. ***The initial segment of the Busemann geodesic.*** As the last application of Theorem 3.2 we calculate the probability distribution of the length of the initial horizontal run of a semi-infinite geodesic.

Let $\boldsymbol{a}_x^\rho = \boldsymbol{b}_1^{\rho,x} - x$ be the first step of the B^ρ Busemann geodesic (2-13) started at x. $\{\boldsymbol{a}_x^\rho\}_{x\in\mathbb{Z}^2}$ is a random configuration with values in $\{-\boldsymbol{e}_1, -\boldsymbol{e}_2\}$. By weight recovery (2-8), $\boldsymbol{a}_x^\rho = -\boldsymbol{e}_1$ if and only if $B^\rho_{x-\boldsymbol{e}_1,x} - Y_x = 0$. Hence by Theorem 3.5(a) with $\lambda = 1$, reversibility holds along a line: $\{\boldsymbol{a}^\rho_{k\boldsymbol{e}_i}\}_{k\in\mathbb{Z}} \overset{d}{=} \{\boldsymbol{a}^\rho_{-k\boldsymbol{e}_i}\}_{k\in\mathbb{Z}}$.

The first part of the theorem below gives a queueing characterization for the process $\{\boldsymbol{a}^\rho_{k\boldsymbol{e}_1}\}_{k\in\mathbb{Z}}$. To that end, for the queueing mapping $\tilde{I} = D(I, \omega)$ of (2-22) define the indicator variables

$$\eta_k = \mathbf{1}_{\tilde{I}_k = \omega_k} = \mathbf{1}\{\text{customer } k \text{ has to wait before entering service}\}. \tag{3-9}$$

Let

$$\xi_x = \inf\{k \in \mathbb{Z}_{\geq 0} : \boldsymbol{a}^\rho_{x-k\boldsymbol{e}_1} = -\boldsymbol{e}_2\}$$

denote the number of consecutive $-\boldsymbol{e}_1$ steps that $\boldsymbol{b}^{\rho,x}$ takes from a deterministic starting point x. Part (b) of the theorem gives the distribution of ξ_x. The *Catalan triangle* $\{C(n,k) : 0 \leq k \leq n\}$ is given by

$$C(n,k) = \frac{(n+k)!(n-k+1)}{k!(n+1)!}. \tag{3-10}$$

Information about $C(n,k)$ is given above Lemma B.3 in Appendix B.

Theorem 3.6. *Let* $1 < \rho < \infty$.

(a) *Let the service and arrival processes satisfy* $(\omega, I) \sim \nu^{(1,\rho)}$ *and define* η_k *by* (3-9). *Then we have the distributional equality*

$$\{\mathbf{1}\{\boldsymbol{a}^\rho_{k\boldsymbol{e}_1} = -\boldsymbol{e}_1\}\}_{k\in\mathbb{Z}} \overset{d}{=} \{\eta_k\}_{k\in\mathbb{Z}}.$$

(b) *Let* $x \in \mathbb{Z}^2$. *Then* $\mathbb{P}\{\xi_x = 0\} = 1 - \rho^{-1}$ *and for* $n \in \mathbb{Z}_{>0}$,

$$\mathbb{P}\{\xi_x = n\} = (1-\rho^{-1}) \sum_{k=0}^{n-1} C(n-1,k) \frac{\rho^k}{(\rho+1)^{n+k}}. \tag{3-11}$$

The distribution in (3-11) is proper; that is, $\sum_{n\in\mathbb{Z}_{\geq 0}} \mathbb{P}\{\xi_x = n\} = 1$. This follows for example from Theorem 2.4(i) according to which the Busemann geodesic has direction strictly off the axes.

Remark 3.7. If we take $(\omega, I) \sim \nu^{(\lambda,\rho)}$ for $1 < \lambda < \rho$ in Theorem 3.6 and define η_k again by (3-9), we get the distributional equality $\{\mathbf{1}\{B^\rho_{(k-1,t),(k,t)} = B^\lambda_{(k-1,t),(k,t)}\}\}_{k\in\mathbb{Z}} \overset{d}{=} \{\eta_k\}_{k\in\mathbb{Z}}$. The calculation that produced part (b) gives the distribution $\mathbb{P}\{\xi_x^{\lambda,\rho} = 0\} = (\rho - \lambda)/\rho$ and

$$\mathbb{P}\{\xi_x^{\lambda,\rho} = n\} = \frac{\rho - \lambda}{\rho} \sum_{k=0}^{n-1} C(n-1,k) \frac{\rho^k \lambda^n}{(\lambda+\rho)^{n+k}} \qquad \text{for } n \in \mathbb{Z}_{>0},$$

for the random variable

$$\xi_x^{\lambda,\rho} = \inf\{k \in \mathbb{Z}_{\geq 0} : B^\rho_{x-(k+1)\boldsymbol{e}_1, x-k\boldsymbol{e}_1} > B^\lambda_{x-(k+1)\boldsymbol{e}_1, x-k\boldsymbol{e}_1}\}.$$

Note that $B^\rho_{x-\boldsymbol{e}_1,x} = B^\lambda_{x-\boldsymbol{e}_1,x}$ tells us that $\boldsymbol{b}_1^{\rho,x} = \boldsymbol{b}_1^{\lambda,x}$ but not which step is chosen.

4. Properties of queueing mappings

This section proves a property of the queueing mapping D (Lemma 4.4) on which the intertwining property that comes in Section 5D rests. To prove Lemma 4.4 we develop a duality in the queueing setting of Section 2C: namely, an LPP process defined in terms of weights (I, ω) can be equivalently described in terms of weights $(\tilde{I}, \widetilde{\omega})$ defined by (2-22). Routine facts about the queueing mappings are collected in Appendix A.

Fix an origin $m \in \mathbb{Z}$. Assume given nonnegative real weights

$$J_m, \quad (I_i)_{i \geq m+1}, \quad \text{and} \quad (\omega_i)_{i \geq m+1}. \tag{4-1}$$

From these define iteratively for $k = m+1, m+2, \ldots$

$$\tilde{I}_k = \omega_k + (I_k - J_{k-1})^+, \quad J_k = \omega_k + (J_{k-1} - I_k)^+, \quad \text{and} \quad \widetilde{\omega}_k = I_k \wedge J_{k-1}. \tag{4-2}$$

There is a duality or reversibility of sorts here. For a fixed k, equations (4-2) are equivalent to

$$I_k = \widetilde{\omega}_k + (\tilde{I}_k - J_k)^+, \quad J_{k-1} = \widetilde{\omega}_k + (J_k - \tilde{I}_k)^+, \quad \text{and} \quad \omega_k = \tilde{I}_k \wedge J_k. \tag{4-3}$$

We turn this reversibility into a lemma as follows. Restrict the given J, I and ω weights in (4-1) to the interval $[\![m, n]\!]$. Then on the interval $[\![-n, -m]\!]$ define the given weights J'_{-n}, $(I'_i)_{-n+1 \leq i \leq -m}$ and $(\omega'_i)_{-n+1 \leq i \leq -m}$ as

$$I'_i = \tilde{I}_{-i+1}, \quad J'_{-n} = J_n, \quad \text{and} \quad \omega'_i = \widetilde{\omega}_{-i+1}. \tag{4-4}$$

Now apply (4-2) to these given weights to compute $(\widetilde{I'_k}, J'_k, \widetilde{\omega}'_k)$ for $k \in [\![-n+1, -m]\!]$. First assume by induction that $J'_{k-1} = J_{-k+1}$. The base case $k-1 = -n$ is covered by the definition in (4-4). Then

$$\begin{aligned} J'_k &= \omega'_k + (J'_{k-1} - I'_k)^+ = \widetilde{\omega}_{-k+1} + (J_{-k+1} - \tilde{I}_{-k+1})^+ \\ &= I_{-k+1} \wedge J_{-k} + (J_{-k} - I_{-k+1})^+ = J_{-k}. \end{aligned}$$

The third equality above used the definition of $\widetilde{\omega}$ in (4-2) and the conservation law

$$I_k + J_k = J_{k-1} + \tilde{I}_k \tag{4-5}$$

that follows from (4-2). Thus $J'_k = J_{-k}$ for all $k \in [\![-n, -m]\!]$. Next

$$\begin{aligned} \tilde{I}'_k &= \omega'_k + (I'_k - J'_{k-1})^+ = \widetilde{\omega}_{-k+1} + (\widetilde{I}_{-k+1} - J_{-k+1})^+ \\ &= I_{-k+1} \wedge J_{-k} + (I_{-k+1} - J_{-k})^+ = I_{-k+1}. \end{aligned}$$

Finally,

$$\widetilde{\omega}'_k = I'_k \wedge J'_{k-1} = \tilde{I}_{-k+1} \wedge J_{-k+1} = \omega_{-k+1}$$

as follows again from (4-2). We summarize this finding as follows.

Lemma 4.1. *Fix $m < n$. Assume given J_m, $(I_i)_{m+1 \leq i \leq n}$ and $(\omega_i)_{m+1 \leq i \leq n}$. Compute $(\tilde{I}_k, J_k, \widetilde{\omega}_k)_{m+1 \leq k \leq n}$ from (4-2). Then define J'_{-n}, $(I'_i)_{-n+1 \leq i \leq -m}$ and $(\omega'_i)_{-n+1 \leq i \leq -m}$ by (4-4) and apply (4-2) to compute $(\tilde{I}'_k, J'_k, \widetilde{\omega}'_k)_{-n+1 \leq k \leq -m}$. The conclusion is that $(\widetilde{I'_k}, J'_k, \widetilde{\omega}'_k) = (I_{-k+1}, J_{-k}, \omega_{-k+1})$ for $k \in [\![-n+1, -m]\!]$.*

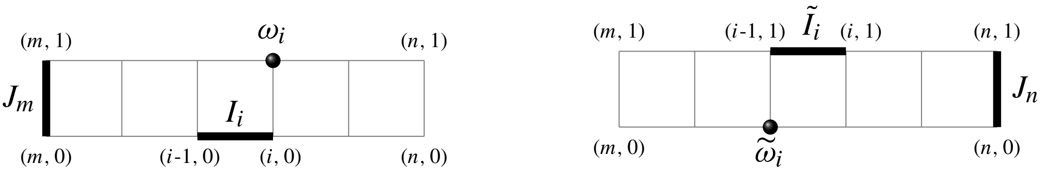


Figure 3. Illustration of the weights (I, J, ω) on the left and weights $(\tilde I, J, \widetilde\omega)$ on the right. Pairs $(k, a) \in [\![m, n]\!] \times [\![0, 1]\!]$ mark vertices of the two-level strip.

Next we use the weights given in (4-1) to construct a last-passage process on the two-level strip $[\![m, \infty[\![\times [\![0, 1]\!]$ in $\mathbb{Z}^2$. In this construction, I_i serves as a weight on the horizontal edge $((i-1, 0), (i, 0))$ on the lower 0-level, J_m is a weight on the vertical edge $((m, 0), (m, 1))$, and ω_i is a weight at vertex $(i, 1)$ on the upper 1-level. (The left diagram of Figure 3 illustrates this.) The last-passage values $H_{(m,0),(n,a)}$ are defined for $(n, a) \in [\![m, \infty[\![\times [\![0, 1]\!]$ as follows:

$$\begin{aligned}
&H_{(m,0),(m,0)} = 0 \quad\text{and}\quad H_{(m,0),(n,0)} = \sum_{i=m+1}^{n} I_i \quad \text{for } n > m,\\
&H_{(m,0),(m,1)} = J_m\,, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (4\text{-}6)\\
&H_{(m,0),(n,1)} = \Big\{J_m + \sum_{i=m+1}^{n} \omega_i\Big\} \vee \max_{m+1\le j\le n}\Big\{\sum_{i=m+1}^{j} I_i + \sum_{i=j}^{n}\omega_i\Big\}, \qquad n > m.
\end{aligned}$$

If the given weights (4-1) come from the queueing setting of Section 2C, then $H_{(m,0),(n,1)} = \widetilde G_n - G_m$. But this connection is not needed for the present.

The next lemma gives alternative formulas for H in terms of the weights calculated in (4-2). Pictorially, imagine $\tilde I_i$ as a weight on the edge $((i-1, 1), (i, 1))$ and $\widetilde\omega_i$ as a weight on the vertex $(i-1, 0)$. (The right diagram of Figure 3 illustrates this.) In (4-7), a sum expression of the form $a_j + \cdots + a_{j-1}$ is interpreted as zero. The equation (4-8) makes sense also for $\ell = n$ in which case the right-hand side simplifies to J_n.

Lemma 4.2. *Let $m \le n$. Then*

$$H_{(m,0),(n,1)} = I_{m+1} + \cdots + I_k + J_k + \tilde I_{k+1} + \cdots + \tilde I_n \quad \textit{for each } k \in [\![m, n]\!]. \tag{4-7}$$

For each $\ell \in [\![m, n-1]\!]$,

$$H_{(m,0),(n,1)} - H_{(m,0),(\ell,0)} = \max_{\ell+1\le j\le n}\Big\{\sum_{i=\ell+1}^{j} \widetilde\omega_i + \sum_{i=j}^{n} \tilde I_i\Big\} \vee \Big\{\sum_{i=\ell+1}^{n} \widetilde\omega_i + J_n\Big\}. \tag{4-8}$$

We make some observations before the proof. By the two top lines of (4-6), equivalent to (4-7) are the increment formulas (for all $n > m$)

$$\tilde I_n = H_{(m,0),(n,1)} - H_{(m,0),(n-1,1)} \quad\text{and}\quad J_n = H_{(m,0),(n,1)} - H_{(m,0),(n,0)}. \tag{4-9}$$

Taking $\ell = m$ in (4-8) gives this dual representation for H:

$$H_{(m,0),(n,1)} = \max_{m+1\le j\le n}\Big\{\sum_{i=m+1}^{j} \widetilde\omega_i + \sum_{i=j}^{n} \tilde I_i\Big\} \vee \Big\{\sum_{i=m+1}^{n} \widetilde\omega_i + J_n\Big\}. \tag{4-10}$$

Proof of Lemma 4.2. Let $m < n$ and develop the definition (4-6). As in (2-2),

$$H_{(m,0),(n,1)} = \Big\{ J_m + \sum_{i=m+1}^{n-1} \omega_i \Big\} \vee \max_{m+1\le j\le n-1} \Big\{ \sum_{i=m+1}^{j} I_i + \sum_{i=j}^{n-1} \omega_i \Big\} \vee \Big\{ \sum_{i=m+1}^{n} I_i \Big\} + \omega_n \tag{4-11}$$
$$= H_{(m,0),(n-1,1)} \vee H_{(m,0),(n,0)} + \omega_n .$$

Set temporarily

$$A_n = H_{(m,0),(n,1)} - H_{(m,0),(n-1,1)} \quad \text{and} \quad B_n = H_{(m,0),(n,1)} - H_{(m,0),(n,0)} .$$

Then (4-11) gives the iterative equations

$$A_n = \omega_n + (I_n - B_{n-1})^+ \quad \text{and} \quad B_n = \omega_n + (B_{n-1} - I_n)^+ .$$

Definition (4-6) gives $B_m = J_m$. This starts an induction. Apply the equations above together with (4-2) to obtain $A_n = \tilde{I}_n$ and $B_n = J_n$ for all $n \ge m+1$. This establishes (4-9), and (4-7) follows.

We prove (4-8) by induction as ℓ decreases. The base case $\ell = n$ comes from the just proved $B_n = J_n$. Assume (4-8) for $\ell + 1$. Then for ℓ the right-hand side of (4-8) equals

$$\widetilde{\omega}_{\ell+1} + \Big\{ \sum_{i=\ell+1}^{n} \tilde{I}_i \Big\} \vee \max_{\ell+2\le j\le n} \Big\{ \sum_{i=\ell+2}^{j} \widetilde{\omega}_i + \sum_{i=j}^{n} \tilde{I}_i \Big\} \vee \Big\{ \sum_{i=\ell+2}^{n} \widetilde{\omega}_i + J_n \Big\}$$
$$= H_{(m,0),(\ell+1,0)} \wedge H_{(m,0),(\ell,1)} - H_{(m,0),(\ell,0)} + \big\{ H_{(m,0),(n,1)} - H_{(m,0),(\ell,1)} \big\} \vee \big\{ H_{(m,0),(n,1)} - H_{(m,0),(\ell+1,0)} \big\}$$
$$= H_{(m,0),(n,1)} - H_{(m,0),(\ell,0)} .$$

In the first equality we used $\widetilde{\omega}_{\ell+1} = I_{\ell+1} \wedge J_\ell$, (4-9) and the induction assumption. □

The last line of (4-6) and formula (4-10) give dual representations of the quantity $H_{(m,0),(n,1)}$. The next lemma shows that equality persists if we drop the terms that involve J from both formulas. This statement is the crucial ingredient of Lemma 4.4 below.

Lemma 4.3. *Let $m \le n$ in $\mathbb{Z}$. Assume given nonnegative weights J_{m-1}, $(I_i)_{m\le i\le n}$ and $(\omega_i)_{m\le i\le n}$. Compute $(\tilde{I}_k, J_k, \widetilde{\omega}_k)_{m\le k\le n}$ from (4-2). Define*

$$T_{m,n} = \max_{m\le j\le n} \Big\{ \sum_{i=m}^{j} I_i + \sum_{i=j}^{n} \omega_i \Big\} \quad \textit{and} \quad \widetilde{T}_{m,n} = \max_{m\le j\le n} \Big\{ \sum_{i=m}^{j} \widetilde{\omega}_i + \sum_{i=j}^{n} \tilde{I}_i \Big\}. \tag{4-12}$$

Then $T_{m,n} = \widetilde{T}_{m,n}$.

Proof. The case $m = n$ is the identity $I_n + \omega_n = \widetilde{\omega}_n + \tilde{I}_n$ that follows from (4-2).

Let $n \ge m+1$ and assume by induction that $\widetilde{T}_{m,n-1} \le T_{m,n-1}$. Develop the definitions.

$$T_{m,n} = \max_{m\le j\le n-1} \Big\{ \sum_{i=m}^{j} I_i + \sum_{i=j}^{n-1} \omega_i \Big\} \vee \Big\{ \sum_{i=m}^{n} I_i \Big\} + \omega_n = T_{m,n-1} \vee \Big\{ \sum_{i=m}^{n} I_i \Big\} + \omega_n . \tag{4-13}$$

Similarly,

$$\begin{aligned}\widetilde{T}_{m,n} &= \widetilde{T}_{m,n-1} \vee \left\{\sum_{i=m}^{n} \widetilde{\omega}_i\right\} + \tilde{I}_n = \widetilde{T}_{m,n-1} \vee \left\{\sum_{i=m}^{n} (I_i \wedge J_{i-1})\right\} + \omega_n + (I_n - J_{n-1})^+ \\ &\le T_{m,n-1} \vee \left\{\sum_{i=m}^{n} I_i\right\} + \omega_n + (I_n - J_{n-1})^+.\end{aligned} \tag{4-14}$$

The induction assumption was used in the last step.

Case 1: $I_n \le J_{n-1}$. This assumption kills the last term of (4-14) and gives

$$\widetilde{T}_{m,n} \le T_{m,n-1} \vee \left\{\sum_{i=m}^{n} I_i\right\} + \omega_n = T_{m,n}.$$

Case 2: $I_n > J_{n-1}$. For this case induction is not needed. We use the last-passage process $H_{(m-1,0),(\bullet,\bullet)}$. Conservation law (4-5) and (4-9) imply

$$I_n > J_{n-1} \iff \tilde{I}_n > J_n \iff H_{(m-1,0),(n-1,1)} < H_{(m-1,0),(n,0)}.$$

Then by (4-11),

$$H_{(m-1,0),(n,1)} = H_{(m-1,0),(n,0)} + \omega_n = \sum_{i=m}^{n} I_i + \omega_n \le T_{m,n}.$$

On the other hand, by definition (4-6),

$$H_{(m-1,0),(n,1)} = \left\{J_{m-1} + \sum_{i=m}^{n} \omega_i\right\} \vee T_{m,n}.$$

Hence $H_{(m-1,0),(n,1)} = T_{m,n}$. By the dual formula (4-10),

$$H_{(m-1,0),(n,1)} = \widetilde{T}_{m,n} \vee \left\{\sum_{i=m}^{n} \widetilde{\omega}_i + J_n\right\} \ge \widetilde{T}_{m,n}.$$

We conclude that in Case 2, $\widetilde{T}_{m,n} \le T_{m,n}$.

We have shown that $\widetilde{T}_{m,n} \le T_{m,n}$. This suffices for the proof by the duality in Lemma 4.1 because the roles of $T_{m,n}$ and $\widetilde{T}_{m,n}$ can be switched around. □

The next lemma is the key property of the queueing mapping D that underlies our results. Its proof relies on Lemma 4.3. Lemma 4.3 applies to the queueing setting described in Section 2C because equations (2-21) and (2-24) ensure that the assumptions of Lemma 4.3 are satisfied.

Lemma 4.4. *Assume given three sequences $I^2, I^1, \omega^1 \in \mathbb{R}^{\mathbb{Z}}_{\ge 0}$ such that the queueing operations below are well-defined. Let $\omega^2 = R(I^1, \omega^1)$ as defined in (2-21). Then we have the identity*

$$D\big(D(I^2, \omega^2), D(I^1, \omega^1)\big) = D\big(D(I^2, I^1), \omega^1\big). \tag{4-15}$$

Proof. Choose G^1 and G^2 so that $I^t_k = G^t_k - G^t_{k-1}$ for $t = 1, 2$. Let

$$H_j = \sup_{\ell:\,\ell\le j}\biggl\{G^2_\ell + \sum_{i=\ell}^{j} I^1_i\biggr\}$$

and then

$$\widetilde{H}_k = \sup_{j:\,j\le k}\biggl\{H_j + \sum_{i=j}^{k}\omega^1_i\biggr\} = \sup_{\ell:\,\ell\le k}\biggl\{G^2_\ell + \max_{j:\ell\le j\le k}\biggl[\sum_{i=\ell}^{j} I^1_i + \sum_{i=j}^{k}\omega^1_i\biggr]\biggr\}. \tag{4-16}$$

The sequence $(\widetilde{H}_k - \widetilde{H}_{k-1})_{k\in\mathbb{Z}}$ is the output $D(D(I^2, I^1), \omega^1)$.

For the left-hand side of (4-15) define first for $D(I^t, \omega^t)$ the sequence

$$\widetilde{G}^t_j = \sup_{\ell:\,\ell\le j}\biggl\{G^t_\ell + \sum_{i=\ell}^{j}\omega^t_i\biggr\}, \quad t\in\{1,2\}.$$

Set $\tilde{I}^1_k = \widetilde{G}^1_k - \widetilde{G}^1_{k-1}$. The output $D(D(I^2, \omega^2), D(I^1, \omega^1))$ is given by the increments of the sequence

$$\widehat{H}_k = \sup_{j:\,j\le k}\biggl\{\widetilde{G}^2_j + \sum_{i=j}^{k}\tilde{I}^1_i\biggr\} = \sup_{\ell:\,\ell\le k}\biggl\{G^2_\ell + \max_{j:\ell\le j\le k}\biggl[\sum_{i=\ell}^{j}\omega^2_i + \sum_{i=j}^{k}\tilde{I}^1_i\biggr]\biggr\}. \tag{4-17}$$

The rightmost members of lines (4-16) and (4-17) are equal because the innermost maxima over the quantities in square brackets $[\cdots]$ agree, by Lemma 4.3. We have shown that $\widetilde{H} = \widehat{H}$ and thereby proved the lemma. □

We extend Lemma 4.4 inductively.

Lemma 4.5. *Let $n \ge 2$ and assume given $n+1$ sequences $I^1, I^2, \dots, I^n, \omega^1 \in \mathbb{R}^{\mathbb{Z}}_{\ge 0}$ such that all the queueing operations below are well-defined. Define iteratively*

$$\omega^j = R(I^{j-1}, \omega^{j-1}) \quad \textit{for } j = 2, \dots, n. \tag{4-18}$$

Then we have these identities for $1 \le k \le n-1$:

$$D^{(n+1)}(I^n, I^{n-1}, \dots, I^1, \omega^1) = D^{(k+1)}\bigl(D^{(n-k+1)}[I^n, \dots, I^{k+1}, \omega^{k+1}], D(I^k, \omega^k), \dots, D(I^1, \omega^1)\bigr). \tag{4-19}$$

Proof. The case $n = 2$ is Lemma 4.4.

Let $n \ge 3$ and assume that the claim of the lemma holds when n is replaced by $n-1$, for $1 \le k \le n-2$. We prove the claim for n.

First the case $k = 1$, beginning with the right-hand side of (4-19):

$$\begin{aligned} D^{(2)}\bigl(D^{(n)}(I^n, \dots, I^2, \omega^2), D(I^1, \omega^1)\bigr) &= D\bigl(D\bigl[D^{(n-1)}(I^n, \dots, I^2), \omega^2\bigr], D(I^1, \omega^1)\bigr) \\ &= D\bigl(D\bigl[D^{(n-1)}(I^n, \dots, I^2), I^1\bigr], \omega^1\bigr) \\ &= D\bigl(D^{(n)}(I^n, \dots, I^1), \omega^1\bigr) = D^{(n+1)}(I^n, \dots, I^1, \omega^1). \end{aligned}$$

The first and last two equalities above are from definition (2-26) of $D^{(n)}$, and the middle equality is Lemma 4.4.

Now let $2 \le k \le n-1$. The first equality below is definition (2-26) for $D^{(k+1)}$. The second equality is the induction assumption:

$$\begin{aligned} &D^{(k+1)}\big(D^{(n-k+1)}(I^n, \dots, I^{k+1}, \omega^{k+1}), D(I^k, \omega^k), \dots, D(I^1, \omega^1)\big) \\ &\qquad = D\big(D^{(k)}\big[D^{(n-k+1)}(I^n, \dots, I^{k+1}, \omega^{k+1}), D(I^k, \omega^k), \dots, D(I^2, \omega^2)\big], D(I^1, \omega^1)\big) \\ &\qquad = D\big(D^{(n)}[I^n, \dots, I^2, \omega^2], D(I^1, \omega^1)\big). \end{aligned}$$

The last line above is the same as the left-hand side of the previous display. The calculation is completed as was done there. □

In particular, for $k = n-1$, (4-19) gives

$$D^{(n+1)}(I^n, \dots, I^1, \omega^1) = D^{(n)}\big(D(I^n, \omega^n), \dots, D(I^1, \omega^1)\big) \tag{4-20}$$

and for $k = 1$,

$$D^{(n+1)}(I^n, \dots, I^1, \omega^1) = D\big(D^{(n)}\big[I^n, \dots, I^2, \omega^2\big], D(I^1, \omega^1)\big). \tag{4-21}$$

5. Multiclass processes

The distribution $\mu^{(1,\rho_1,\dots,\rho_n)}$ of the $(n+1)$-tuple $(\overline{Y}_t, \overline{B}_t^{\rho_1,e_1}, \dots, \overline{B}_t^{\rho_n,e_1})$ given in Theorem 3.2 is deduced through studying two multiclass LPP processes. Fix a positive integer n, the number of levels or classes. We define two discrete-time Markov processes on n-tuples of sequences, the *multiline process* and the *coupled process*. Their state space is

$$\mathcal{A}_n = \left\{ I = (I^1, I^2, \dots, I^n) \in (\mathbb{R}_{\ge 0}^{\mathbb{Z}})^n : \forall i \in [n], \lim_{m \to -\infty} \frac{1}{|m|} \sum_{k=-m}^{0} I_k^i > 1 \right\}. \tag{5-1}$$

At each step their evolution is driven by an independent sequence of i.i.d. exponential weights, so assume that

$$\omega = (\omega_k)_{k \in \mathbb{Z}} \text{ is a sequence of i.i.d. variables } \omega_k \sim \text{Exp}(1). \tag{5-2}$$

5A. *Multiline process.* At time $t \in \mathbb{Z}_{\ge 0}$, the state of the *multiline process* is denoted by

$$I(t) = (I^1(t), \dots, I^n(t)) \in \mathcal{A}_n.$$

The one-step evolution from time t to $t+1$ is defined as follows in terms of the mappings (2-22). Given the time t configuration $I(t) = I = (I^1, I^2, \dots, I^n)$ in the space $\mathcal{A}_n$ and independent driving weights ω, define the time $t+1$ configuration $I(t+1) = \overline{I} = (\overline{I}^1, \overline{I}^2, \dots, \overline{I}^n)$ iteratively as follows:

$$\begin{aligned} &\text{set } \omega^1 = \omega \text{ and } \overline{I}^1 = D(I^1, \omega^1); \\ &\text{for } i = 2, 3, \dots, n: \\ &\qquad \text{set } \omega^i = R(I^{i-1}, \omega^{i-1}) \text{ and } \overline{I}^i = D(I^i, \omega^i). \end{aligned} \tag{5-3}$$

Thus the driving sequence ω acts on the first line I^1 directly, and is then transformed at each stage before it is passed to the next line. Lemma A.3 guarantees that, for almost every ω from (5-2), the Cesàro limit $\lim_{m\to-\infty}|m|^{-1}\sum_{k=m}^{0}\omega_k^i = 1$ holds for each $i\in[n]$ and the new state $\bar I$ lies in $\mathcal{A}_n$.

Theorem 5.1. *Assume (5-2). Then for each $\rho=(\rho_1,\dots,\rho_n)\in(1,\infty)^n$, the product measure ν^ρ defined in (3-4) is invariant for the multiline process $(I(t))_{t\in\mathbb{Z}_{\ge 0}}$.*

Theorem 5.1 follows from Lemma B.2 in Appendix B: induction on k shows that $\bar I^1,\dots,\bar I^k,\ \omega^{k+1}, I^{k+1},\dots,I^n$ are independent with $\bar I^i\sim\nu^{\rho_i},\ \omega^{k+1}\sim\nu^1,\ I^j\sim\nu^{\rho_j}$. We do not have proof that ν^ρ is the unique translation-ergodic stationary distribution with mean vector ρ, but have no reason to doubt this either.

5B. *Coupled process.* At time $t\in\mathbb{Z}_{\ge 0}$, the state of the *coupled process* is denoted by

$$\eta(t)=(\eta^1(t),\dots,\eta^n(t))\in\mathcal{A}_n$$

where again $\eta^i(t)=(\eta^i_k(t))_{k\in\mathbb{Z}}$. The evolution is simple: the queueing operator D acts on each sequence η^i with service times ω:

$$\eta(t+1)=\big(D(\eta^1(t),\omega),\,D(\eta^2(t),\omega),\,\dots,\,D(\eta^n(t),\omega)\big). \tag{5-4}$$

We call $\eta(t)$ the coupled process because it lives also on the smaller state space $\mathcal{X}_n\cap\mathcal{A}_n$ (recall (3-2)) where the sequences η^i are coupled so that $\eta^{i-1}\le\eta^i$. This is the case relevant for the Busemann processes because the latter are monotone (recall (2-6)). Inequality (2-28) and Lemma A.1 ensure that the Markovian evolution $\eta(\cdot)$ is well-defined on $\mathcal{X}_n\cap\mathcal{A}_n$. However, since the mapping (5-4) is well-defined for more general states, we consider it on the larger state space $\mathcal{A}_n$ of (5-1).

To state an invariance and uniqueness theorem for all parameter vectors $\rho\in(1,\infty)^n$ we extend μ^ρ of (3-5), by ordering ρ and by requiring that $\eta^i=\eta^{i+1}$ if $\rho_i=\rho_{i+1}$. This is necessary because the mapping $\mathcal{D}^{(n)}$ in (3-5) cannot be applied if some $\rho_i=\rho_{i+1}$. For if I and ω are both i.i.d. $\text{Exp}(\rho^{-1})$ sequences, then $\widetilde G$ in (2-18) is identically infinite because it equals a random constant plus the supremum of a symmetric random walk.

Definition 5.2. Let $\rho=(\rho_1,\rho_2,\dots,\rho_n)\in(0,\infty)^n$. The probability measure μ^ρ on the space $(\mathbb{R}_{\ge 0}^{\mathbb{Z}})^n$ is defined as follows.

(i) If $0<\rho_1<\rho_2<\dots<\rho_n$ then apply (3-5).

(ii) If $0<\rho_1\le\rho_2\le\dots\le\rho_n$, there exist $m\in[n]$, a vector $\sigma=(\sigma_1,\dots,\sigma_m)$ such that $0<\sigma_1<\dots<\sigma_m$, and indices $1=i_1<i_2<\dots<i_m<i_{m+1}=n+1$ such that $\rho_{i_\ell}=\dots=\rho_{i_{\ell+1}-1}=\sigma_\ell$ for $\ell=1,\dots,m$. Let $I\sim\nu^\sigma,\ \zeta=\mathcal{D}^{(m)}(I)$, and then define $\eta=(\eta^1,\dots,\eta^n)\in\mathcal{X}_n$ by $\eta^{i_\ell}=\dots=\eta^{i_{\ell+1}-1}=\zeta^\ell$ for $\ell=1,\dots,m$. Define μ^ρ to be the distribution of η.

(iii) For general $\rho=(\rho_1,\dots,\rho_n)\in(0,\infty)^n$, choose a permutation π such that $\pi\rho=(\rho_{\pi(1)},\dots,\rho_{\pi(n)})$ satisfies $\rho_{\pi(1)}\le\rho_{\pi(2)}\le\dots\le\rho_{\pi(n)}$. Let π act on weight configurations $\eta=(\eta^1,\dots,\eta^n)$ by $\pi\eta=(\eta^{\pi(1)},\dots,\eta^{\pi(n)})$. Define $\mu^\rho=\mu^{\pi\rho}\circ\pi^{-1}$, or more explicitly

$$E^{\mu^\rho}[f]=E^{\mu^{\pi\rho}}[f(\pi\eta)]=E^{\mu^{\pi\rho}\circ\pi^{-1}}[f]$$

for bounded Borel functions f on $(\mathbb{R}_{\ge 0}^{\mathbb{Z}})^n$, where the measure $\mu^{\pi\rho}$ is the one defined in step (ii).

If there is more than one ordering permutation in step (iii), there are identical sequences whose ordering among themselves is immaterial. If $\rho \in (1,\infty)^n$ then μ^ρ is supported on the space $\mathcal{A}_n$ of (5-1). The next existence and uniqueness theorem is proved in Section 5D.

Theorem 5.3. *Assume* (5-2).

(i) Invariance. *Let* $\rho = (\rho_1, \rho_2, \dots, \rho_n) \in (1,\infty)^n$. *Then the probability measure* μ^ρ *of* Definition 5.2 *is invariant for the Markov chain* $(\eta(t))_{t\in\mathbb{Z}_{\geq 0}}$ *defined by* (5-4).

(ii) Uniqueness. *Let* $\widetilde{\mu}$ *be a translation-ergodic probability measure on* $\mathcal{A}_n$ *under which coordinates* η^i_k *have finite means* $\rho_i = E^{\widetilde{\mu}}[\eta^i_k] > 1$. *If* $\widetilde{\mu}$ *is invariant for the process* $\eta(t)$, *then* $\widetilde{\mu} = \mu^\rho$ *for* $\rho = (\rho_1, \dots, \rho_n) \in (1,\infty)^n$.

5C. *Stationary multiclass LPP on the upper half-plane.* We reformulate the coupled process as a multiclass CGM on the upper half-plane. Fix the number n of classes. Assume given i.i.d. Exp(1) random weights $\{\omega_x\}_{x\in\mathbb{Z}\times\mathbb{Z}_{>0}}$, and an initial configuration $\eta(0) = (\eta^1(0), \dots, \eta^n(0)) \in \mathcal{A}_n$ independent of ω. Define a vector of LPP processes $G_x = (G^1_x, \dots, G^n_x)$ for $x \in \mathbb{Z}\times\mathbb{Z}_{\geq 0}$ as follows. First choose initial functions $\{G^i_{(k,0)}\}_{k\in\mathbb{Z}}$ with the property $\eta^i_k(0) = G^i_{(k,0)} - G^i_{(k-1,0)}$. Then for $(k,t) \in \mathbb{Z}\times\mathbb{Z}_{>0}$ define

$$G^i_{(k,t)} = \sup_{j:\, j\leq k} \{G^i_{(j,0)} + G_{(j,1),(k,t)}\}, \tag{5-5}$$

where $G_{x,y}$ is the usual LPP process of (2-1) with weights $Y_x(\omega) = \omega_x$. Then lastly define the process $\eta(t) = (\eta^1(t), \dots, \eta^n(t))$ for $t \in \mathbb{Z}_{>0}$ as the increments

$$\eta^i_k(t) = G^i_{(k,t)} - G^i_{(k-1,t)} \qquad \text{for } i \in [n] \text{ and } k \in \mathbb{Z}. \tag{5-6}$$

Theorem 5.4. *Let* $\rho = (\rho_1, \rho_2, \dots, \rho_n) \in (1,\infty)^n$. *Then* μ^ρ *of* Definition 5.2 *is an invariant measure of the increment process* $\eta(\cdot)$ *defined above by* (5-6) *in the multiclass exponential corner growth model. Measure* μ^ρ *is the unique invariant measure for* $\eta(\cdot)$ *among translation-ergodic probability measures on* $\mathcal{A}_n$ *with means given by* ρ.

This follows from Theorem 5.3 simply by noting that (5-6) can be reformulated inductively as

$$\eta(t) = \bigl(D(\eta^1(t-1), \overline{\omega}_t), D(\eta^2(t-1), \overline{\omega}_t), \dots, D(\eta^n(t-1), \overline{\omega}_t)\bigr), \qquad t \in \mathbb{Z}_{>0}, \tag{5-7}$$

where $\overline{\omega}_t = \{\omega_{(k,t)}\}_{k\in\mathbb{Z}}$ is the sequence of weights on level t.

5D. *Invariant distribution for the coupled process.* This section proves Theorem 5.3. We separate the invariance of μ^ρ and the uniqueness in Theorems 5.5 and 5.6 below. Their combination establishes Theorem 5.3. The proof of the next theorem shows how the invariance of μ^ρ for $\eta(t)$ follows from the invariance of ν^ρ for $I(t)$ and the fact that the mapping $\mathcal{D}^{(n)}$ intertwines the evolutions of $I(t)$ and $\eta(t)$.

Theorem 5.5. *Let* $\rho = (\rho_1, \rho_2, \dots, \rho_n) \in (1,\infty)^n$. *Then* μ^ρ *of* Definition 5.2 *is an invariant distribution for the* $(\mathbb{R}^{\mathbb{Z}}_{\geq 0})^n$*-valued Markov chain* $\eta(t)$ *defined by* (5-4).

Proof. The general claim follows from the case $1 < \rho_1 < \rho_2 < \cdots < \rho_n$ because permuting the $\{\eta^i\}$ or setting $\eta^i = \eta^j$ produces the exact same change in the image of the mapping in (5-4).

So assume $1 < \rho_1 < \rho_2 < \cdots < \rho_n$. Given a driving sequence ω, denote by $\mathcal{S}^\omega$ and $\mathcal{T}^\omega$ the mappings on the state spaces that encode a single temporal evolution step of the processes $I(\cdot)$ and $\eta(\cdot)$. In other words, the mapping from time t to $t+1$ defined by (5-3) for the multiline process is encoded as $I(t+1) = \mathcal{S}^\omega(I(t))$. For the coupled process the step in (5-4) is encoded as $\eta(t+1) = \mathcal{T}^\omega(\eta(t))$. Let $\mathcal{D} = \mathcal{D}^{(n)}$ denote the mapping (3-3) that constructs the coupled configuration from the multiline configuration. Let $\mathcal{D}_k$, $\mathcal{S}^\omega_k$ and $\mathcal{T}^\omega_k$ denote the k-th $\mathbb{R}^{\mathbb{Z}}_{\geq 0}$-valued coordinates of the images of these mappings.

Let $I \sim \nu^\rho$ be a multiline configuration with product exponential distribution ν^ρ. We need to show that if η has the distribution μ^ρ of $\mathcal{D}(I)$, then so does $\mathcal{T}^\omega(\eta)$ when ω is an independent sequence of i.i.d. Exp(1) weights. For the argument we can assume that $\eta = \mathcal{D}(I)$. As before let $\omega^1 = \omega$ and iteratively $\omega^j = R(I^{j-1}, \omega^{j-1})$ for $j = 2, 3, \ldots, n$. The fourth equality below is (4-20). The other equalities are consequences of definitions:

$$\begin{aligned}
\mathcal{T}^\omega_k(\eta) &= D(\eta^k, \omega) = D\big(D^{(k)}(I^k, \ldots, I^1), \omega^1\big) = D^{(k+1)}(I^k, \ldots, I^1, \omega^1) \\
&\overset{(4\text{-}20)}{=} D^{(k)}\big(D(I^k, \omega^k), D(I^{k-1}, \omega^{k-1}), \ldots, D(I^1, \omega^1)\big) \\
&= D^{(k)}\big(\mathcal{S}^\omega_k(I), \mathcal{S}^\omega_{k-1}(I), \ldots, \mathcal{S}^\omega_1(I)\big) = \mathcal{D}_k(\mathcal{S}^\omega(I)).
\end{aligned}$$

Since the above works for all coordinates $k \in [n]$, we have $\mathcal{T}^\omega(\eta) = \mathcal{D}(\mathcal{S}^\omega(I))$. Since $\eta = \mathcal{D}(I)$, we have verified the intertwining

$$\mathcal{T}^\omega(\mathcal{D}(I)) = \mathcal{D}(\mathcal{S}^\omega(I)). \tag{5-8}$$

By Theorem 5.1, $\mathcal{S}^\omega(I) \overset{d}{=} I \sim \nu^\rho$. Consequently $\mathcal{T}^\omega(\eta) \overset{d}{=} \mathcal{D}(I) \sim \mu^\rho$. □

Theorem 5.6. *Assume (5-2). Let $\widetilde{\mu}$ be a translation-ergodic probability measure on $\mathcal{X}_n$ under which each coordinate η^i_k has a finite mean. If $\widetilde{\mu}$ is invariant for the coupled process $\eta(t)$, then $\widetilde{\mu} = \mu^\rho$ for the mean vector ρ of $\widetilde{\mu}$.*

We prove Theorem 5.6 following [Chang 1994], by showing that the evolution contracts the $\bar{\rho}$ distance between stationary and ergodic sequences. Let $\eta = (\eta_k)_{k\in\mathbb{Z}}$ and $\xi = (\xi_k)_{k\in\mathbb{Z}}$ be stationary processes taking values in $\mathbb{R}^n_{\geq 0}$. Their $\bar{\rho}$ distance is defined by

$$\bar{\rho}(\eta, \xi) = \inf_{(X,Y)\in\mathcal{M}} \boldsymbol{E}[\,|X_0 - Y_0|_1], \tag{5-9}$$

where $\mathcal{M}$ is the set of jointly defined stationary sequences $(X, Y) = (X_k, Y_k)_{k\in\mathbb{Z}}$ such that $X \overset{d}{=} \eta$ and $Y \overset{d}{=} \xi$, $\boldsymbol{E}$ is the expectation on the probability space on which the coupling (X, Y) is defined, and $|\cdot|_1$ is the ℓ^1 distance on $\mathbb{R}^n_{\geq 0}$.

From [Gray 2009, Theorem 9.2], we know that (i) $\bar{\rho}$ induces a metric on the space of translation-invariant distributions and (ii) if η and ξ are both ergodic, there exists a jointly stationary and ergodic pair (X, Y) at which the infimum in (5-9) is attained.

The following is a straightforward generalization of Theorem 2.4 of [Chang 1994] to $\mathbb{R}^n_{\geq 0}$-valued stationary and ergodic sequences $\eta = (\eta^1, \ldots, \eta^n)$ and $\xi = (\xi^1, \ldots, \xi^n)$ where $\eta^i = (\eta^i_k)_{k\in\mathbb{Z}}$ and $\xi^i = (\xi^i_k)_{k\in\mathbb{Z}}$

are random elements of $\mathbb{R}^{\mathbb{Z}}_{\geq 0}$. Let

$$\widetilde{\eta} = (\widetilde{\eta}^1, \dots, \widetilde{\eta}^n) = \bigl(D(\eta^1, \omega), \dots, D(\eta^n, \omega)\bigr)$$

and similarly $\widetilde{\xi} = (\widetilde{\xi}^1, \dots, \widetilde{\xi}^n)$ denote the outcome of applying the queueing map $D(\cdot, \omega)$ to each sequence-valued coordinate.

Proposition 5.7. *Let ω satisfy (5-2). Let the $\mathbb{R}^n_{\geq 0}$-valued stationary and ergodic processes η and ξ be independent of ω and have finite means that satisfy $\mathbb{E}[\eta^i_k] = \mathbb{E}[\xi^i_k] = \lambda_i > 1$ for $i \in [n]$ and $k \in \mathbb{Z}$. Then*

$$\bar{\rho}(\widetilde{\eta}, \widetilde{\xi}) \leq \bar{\rho}(\eta, \xi). \tag{5-10}$$

If η and ξ have different distributions the inequality in (5-10) is strict.

Before the proof we complete the proof of Theorem 5.6. Let $\rho = E^{\widetilde{\mu}}[\eta_0]$ be the mean vector of $\widetilde{\mu}$. Let $\eta \sim \mu^\rho$ and $\xi \sim \widetilde{\mu}$. By the known invariance of μ^ρ and the assumed invariance of $\widetilde{\mu}$, $\widetilde{\eta} \overset{d}{=} \eta$ and $\widetilde{\xi} \overset{d}{=} \xi$. Hence $\bar{\rho}(\widetilde{\eta}, \widetilde{\xi}) = \bar{\rho}(\eta, \xi)$. The last statement of Proposition 5.7 forces $\widetilde{\mu} = \mu^\rho$.

Proof of Proposition 5.7. Let $(X, Y) = ((X^1, \dots, X^n), (Y^1, \dots, Y^n))$ be an arbitrary $\mathbb{R}^{2n}_{\geq 0}$-valued jointly stationary and ergodic process with marginals $X \overset{d}{=} \eta$ and $Y \overset{d}{=} \xi$, independent of the weights ω, with (X, Y, ω) coupled together under a probability measure $\boldsymbol{P}$ with expectation $\boldsymbol{E}$. As above, write $\widetilde{X}^i = (\widetilde{X}^i_k)_{k \in \mathbb{Z}} = D(X^i, \omega)$ and $\widetilde{Y}^i = (\widetilde{Y}^i_k)_{k \in \mathbb{Z}} = D(Y^i, \omega)$ for the action of the queueing operator on the individual sequences $X^i = (X^i_k)_{k \in \mathbb{Z}}$ and $Y^i = (Y^i_k)_{k \in \mathbb{Z}}$. Inequality (5-10) follows from showing

$$\boldsymbol{E}[|\widetilde{X}_0 - \widetilde{Y}_0|_1] \leq \boldsymbol{E}[|X_0 - Y_0|_1]. \tag{5-11}$$

Define the process Z by $Z^i_k = X^i_k \vee Y^i_k$. Then

$$|X_0 - Y_0|_1 = \sum_{i=1}^n |X^i_0 - Y^i_0| = \sum_{i=1}^n (2Z^i_0 - X^i_0 - Y^i_0). \tag{5-12}$$

Let $\widetilde{Z}^i = D(Z^i, \omega)$. Then $\widetilde{Z}^i \geq \widetilde{X}^i \vee \widetilde{Y}^i$ by monotonicity (2-28). Hence

$$|\widetilde{X}_0 - \widetilde{Y}_0|_1 = \sum_{i=1}^n |\widetilde{X}^i_0 - \widetilde{Y}^i_0| = \sum_{i=1}^n \bigl(2(\widetilde{X}^i_0 \vee \widetilde{Y}^i_0) - \widetilde{X}^i_0 - \widetilde{Y}^i_0\bigr) \leq \sum_{i=1}^n (2\widetilde{Z}^i_0 - \widetilde{X}^i_0 - \widetilde{Y}^i_0). \tag{5-13}$$

The triple (X, Y, ω) is jointly stationary and ergodic because ω is an i.i.d. process independent of the ergodic process (X, Y). Consequently, as translation-respecting mappings of ergodic processes, both (X, Y, Z, ω) and $(\widetilde{X}, \widetilde{Y}, \widetilde{Z})$ are jointly stationary and ergodic. The queueing stability condition $\boldsymbol{E}(X^i_0) > \boldsymbol{E}(\omega_0)$ implies $\boldsymbol{E}(\widetilde{X}^i_0) = \boldsymbol{E}(X^i_0)$, and by the same token $\boldsymbol{E}(\widetilde{Y}^i_0) = \boldsymbol{E}(Y^i_0)$ and $\boldsymbol{E}(\widetilde{Z}^i_0) = \boldsymbol{E}(Z^i_0)$. This goes back to Loynes [1962] and follows also from Lemma A.3 in Appendix A. Taking expectations on both sides of (5-12) and (5-13) gives (5-11).

For the strict inequality assume that η and ζ are not equal in distribution and let (X, Y) be a jointly ergodic pair that gives the minimum in (5-9). To deduce the strict inequality

$$\sum_{i=1}^n \boldsymbol{E}[\widetilde{X}^i_0 \vee \widetilde{Y}^i_0] < \sum_{i=1}^n \boldsymbol{E}(\widetilde{Z}^i_0). \tag{5-14}$$

we can tap directly into the proof of part (ii) of Theorem 2.4 in [Chang 1994], once we show that X^i and Y^i must cross for some $i \in [n]$. X^i and Y^i *cross* if with probability one there exist $k, \ell \in \mathbb{Z}$ such that $X^i_k > Y^i_k$ and $X^i_\ell < Y^i_\ell$.

Suppose X^i and Y^i do not cross. Then $\boldsymbol{P}(\{X^i \geq Y^i\} \cup \{X^i \leq Y^i\}) = 1$. We show that this implies $X^i = Y^i$ a.s. This gives us the contradiction needed, since $X^i = Y^i$ for all $i \in [n]$ implies that $\eta \overset{d}{=} \xi$.

To show $X^i = Y^i$ a.s., write $\{X^i = Y^i\}^c = A^+ \cup A^-$ a.s. for

$$\begin{aligned} A^+ &= \{X^i \geq Y^i \text{ and } X^i_k > Y^i_k \text{ for some } k \in \mathbb{Z}\},\\ A^- &= \{X^i \leq Y^i \text{ and } X^i_k < Y^i_k \text{ for some } k \in \mathbb{Z}\}. \end{aligned}$$

A^+ is a shift-invariant event. By the joint ergodicity of (X, Y) and $\boldsymbol{E}(X^i_0 - Y^i_0) = 0$,

$$\begin{aligned} 0 &= \mathbf{1}_{A^+} \cdot \lim_{n\to\infty} \frac{1}{2n+1} \sum_{-n\leq k\leq n} (X^i_k - Y^i_k)\\ &= \lim_{n\to\infty} \frac{1}{2n+1} \sum_{-n\leq k\leq n} (X^i_k - Y^i_k) \cdot \mathbf{1}_{\theta_{-k}A^+} = \boldsymbol{E}[(X^i_0 - Y^i_0) \cdot \mathbf{1}_{A^+}] \quad \text{a.s.} \end{aligned}$$

Thus $X^i_0 = Y^i_0$ a.s. on A^+. By the shift-invariance of A^+, $X^i_k = Y^i_k$ a.s. on A^+ for all $k \in \mathbb{Z}$. But then it must be that $\boldsymbol{P}(A^+) = 0$. Similarly $\boldsymbol{P}(A^-) = 0$.

To summarize, we have shown that some X^i and Y^i must cross. Following the proof on page 1131–1132 of [Chang 1994] gives the strict inequality (5-14). The connection between the notation of [Chang 1994] and ours is $S_k = \omega_k$, $(T^1_{1,k-1}, T^1_{2,k-1}) = (X^i_k, \widetilde{X}^i_k)$ and $(T^2_{1,k-1}, T^2_{2,k-1}) = (Y^i_k, \widetilde{Y}^i_k)$. □

6. Proofs of the results for Busemann functions

We prove the theorems of Section 3 in the order in which they were stated.

6A. *Continuity of μ^ρ and distribution of the Busemann process.*

Proof of the continuity claim of Theorem 3.1. Fix $\rho = (\rho_1, \dots, \rho_n)$ such that $0 < \rho_1 < \cdots < \rho_n$. Let $\{\rho^h\}_{h\in\mathbb{Z}_{>0}}$ be a sequence of parameter vectors such that $\rho^h = (\rho^h_1, \dots, \rho^h_n) \to (\rho_1, \dots, \rho_n)$ as $h \to \infty$. We construct variables $\eta^h \sim \mu^{\rho^h}$ and $\eta \sim \mu^\rho$ such that $\eta^h \to \eta$ coordinatewise almost surely.

Let $I = (I^1, \dots, I^n) \sim \nu^\rho$ and define $I^{h,i}_k = (\rho^h_i/\rho_i) I^i_k$. Then $I^h = (I^{h,1}, \dots, I^{h,n}) \sim \nu^{\rho^h}$ and we have the pointwise limits $I^{h,i}_k \to I^i_k$ for all $i \in [n]$ and $k \in \mathbb{Z}$ as $h \to \infty$. Furthermore, the assumption in (A-2) holds:

$$\varlimsup_{\substack{m\to-\infty\\ h\to\infty}} \left| \frac{1}{|m|} \sum_{j=m}^{0} I^{h,i}_j - \rho_i \right| = 0 \quad \text{almost surely for all } i \in [n]. \tag{6-1}$$

Let $\eta^h = \mathcal{D}^{(n)}(I^h)$ and $\eta = \mathcal{D}^{(n)}(I)$. Apply Lemma A.2 repeatedly to show that $\eta^h \to \eta$ coordinatewise almost surely:

(1) $\eta^{h,1} = I^{h,1} \to I^1 = \eta^1$ needs no proof.

(2) Lemma A.2 gives the limit $\eta^{h,2} = D(I^{h,2}, I^{h,1}) \to D(I^2, I^1) = \eta^2$ and that $D(I^{h,2}, I^{h,1})$ satisfies the hypotheses of the lemma.

(3) For $\eta^{h,3} = D^{(3)}(I^{h,3}, I^{h,2}, I^{h,1}) = D(D(I^{h,3}, I^{h,2}), I^{h,1})$, by case (2), $D(I^{h,3}, I^{h,2})$ satisfies the hypotheses of Lemma A.2. Then Lemma A.2 gives $D(D(I^{h,3}, I^{h,2}), I^{h,1}) \to D(D(I^3, I^2), I^1)$ and that $D^{(3)}(I^{h,3}, I^{h,2}, I^{h,1})$ satisfies the hypotheses of Lemma A.2.

(4) Proceed by induction. From the case of $i-1$ sequences, $D^{(i-1)}(I^{h,i}, I^{h,i-1}, \dots, I^{h,2})$ satisfies the hypotheses of Lemma A.2. Apply the Lemma to conclude that the mapping for i sequences obeys the limit

$$\eta^{h,i} = D^{(i)}(I^{h,i}, \dots, I^{h,2}, I^{h,1}) = D(D^{(i-1)}(I^{h,i}, I^{h,i-1}, \dots, I^{h,2}), I^{h,1}) \to \eta^i$$

and also satisfies the assumptions of Lemma A.2. This is then passed on to be used for the case of $i+1$ sequences.

This completes the proof of $\eta^h \to \eta$. □

Proof of Theorem 3.2. Introduce an $(n+1)$-st parameter value $\rho_0 \in (1, \rho_1)$. By Lemma 2.3, the $\mathbb{R}^{n+1}_{\geq 0}$-valued $\mathbb{Z}$-indexed process

$$\bar{B}_t^{\rho_0,\dots,\rho_n,\boldsymbol{e}_1} = \{(B^{\rho_0}_{(k-1,t),(k,t)}, B^{\rho_1}_{(k-1,t),(k,t)}, \dots, B^{\rho_n}_{(k-1,t),(k,t)})\}_{k\in\mathbb{Z}} \tag{6-2}$$

is stationary and ergodic under translation of the k-index and furthermore $\bar{B}_t^{\rho_0,\dots,\rho_n,\boldsymbol{e}_1}$ has the same distribution as the sequence $\bar{B}_{t-1}^{\rho_0,\dots,\rho_n,\boldsymbol{e}_1}$ on the previous level $t-1$. Lemma 3.3 gives $\bar{B}_t^{\rho_0,\dots,\rho_n,\boldsymbol{e}_1} = D(\bar{B}_{t-1}^{\rho_0,\dots,\rho_n,\boldsymbol{e}_1}, \bar{Y}_t)$. By the uniqueness given in Theorem 5.3, the distribution of $\bar{B}_t^{\rho_0,\dots,\rho_n,\boldsymbol{e}_1}$ must be the invariant distribution $\mu^{(\rho_0,\dots,\rho_n)}$.

Let $\rho_0 \searrow 1$. By Lemma 2.2, almost surely,

$$\lim_{\rho_0\searrow 1} \bar{B}_t^{\rho_0,\rho_1,\dots,\rho_n,\boldsymbol{e}_1} = \{(Y_{(k,t)}, B^{\rho_1}_{(k-1,t),(k,t)}, \dots, B^{\rho_n}_{(k-1,t),(k,t)})\}_{k\in\mathbb{Z}},$$

while Theorem 3.1 gives the weak convergence $\mu^{(\rho_0,\rho_1,\dots,\rho_n)} \to \mu^{(1,\rho_1,\dots,\rho_n)}$ as $\rho_0 \searrow 1$. □

The proof of Lemma 3.3 below relies on the iterative equations (2-24). Since these equations can have solutions other than the one coming from the queuing mapping, additional conditions are needed as specified in Lemma A.4 in Appendix A.

Proof of Lemma 3.3. We show that there is an event Ω_0 of full probability on which the assumptions of Lemma A.4 hold for the sequences $(\tilde{I}, J, I, \omega) = (\bar{B}_t^{\rho,\boldsymbol{e}_1}, \bar{B}_t^{\rho,\boldsymbol{e}_2}, \bar{B}_{t-1}^{\rho,\boldsymbol{e}_1}, \bar{Y}_t)$ simultaneously for *all* uncountably many $\rho \in (1,\infty)$ and $t \in \mathbb{Z}$.

Assumption (A-14) requires

$$\lim_{m\to-\infty} \sum_{k=m}^{0} (Y_{(k,t)} - B^{\rho}_{(k,t-1),(k+1,t-1)}) = -\infty \quad \text{for all } t \in \mathbb{Z}.$$

This holds almost surely simultaneously for all ρ in a dense countable subset of $(1,\infty)$. By the monotonicity (2-6) this extends to all $\rho \in (1,\infty)$ on a single event of full probability.

Utilizing the recovery property (2-8) and additivity (2-7),

$$\begin{aligned} Y_{(k,t)} &+ (B^{\rho}_{(k-1,t-1),(k,t-1)} - B^{\rho}_{(k-1,t-1),(k-1,t)})^{+} \\ &= B^{\rho}_{(k-1,t),(k,t)} \wedge B^{\rho}_{(k,t-1),(k,t)} + (B^{\rho}_{(k-1,t),(k,t)} - B^{\rho}_{(k,t-1),(k,t)})^{+} \\ &= B^{\rho}_{(k-1,t),(k,t)} \end{aligned}$$

and

$$\begin{aligned} Y_{(k,t)} &+ (B^{\rho}_{(k-1,t-1),(k-1,t)} - B^{\rho}_{(k-1,t-1),(k,t-1)})^{+} \\ &= B^{\rho}_{(k-1,t),(k,t)} \wedge B^{\rho}_{(k,t-1),(k,t)} + (B^{\rho}_{(k,t-1),(k,t)} - B^{\rho}_{(k-1,t),(k,t)})^{+} \\ &= B^{\rho}_{(k,t-1),(k,t)}. \end{aligned}$$

These equations are valid for all ρ and all (k,t) on a single event of full probability because this is true of properties (2-8) and (2-7). Assumption (A-15) has been verified.

Lemma A.5 implies that with probability one, for all ρ in a dense countable subset of $(1,\infty)$, $Y_{(k,t)} = B^{\rho}_{(k,t-1),(k,t)}$ for infinitely many $k < 0$. Monotonicity (2-6) and recovery (2-8) extend this property to all $\rho \in (1,\infty)$ on the same event. □

6B. *Triangular arrays and independent increments.* To extract further properties of the distribution μ^{ρ}, we develop an alternative representation for $\eta = \mathcal{D}^{(n)}(I)$ of (3-3). Assume given $I = (I^1, \dots, I^n) \in \mathcal{Y}_n$. Define arrays $\{\eta^{i,j} : 1 \le j \le i \le n\}$ and $\{\xi^{i,j} : 1 \le j \le i \le n\}$ of elements of $\mathbb{R}^{\mathbb{Z}}_{\ge 0}$ as follows. The ξ variables are passed from one i level to the next.

(i) For $i = 1$, set $\eta^{1,1} = I^1 = \xi^{1,1}$.

(ii) For $i = 2, 3, \dots, n$,

$$\begin{aligned} \eta^{i,1} &= I^i, \\ \eta^{i,j} &= D(\eta^{i,j-1}, \xi^{i-1,j-1}) \qquad \text{for } j = 2,3,\dots,i, \\ \xi^{i,j-1} &= R(\eta^{i,j-1}, \xi^{i-1,j-1}) \qquad \text{for } j = 2,3,\dots,i, \\ \xi^{i,i} &= \eta^{i,i}. \end{aligned} \tag{6-3}$$

Step i takes inputs from two sources: from the outside it takes I^i, and from step $i-1$ it takes the configuration $\xi^{i-1,\bullet} = (\xi^{i-1,1}, \xi^{i-1,2}, \dots, \xi^{i-1,i-2}, \xi^{i-1,i-1} = \eta^{i-1,i-1})$.

Lemma A.3 ensures that the arrays are well-defined for $I \in \mathcal{Y}_n$. The inputs $I^1, \dots, I^n$ enter the algorithm one by one in order. If the process is stopped after the step $i = m$ is completed for some $m < n$, it produces the arrays for $(I^1, \dots, I^m) \in \mathcal{Y}_m$.

The arrays are illustrated in Figure 4. The following properties of the arrays come from Lemmas 6.1 and 6.2 and their proofs.

(i) The input of the $\mathcal{D}^{(n)}$-mapping lies on the left edge of the η-array: $(\eta^{1,1}, \dots, \eta^{n,1}) = (I^1, \dots, I^n)$. The output of the $\mathcal{D}^{(n)}$-mapping lies on the right-hand diagonal edges of both arrays:

$$(\eta^{1,1}, \eta^{2,2}, \dots, \eta^{n,n}) = (\xi^{1,1}, \xi^{2,2}, \dots, \xi^{n,n}) = \mathcal{D}^{(n)}(I^1, \dots, I^n) \sim \mu^{(\rho_1, \rho_2, \dots, \rho_n)}.$$

$$\begin{array}{lllll}
\eta^{1,1} & & & & \\
\eta^{2,1} & \eta^{2,2} & & & \\
\eta^{3,1} & \eta^{3,2} & \eta^{3,3} & & \\
\vdots & \vdots & \vdots & \ddots & \\
\eta^{n,1} & \eta^{n,2} & \eta^{n,3} & \cdots & \eta^{n,n}
\end{array}
\qquad\qquad
\begin{array}{lllll}
\xi^{1,1} & & & & \\
\xi^{2,1} & \xi^{2,2} & & & \\
\xi^{3,1} & \xi^{3,2} & \xi^{3,3} & & \\
\vdots & \vdots & \vdots & \ddots & \\
\xi^{n,1} & \xi^{n,2} & \xi^{n,3} & \cdots & \xi^{n,n}
\end{array}$$

Figure 4. Arrays $\{\eta^{i,j} : 1 \le j \le i \le n\}$ and $\{\xi^{i,j} : 1 \le j \le i \le n\}$.

(ii) The j-th column $(\eta^{j,j}, \eta^{j+1,j}, \dots, \eta^{n,j})$ of the η-array has the product distribution $\nu^{(\rho_j,\,\rho_{j+1},\dots,\,\rho_n)}$. It is obtained from the $(j-1)$-st column $(\eta^{j,j-1}, \eta^{j+1,j-1}, \dots, \eta^{n,j-1})$ by the mapping (5-3) with $\eta^{j-1,j-1} = \xi^{j-1,j-1}$ as the external driving weights.

(iii) Row $(\xi^{i,1}, \xi^{i,2}, \dots, \xi^{i,i})$ of the ξ-array has the product distribution $\nu^{(\rho_1,\,\rho_2,\dots,\,\rho_i)}$.

Lemma 6.1. *Let $I = (I^1, \dots, I^n) \in \mathcal{Y}_n$. Let $(\widetilde{\eta}^1, \dots, \widetilde{\eta}^n) = \mathcal{D}^{(n)}(I^1, \dots, I^n)$ be given by the mapping (3-3). Let $\{\eta^{i,j}\}$ be the array defined above. Then $\widetilde{\eta}^i = \eta^{i,i}$ for $i = 1, \dots, n$.*

Proof. It suffices to prove $\tilde{\eta}^n = \eta^{n,n}$ because the same proof applies to all i. The construction of the array can be reimagined as follows. Start with $(\eta^{1,1}, \eta^{2,1}, \dots, \eta^{n,1}) = (I^1, I^2, \dots, I^n)$. Then for $\ell = 2, 3, \dots, n-1$ iterate the following step that maps the $(n-\ell+2)$-vector

$$(\eta^{n,\ell-1}, \eta^{n-1,\ell-1}, \dots, \eta^{\ell,\ell-1}, \eta^{\ell-1,\ell-1})$$

to the $(n-\ell+1)$-vector

$$\begin{aligned}
&(\eta^{n,\ell}, \eta^{n-1,\ell}, \dots, \eta^{\ell+1,\ell}, \eta^{\ell,\ell}) \\
&\qquad = \bigl(D(\eta^{n,\ell-1}, \xi^{n-1,\ell-1}), D(\eta^{n-1,\ell-1}, \xi^{n-2,\ell-1}), \dots, D(\eta^{\ell+1,\ell-1}, \xi^{\ell,\ell-1}), D(\eta^{\ell,\ell-1}, \eta^{\ell-1,\ell-1})\bigr).
\end{aligned}$$

The ξ-variables above satisfy

$$\begin{aligned}
\xi^{\ell,\ell-1} &= R(\eta^{\ell,\ell-1}, \xi^{\ell-1,\ell-1}) = R(\eta^{\ell,\ell-1}, \eta^{\ell-1,\ell-1}) \\
\xi^{\ell+1,\ell-1} &= R(\eta^{\ell+1,\ell-1}, \xi^{\ell,\ell-1}) \\
&\ \ \vdots \\
\xi^{n-1,\ell-1} &= R(\eta^{n-1,\ell-1}, \xi^{n-2,\ell-1}).
\end{aligned}$$

Thus (4-20) implies that

$$D^{(n-\ell+2)}(\eta^{n,\ell-1}, \eta^{n-1,\ell-1}, \dots, \eta^{\ell,\ell-1}, \eta^{\ell-1,\ell-1}) = D^{(n-\ell+1)}(\eta^{n,\ell}, \eta^{n-1,\ell}, \dots, \eta^{\ell+1,\ell}, \eta^{\ell,\ell}). \tag{6-4}$$

In the derivation below, use the first line of (6-3) to replace each I^i with $\eta^{i,1}$. Then iterate (6-4) from $\ell = 2$ to $\ell = n-1$ to obtain

$$\begin{aligned}
\widetilde{\eta}^n &= D^{(n)}(I^n, I^{n-1}, \dots, I^3, I^2, I^1) = D^{(n)}(\eta^{n,1}, \eta^{n-1,1}, \dots, \eta^{3,1}, \eta^{2,1}, \eta^{1,1}) \\
&= D^{(n-1)}\bigl(\eta^{n,2}, \eta^{n-1,2}, \dots, \eta^{3,2}, \eta^{2,2}\bigr) \\
&= \cdots = D^{(3)}(\eta^{n,n-2}, \eta^{n-1,n-2}, \eta^{n-2,n-2}) = D(\eta^{n,n-1}, \eta^{n-1,n-1}) = \eta^{n,n}.
\end{aligned}$$

□

The next two lemmas describe the distributions of the arrays.

Lemma 6.2. *Fix $0 < \rho_1 < \cdots < \rho_n$ and let the multiline configuration $I = (I^1, \dots, I^n)$ have distribution $\nu^{(\rho_1,\dots,\rho_n)}$. Let $\{\eta^{i,j}\}_{1\le j\le i\le n}$ and $\{\xi^{i,j}\}_{1\le j\le i\le n}$ be the arrays defined above. Then for each $1 \le i, j \le n$, configuration $(\eta^{j,j}, \eta^{j+1,j}, \dots, \eta^{n,j})$ has distribution $\nu^{(\rho_j,\, \rho_{j+1},\dots,\, \rho_n)}$ and configuration $(\xi^{i,1}, \xi^{i,2}, \dots, \xi^{i,i})$ has distribution $\nu^{(\rho_1,\, \rho_2,\dots,\, \rho_i)}$. In particular, each $\eta^{i,j}$ has distribution ν^{ρ_i} and each $\xi^{i,j}$ has distribution ν^{ρ_j}.*

Proof. First we prove the claim for $(\eta^{j,j}, \eta^{j+1,j}, \dots, \eta^{n,j})$. Recall that by definition $\xi^{j,j} = \eta^{j,j}$.

For $j = 1$, the definitions give $(\xi^{1,1} = \eta^{1,1}, \eta^{2,1}, \dots, \eta^{n,1}) = (I^1, I^2, \dots, I^n) \sim \nu^{(\rho_1,\, \rho_2,\dots,\, \rho_n)}$.

Let $j \in [\![2, n]\!]$. Assume inductively that

$$(\xi^{j-1,j-1} = \eta^{j-1,j-1}, \eta^{j,j-1}, \dots, \eta^{n,j-1}) \sim \nu^{(\rho_{j-1},\, \rho_j,\dots,\, \rho_n)}.$$

The mapping from $(\eta^{j,j-1}, \dots, \eta^{n,j-1})$ to $(\eta^{j,j}, \dots, \eta^{n,j})$ is the mapping (5-3) of the multiline process, with $\xi^{j-1,j-1}$ as the external driving weights ω. Namely, this mapping is carried out by iterating

$$\eta^{j+k,j} = D(\eta^{j+k,j-1}, \xi^{j+k-1,j-1}), \qquad \xi^{j+k,j-1} = R(\eta^{j+k,j-1}, \xi^{j+k-1,j-1})$$

for $k = 0, 1, \dots, n - j$. Then $(\eta^{j,j}, \eta^{j+1,j}, \dots, \eta^{n,j}) \sim \nu^{(\rho_j,\, \rho_{j+1},\dots,\, \rho_n)}$ follows from the invariance in Theorem 5.1.

Next the proof for $(\xi^{i,1}, \xi^{i,2}, \dots, \xi^{i,i})$. The claim is immediate for $i = 1$ because there is just one sequence $\eta^{1,1} = I^1 = \xi^{1,1} \sim \nu^{\rho_1}$. Let $i \in [\![2, n]\!]$ and assume inductively that $(\xi^{i-1,1}, \xi^{i-1,2}, \dots, \xi^{i-1,i-1}) \sim \nu^{(\rho_1,\, \rho_2,\dots,\, \rho_{i-1})}$. By construction, $\eta^{i,1} = I^i \sim \nu^{\rho_i}$ is independent of $\xi^{i-1,\bullet}$, and hence

$$(\eta^{i,1}, \xi^{i-1,1}, \xi^{i-1,2}, \dots, \xi^{i-1,i-1}) \sim \nu^{(\rho_i,\, \rho_1,\, \rho_2,\dots,\, \rho_{i-1})}.$$

Now we transform the sequence above by repeated application of the mapping $(\eta^{i,\ell}, \xi^{i-1,\ell}) \mapsto (\xi^{i,\ell}, \eta^{i,\ell+1})$ defined by (6-3):

$$\begin{aligned} \eta^{i,\ell+1} &= D(\eta^{i,\ell}, \xi^{i-1,\ell}), \\ \xi^{i,\ell} &= R(\eta^{i,\ell}, \xi^{i-1,\ell}) \end{aligned}$$

for $\ell = 1, \dots, i - 1$. The pair to be transformed next slides successively to the right. The succession of sequences produced by this process is displayed below, beginning with the first one from above. The pair to which the mapping is applied next is enclosed in the box. The distribution follows from Lemma B.2:

$$\begin{aligned}
&(\boxed{\eta^{i,1}, \xi^{i-1,1}}, \xi^{i-1,2}, \xi^{i-1,3}, \dots, \xi^{i-1,i-1}) \sim \nu^{(\rho_i,\, \rho_1,\, \rho_2,\, \rho_3,\dots,\, \rho_{i-1})} \\
&(\xi^{i,1}, \boxed{\eta^{i,2}, \xi^{i-1,2}}, \xi^{i-1,3}, \dots, \xi^{i-1,i-1}) \sim \nu^{(\rho_1,\, \rho_i,\, \rho_2,\, \rho_3,\dots,\, \rho_{i-1})} \\
&\qquad\qquad \ddots \\
&(\xi^{i,1}, \dots, \xi^{i,\ell-1}, \boxed{\eta^{i,\ell}, \xi^{i-1,\ell}}, \xi^{i-1,\ell+1}, \dots, \xi^{i-1,i-1}) \sim \nu^{(\rho_1,\dots,\rho_{\ell-1},\, \rho_i,\, \rho_\ell,\, \rho_{\ell+1}\dots,\, \rho_{i-1})} \\
&\qquad\qquad \ddots \\
&(\xi^{i,1}, \dots, \xi^{i,i-1}, \eta^{i,i}) \sim \nu^{(\rho_1,\dots,\, \rho_{i-1},\, \rho_i)}.
\end{aligned}$$

To complete the induction from $i - 1$ to i, set $\xi^{i,i} = \eta^{i,i}$. □

Remark 6.3 (notation). To keep track of the inputs when processes are constructed by queueing mappings (2-22), superscripts indicate the arrival and service processes used in the construction. This works as follows when the arrival process is I and the service process is ω:

- G^I denotes a function that satisfies $I_k = G^I_k - G^I_{k-1}$.
- $\widetilde{G}^{I,\omega}$ is the process defined by (2-18) whose increments are the output $\tilde{I}^{I,\omega}_k = \widetilde{G}^{I,\omega}_k - \widetilde{G}^{I,\omega}_{k-1}$, and so $\tilde{I}^{I,\omega} = D(I,\omega)$.
- $J^{I,\omega} = S(I,\omega)$ is the process defined by (2-20) as $J^{I,\omega}_k = \widetilde{G}^{I,\omega}_k - G^I_k$.
- $\widetilde{\omega}^{I,\omega} = R(I,\omega)$.

Lemma 6.4. *Fix $0 < \rho_1 < \cdots < \rho_n$ and let the multiline configuration $I = (I^1, \ldots, I^n)$ have distribution $\nu^{(\rho_1,\ldots,\rho_n)}$. Let $\eta = (\eta^1, \ldots, \eta^n) = \mathcal{D}^{(n)}(I)$ and let $\{\eta^{i,j}\}$ and $\{\xi^{i,j}\}$ be the arrays constructed above. Then for each $m \in [\![2, n]\!]$ and $k \in \mathbb{Z}$, the following random variables are independent*:

$$\{\xi^{m,1}_i\}_{i\le k},\ \{\xi^{m,2}_i\}_{i\le k}, \ldots, \{\xi^{m,m-1}_i\}_{i\le k},\ \{\eta^m_i\}_{i\le k-1},\ \eta^m_k - \eta^{m-1}_k,\ \eta^{m-1}_k - \eta^{m-2}_k, \ldots, \eta^2_k - \eta^1_k,\ \eta^1_k.$$

Proof. Index k is fixed throughout the proof. We begin with the case $m = 2$.

By the definitions, $\eta^1 = I^1$,

$$\xi^{2,1} = R(\eta^{2,1}, \xi^{1,1}) = R(I^2, I^1) = \widetilde{\omega}^{I^2,I^1} \quad\text{and}\quad \eta^2 = D(I^2, I^1) = \tilde{I}^{I^2,I^1}.$$

Hence $\xi^{2,1}_i = I^2_i \wedge J^{I^2,I^1}_{i-1}$ and $\eta^2_k - \eta^1_k = (I^2_k - J^{I^2,I^1}_{k-1})^+$. By Lemma B.2(a), $\{\tilde{I}^{I^2,I^1}_i\}_{i\le k-1}$, J^{I^2,I^1}_{k-1}, $\{\widetilde{\omega}^{I^2,I^1}_i\}_{i\le k-1}$, I^2_k, I^1_k are independent. To be precise, Lemma B.2(a) gives the independence of $\{\widetilde{\omega}^{I^2,I^1}_i\}_{i\le k-1}$, $\{\tilde{I}^{I^2,I^1}_i\}_{i\le k-1}$, and J^{I^2,I^1}_{k-1}. These are functions of $\{I^2_i, I^1_i\}_{i\le k-1}$, and thereby independent of I^2_k, I^1_k. Properties of independent exponentials (Lemma B.1(i)) imply that

$$\xi^{2,1}_k = I^2_k \wedge J^{I^2,I^1}_{k-1} \text{ and } \eta^2_k - \eta^1_k = (I^2_k - J^{I^2,I^1}_{k-1})^+ \text{ are mutually independent.} \tag{6-5}$$

Altogether we have that $\{\xi^{2,1}_i\}_{i\le k}$, $\{\eta^2_i\}_{i\le k-1}$, $\eta^2_k - \eta^1_k$, η^1_k are independent.

Let $m \ge 3$ and make an induction assumption:

$$\{\xi^{m-1,1}_i\}_{i\le k}, \ldots, \{\xi^{m-1,m-2}_i\}_{i\le k},\ \{\eta^{m-1}_i\}_{i\le k-1},\ \eta^{m-1}_k - \eta^{m-2}_k, \ldots, \eta^2_k - \eta^1_k, \eta^1_k \text{ are independent.} \tag{6-6}$$

The previous paragraph verified this assumption for $m = 3$.

Since $\eta^{m,1} = I^m$ is independent of all the variables in (6-6), apply Lemma B.2(a) to the pair $\xi^{m,1} = R(\eta^{m,1}, \xi^{m-1,1})$, $\eta^{m,2} = D(\eta^{m,1}, \xi^{m-1,1})$ to conclude the independence of

$$\begin{gathered}\{\xi^{m,1}_i\}_{i\le k},\ \{\eta^{m,2}_i\}_{i\le k},\ \{\xi^{m-1,2}_i\}_{i\le k}, \ldots, \{\xi^{m-1,m-2}_i\}_{i\le k},\\ \{\eta^{m-1}_i\}_{i\le k-1},\ \eta^{m-1}_k - \eta^{m-2}_k, \ldots, \eta^2_k - \eta^1_k, \eta^1_k.\end{gathered} \tag{6-7}$$

This starts an induction on $j = 2, 3, \ldots, m-1$, whose induction assumption is the independence of

$$\begin{gathered}\{\xi^{m,1}_i\}_{i\le k}, \ldots, \{\xi^{m,j-1}_i\}_{i\le k}, \{\eta^{m,j}_i\}_{i\le k}, \{\xi^{m-1,j}_i\}_{i\le k}, \ldots, \{\xi^{m-1,m-2}_i\}_{i\le k},\\ \{\eta^{m-1}_i\}_{i\le k-1}, \eta^{m-1}_k - \eta^{m-2}_k, \ldots, \eta^2_k - \eta^1_k, \eta^1_k.\end{gathered} \tag{6-8}$$

The induction step is the application of Lemma B.2(a) to the pair $\xi^{m,j} = R(\eta^{m,j}, \xi^{m-1,j})$, $\eta^{m,j+1} = D(\eta^{m,j}, \xi^{m-1,j})$ to conclude the independence of

$$\begin{gathered}\{\xi_i^{m,1}\}_{i\le k}, \dots, \{\xi_i^{m,j-1}\}_{i\le k},\ \{\xi_i^{m,j}\}_{i\le k},\ \{\eta_i^{m,j+1}\}_{i\le k},\\ \{\xi_i^{m-1,j+1}\}_{i\le k}, \dots, \{\xi_i^{m-1,m-1}\}_{i\le k},\ \{\eta_i^{m-1}\}_{i\le k-1},\ \eta_k^{m-1}-\eta_k^{m-2}, \dots, \eta_k^2-\eta_k^1,\ \eta_k^1.\end{gathered} \tag{6-9}$$

Thus the induction assumption (6-8) for j has been advanced to $j+1$ in (6-9).

At the end of the j-induction we have the independence of

$$\begin{gathered}\{\xi_i^{m,1}\}_{i\le k}, \dots, \{\xi_i^{m,m-2}\}_{i\le k}, \{\eta_i^{m,m-1}\}_{i\le k},\\ \{\eta_i^{m-1}\}_{i\le k-1},\ \eta_k^{m-1}-\eta_k^{m-2}, \dots, \eta_k^2-\eta_k^1, \eta_k^1.\end{gathered} \tag{6-10}$$

Split $\{\eta_i^{m,m-1}\}_{i\le k}$ into the independent pieces $\{\eta_i^{m,m-1}\}_{i\le k-1}$ and $\eta_k^{m,m-1}$. Combine the former with $\{\eta_i^{m-1}\}_{i\le k-1}$, Lemma B.2(a), and the transformations $\xi^{m,m-1} = R(\eta^{m,m-1}, \eta^{m-1})$, $\eta^m = D(\eta^{m,m-1}, \eta^{m-1})$ to form the independent variables $\{\xi_i^{m,m-1}\}_{i\le k-1}$, $\{\eta_i^m\}_{i\le k-1}$ and $J_{k-1}^{\eta^{m,m-1},\eta^{m-1}}$. Transform the independent pair $(\eta_k^{m,m-1}, J_{k-1}^{\eta^{m,m-1},\eta^{m-1}})$ into the independent pair of $\xi_k^{m,m-1} = \eta_k^{m,m-1} \wedge J_{k-1}^{\eta^{m,m-1},\eta^{m-1}}$ and $\eta_k^m - \eta_k^{m-1} = (\eta_k^{m,m-1} - J_{k-1}^{\eta^{m,m-1},\eta^{m-1}})^+$. Attach $\xi_k^{m,m-1}$ to the sequence $\{\xi_i^{m,m-1}\}_{i\le k-1}$. After these steps, we have the independence of

$$\begin{gathered}\{\xi_i^{m,1}\}_{i\le k}, \dots, \{\xi_i^{m,m-2}\}_{i\le k}, \{\xi_i^{m,m-1}\}_{i\le k},\\ \{\eta_i^m\}_{i\le k-1},\ \eta_k^m-\eta_k^{m-1}, \eta_k^{m-1}-\eta_k^{m-2}, \dots, \eta_k^2-\eta_k^1, \eta_k^1.\end{gathered} \tag{6-11}$$

Thus the induction assumption (6-6) has been advanced from $m-1$ to m. □

Proof of Theorem 3.4. Fix $1 < \rho_1 < \cdots < \rho_n$ and let the multiline configuration $I = (I^0, \dots, I^n)$ have distribution $\nu^{(1,\rho_1,\dots,\rho_n)}$. Let $\eta = (\eta^0, \dots, \eta^n) = \mathcal{D}^{(n+1)}(I)$. By Theorem 3.2 proved above, $(\overline{Y}_t, \overline{B}_t^{\rho_1,\boldsymbol{e}_1}, \dots, \overline{B}_t^{\rho_n,\boldsymbol{e}_1}) \overset{d}{=} \eta \sim \mu^{(1,\rho_1,\dots,\rho_n)}$. Lemma 6.4 gives the independence of the components of the vector

$$(\eta_k^1, \eta_k^2-\eta_k^1, \dots, \eta_k^n-\eta_k^{n-1}) \overset{d}{=} (Y_x, B_{x-\boldsymbol{e}_1,x}^{\rho_1}-Y_x, B_{x-\boldsymbol{e}_1,x}^{\rho_2}-B_{x-\boldsymbol{e}_1,x}^{\rho_1}, \dots, B_{x-\boldsymbol{e}_1,x}^{\rho_n}-B_{x-\boldsymbol{e}_1,x}^{\rho_{n-1}}).$$

(Above $k \in \mathbb{Z}$ and $x \in \mathbb{Z}^2$ are arbitrary.)

The distribution of an increment $\eta_k^m - \eta_k^{m-1}$ can be computed from the 2-component mapping $(\eta^{m-1}, \eta^m) = \mathcal{D}^{(2)}(I^{m-1}, I^m) = (I^{m-1}, D(I^m, I^{m-1}))$ where $(I^{m-1}, I^m) \sim \nu^{\rho_{m-1},\rho_m}$. The first equation of (2-24) gives

$$\eta_k^m - \eta_k^{m-1} = \eta_k^m - I_k^{m-1} = (I_k^m - J_{k-1}^{I^m,I^{m-1}})^+.$$

The right-hand side has the distribution in (3-7) with $(\lambda, \rho) = (\rho_{m-1}, \rho_m)$ because, by the structure of the queueing mapping, I_k^m and $J_{k-1}^{I^m,I^{m-1}}$ are independent exponentials with parameters ρ_m^{-1} and $\rho_{m-1}^{-1} - \rho_m^{-1}$.

A computation of the Laplace transform of the increment $X(\rho) - X(\lambda)$ of the process defined by (3-6) gives, for $\rho > \lambda \ge 1$ and $\alpha > 0$,

$$E[e^{-\alpha(X(\rho)-X(\lambda))}] = \frac{1+\lambda\alpha}{1+\rho\alpha}. \tag{6-12}$$

This is the Laplace transform of the distribution in (3-7). Thus $\eta_k^m - \eta_k^{m-1}$ has the same distribution as $X(\rho_m) - X(\rho_{m-1})$.

To summarize, the nondecreasing cadlag processes $B^{\bullet}_{x-\boldsymbol{e}_1,x}$ and $X(\cdot)$ have identically distributed initial values (both $B^1_{x-\boldsymbol{e}_1,x} = Y_x$ and $X(1)$ are Exp(1)-distributed) and identically distributed independent increments. Hence the processes are equal in distribution. □

6C. ***Bivariate Busemann process on a line.*** The remainder of this section proves statements for the sequence $\{(B^{\lambda}_{(k-1,t),(k,t)}, B^{\rho}_{(k-1,t),(k,t)})\}_{k\in\mathbb{Z}}$ that has distribution $\mu^{(\lambda,\rho)}$. We use the following notation: Let $\rho > \lambda > 0$, $(I^1, I^2) \sim \nu^{(\lambda,\rho)}$ and $(\eta^1, \eta^2) = \mathcal{D}^{(2)}(I^1, I^2) = (I^1, D(I^2, I^1))$. Then $(\eta^1, \eta^2) \sim \mu^{(\lambda,\rho)}$. Let $J = J^{I^2,I^1} = S(I^2, I^1)$.

Proof of Theorem 3.5. The next auxiliary lemma identifies a reversible Markov chain.

Lemma 6.5. *Let $X_i = J_{i-1} - I^2_i$. Then $\{X_i\}_{i\in\mathbb{Z}}$ and $\{X_i^+\}_{i\in\mathbb{Z}}$ are stationary reversible Markov chains. $\{X_i^-\}_{i\in\mathbb{Z}}$ is not a Markov chain.*

Proof. From the second equation of (2-24),

$$X_{i+1} = J_i - I^2_{i+1} = I^1_i + (J_{i-1} - I^2_i)^+ - I^2_{i+1} = X_i^+ + I^1_i - I^2_{i+1}.$$

Since J_{i-1} is a function of $(I^1_k, I^2_k)_{k\le i-1}$, X_i is independent of (I^1_i, I^2_{i+1}). Schematically, we can express the transition probability as $X_{i+1} = X_i^+ + \mathrm{Exp}(\lambda^{-1}) - \mathrm{Exp}(\rho^{-1})$, where the three terms on the right-hand side are independent.

Similarly, using conservation (2-25) and the dual equations (4-3),

$$X_i = J_{i-1} - I^2_i = J_i - \eta^2_i = \widetilde{\omega}_{i+1} + (J_{i+1} - \eta^2_{i+1})^+ - \eta^2_i = X^+_{i+1} + \widetilde{\omega}_{i+1} - \eta^2_i. \tag{6-13}$$

J_i and η^2_i are independent by Lemma B.2(a), and hence the triple $(J_i, \eta^2_i, I^2_{i+1})$ is independent. Consequently so is the triple

$$(X_{i+1}, \widetilde{\omega}_{i+1}, \eta^2_i) = (J_i - I^2_{i+1}, J_i \wedge I^2_{i+1}, \eta^2_i)$$

and we can express (6-13) as $X_i = X^+_{i+1} + \mathrm{Exp}(\lambda^{-1}) - \mathrm{Exp}(\rho^{-1})$ where again the three terms on the right-hand side are independent. The transitions from X_i to X_{i+1} and back are the same.

From the equations above we obtain equations that show X_i^+ as a reversible Markov chain.

Writing temporarily $U_i = I^1_{i-1} - I^2_i$, we get these equations for X^-_{i+1}:

$$X^-_{i+1} = (X_i^+ + U_{i+1})^- = \big((X^+_{i-1} + U_i)^+ + U_{i+1}\big)^-.$$

Conditioned on $X_i \ge 0$, $X_i \sim \mathrm{Exp}(\lambda^{-1} - \rho^{-1})$. Thus

$$\begin{aligned} P(X^-_{i+1} = 0 \mid X^-_i = 0) &= P(X^+_i + U_{i+1} \ge 0 \mid X_i \ge 0) \\ &= P\{\mathrm{Exp}(\lambda^{-1} - \rho^{-1}) + U_{i+1} \ge 0\}. \end{aligned} \tag{6-14}$$

For the next calculation, note that $X_{i-1} < 0$ implies $X_i = X^+_{i-1} + U_i = U_i$ and then $X_{i+1} = U^+_i + U_{i+1}$.

$$\begin{aligned} P(X^-_{i+1} = 0 \mid X^-_i = 0, X^-_{i-1} > 0) &= \frac{P(X^-_{i+1} = 0, X^-_i = 0, X^-_{i-1} > 0)}{P(X^-_i = 0, X^-_{i-1} > 0)} \\ &= \frac{P(U_i + U_{i+1} \ge 0, U_i \ge 0, X_{i-1} < 0)}{P(U_i \ge 0, X_{i-1} < 0)} \\ &= P(U_i + U_{i+1} \ge 0 \mid U_i \ge 0) \\ &= P\{\mathrm{Exp}(\lambda^{-1}) + U_{i+1} \ge 0\}. \end{aligned} \tag{6-15}$$

We used above the independence of X_{i-1} from (U_i, U_{i+1}) and then the conditional distribution $U_i \sim \mathrm{Exp}(\lambda^{-1})$, given that $U_i \ge 0$. The conditional distributions in (6-14) and (6-15) do not agree, and consequently X^-_i is not a Markov chain. □

Since $\eta^2_k - \eta^1_k = \eta^2_k - I^1_k = (I^2_k - J_{k-1})^+ = X^-_k$, we conclude that $\eta^2_k - \eta^1_k$ is not a Markov chain, but it is a function of a reversible Markov chain. Part (a) of Theorem 3.5 has been proved.

We give here two more auxiliary lemmas.

Lemma 6.6. *The process $(\eta^1_k, \eta^2_k)_{k\in\mathbb{Z}}$ is not a Markov chain.*

Proof. The construction gives $\eta^2_{k+1} = I^1_{k+1} + (I^2_{k+1} - J_k)^+$. On the right, the variables I^1_{k+1} and I^2_{k+1} are independent and independent of J_k and $(\eta^1_j, \eta^2_j)_{j\le k}$. The conclusion of the lemma follows from showing that conditioning on $\eta^1_k = \eta^2_k$ gives J_k an unbounded distribution, while conditioning on $\eta^1_{k-1} < \eta^2_{k-1}$ and $\eta^1_k = \eta^2_k$ implies $J_k \le \eta^1_{k-1} + \eta^1_k$. Thus conditioning on (η^1_k, η^2_k) does not completely decouple η^2_{k+1} from the earlier past.

From the three independent variables (J_{k-1}, I^1_k, I^2_k) the queueing formulas define

$$\eta^1_k = I^1_k, \quad \eta^2_k = I^1_k + (I^2_k - J_{k-1})^+ \quad \text{and} \quad J_k = I^1_k + (J_{k-1} - I^2_k)^+. \tag{6-16}$$

The condition $\eta^1_k = \eta^2_k$ is equivalent to $J_{k-1} \ge I^2_k$, and conditioning on this implies

$$J_{k-1} - I^2_k \sim \mathrm{Exp}(\lambda^{-1} - \rho^{-1}).$$

Thus J_k is unbounded.

For the second scenario consider the five independent variables $(J_{k-2}, I^1_{k-1}, I^2_{k-1}, I^1_k, I^2_k)$ and augment (6-16) with the equations of the prior step:

$$\eta^1_{k-1} = I^1_{k-1}, \quad \eta^2_{k-1} = I^1_{k-1} + (I^2_{k-1} - J_{k-2})^+ \quad \text{and} \quad J_{k-1} = I^1_{k-1} + (J_{k-2} - I^2_{k-1})^+. \tag{6-17}$$

Now $\eta^1_{k-1} < \eta^2_{k-1}$ implies $J_{k-1} = I^1_{k-1}$ and then $\eta^1_k = \eta^2_k$ implies $J_k = I^1_k + J_{k-1} - I^2_k = I^1_k + I^1_{k-1} - I^2_k$. Hence $J_k \le I^1_k + I^1_{k-1} = \eta^1_k + \eta^1_{k-1}$. The lemma is proved.

With service process $I^1 = \eta^1$, arrival process I^2 and departure process η^2, the queueing explanation of the proof is that $\eta^1_k = \eta^2_k$ implies that customer k had to wait before entering service, and hence delays from the past can influence the next interdeparture time η^2_{k+1}. □

Lemma 6.7. *The pair $((\eta^1_k, \eta^2_k), (\eta^1_{k+1}, \eta^2_{k+1}))$ and its transpose $((\eta^1_{k+1}, \eta^2_{k+1}), (\eta^1_k, \eta^2_k))$ are not equal in distribution.*

Proof. By the queueing construction, $\eta^1_{k+1}=I^1_{k+1}$ is independent of (η^1_k,η^2_k) because the latter pair is a function of $(I^1_i,I^2_i)_{i\le k}$. To see that $\eta^1_k=I^1_k$ is not independent of $(\eta^1_{k+1},\eta^2_{k+1})$, write

$$\eta^2_{k+1}-\eta^1_{k+1}=(I^2_{k+1}-J_k)^+=(I^2_{k+1}-I^1_k-[J_{k-1}-I^2_k]^+)^+$$

where all four variables in the last expression are independent. □

Part (b) of Theorem 3.5 follows from the two lemmas above. □

Proof of Theorem 3.6 and Remark 3.7. Part (a) comes from translating the condition $B^\rho_{(k-1)\boldsymbol{e}_1,k\boldsymbol{e}_1}=Y_{k\boldsymbol{e}_1}$ into a statement about the queueing mapping $\tilde I=D(I,\omega)$.

By (2-13),

$$\mathbb{P}\{\xi_x=0\}=\mathbb{P}\{B^\rho_{x-\boldsymbol{e}_2,x}<B^\rho_{x-\boldsymbol{e}_1,x}\}=1-\rho^{-1}$$

from the independence and exponential distributions in Theorem 2.1(ii). From (3-7),

$$\mathbb{P}\{\xi^{\lambda,\rho}_x=0\}=\mathbb{P}\{B^\rho_{x-\boldsymbol{e}_1,x}>B^\lambda_{x-\boldsymbol{e}_1,x}\}=\frac{\rho-\lambda}{\rho}.$$

To calculate $\mathbb{P}\{\xi^{\lambda,\rho}_x=n\}$ for $n\ge 1$ we put x on the x-axis and use the distribution $(\bar B^{\lambda,\boldsymbol{e}_1}_0,\bar B^{\rho,\boldsymbol{e}_1}_0)\overset{d}{=}(\eta^1,\eta^2)\sim\mu^{(\lambda,\rho)}$ given by Theorem 3.2, with the notation from the start of Section 6C. By setting $\lambda=1$ the same calculation gives $\mathbb{P}\{\xi_x=n\}$ because $\overline{Y}_0=\bar B^{1,\boldsymbol{e}_1}_0$.

$$\begin{aligned}\mathbb{P}\{\xi^{\lambda,\rho}_{n\boldsymbol{e}_1}=n\}&=\mathbb{P}\{B^\rho_{-\boldsymbol{e}_1,0}>B^\lambda_{-\boldsymbol{e}_1,0},\,B^\rho_{0,\boldsymbol{e}_1}=B^\lambda_{0,\boldsymbol{e}_1},\,B^\rho_{\boldsymbol{e}_1,2\boldsymbol{e}_1}=B^\lambda_{\boldsymbol{e}_1,2\boldsymbol{e}_1},\dots,B^\rho_{(n-1)\boldsymbol{e}_1,n\boldsymbol{e}_1}=B^\lambda_{(n-1)\boldsymbol{e}_1,n\boldsymbol{e}_1}\}\\&=P\{\eta^2_0>I^1_0,\,\eta^2_1=I^1_1,\,\eta^2_2=I^1_2,\dots,\eta^2_n=I^1_n\}\\&=P\{I^2_0>J_{-1},\,I^2_1\le J_0,\,I^2_2\le J_1,\dots,I^2_n\le J_{n-1}\}\end{aligned}\tag{6-18}$$

The last equality used $\eta^2_i=I^1_i+(I^2_i-J_{i-1})^+$ repeatedly: $\eta^2_i>I^1_i$ is equivalent to $I^2_i>J_{i-1}$.

Next apply repeatedly the equation $J_i=I^1_i+(J_{i-1}-I^2_i)^+$ inside the last probability in (6-18). $I^2_0>J_{-1}$ implies $J_0=I^1_0$. Then $I^2_1\le J_0$ implies $J_1=I^1_1+J_0-I^2_1=I^1_1+I^1_0-I^2_1$. Assume inductively that

$$J_i=I^1_i+\cdots+I^1_0-I^2_1-\cdots-I^2_i.\tag{6-19}$$

Then $I^2_{i+1}\le J_i$ implies

$$J_{i+1}=I^1_{i+1}+J_i-I^2_{i+1}=I^1_{i+1}+(I^1_i+\cdots+I^1_0-I^2_1-\cdots-I^2_i)-I^2_{i+1}$$

and the induction goes from i to $i+1$. Substitute (6-19) for $J_0,\dots,J_{n-1}$ in the last probability in (6-18). Use the independence of the variables J_{-1}, $\{I^1_i,I^2_i\}_{i\ge 0}$. Let S^α_m denote the sum of m i.i.d. Exp(α) random variables, with S and $\widetilde S$ denoting independent sums.

$$\begin{aligned}\mathbb{P}\{\xi^{\lambda,\rho}_{n\boldsymbol{e}_1}=n\}&=P\{I^2_0>J_{-1},\,I^2_1\le I^1_0,\,I^2_2+I^2_1\le I^1_1+I^1_0,\dots,I^2_n+\cdots+I^2_1\le I^1_{n-1}+\cdots+I^1_0\}\\&=P\{I^2_0>J_{-1}\}\,P\{S^{\rho^{-1}}_m\le\widetilde S^{\lambda^{-1}}_m\ \forall m\in[n]\}=\frac{\rho-\lambda}{\rho}\sum_{k=0}^{n-1}C(n-1,k)\frac{\rho^k\lambda^n}{(\lambda+\rho)^{n+k}}.\end{aligned}\tag{6-20}$$

The last line comes from the independence of I^2_0 and J_{-1}, their distributions $I^2_0\sim\mathrm{Exp}(\rho^{-1})$ and $J_{-1}\sim\mathrm{Exp}(\lambda^{-1}-\rho^{-1})$, and Lemma B.3. □

Appendix A. Queues

We prove elementary lemmas about the queueing mappings. Unless otherwise stated, the weights are real numbers without any probability distributions.

Lemma A.1. *Fix $0 \le a < b$. Let $I = (I_k)_{k\in\mathbb{Z}}$ and $\omega = (\omega_j)_{j\in\mathbb{Z}}$ in $\mathbb{R}_{\ge 0}^{\mathbb{Z}}$ satisfy*

$$\varliminf_{m\to-\infty} \frac{1}{|m|}\sum_{i=m}^{0} I_i \ge b \quad \textit{and} \quad \lim_{m\to-\infty} \frac{1}{|m|}\sum_{i=m}^{0} \omega_i = a. \tag{A-1}$$

Then $\tilde{I} = D(I, \omega)$ is well-defined and satisfies $\varliminf_{m\to-\infty} |m|^{-1}\sum_{i=m}^{0} \tilde{I}_i \ge b$.

Proof. Assumption (2-17) is obviously satisfied. Without loss of generality assume $G_0 = 0$. Let $0 < \varepsilon < (b-a)/3$. Then for large enough n,

$$\begin{aligned}\widetilde{G}_{-n} &= \sup_{k:\,k\le -n}\Biggl\{G_k + \sum_{i=k}^{-n}\omega_i\Biggr\} = \sup_{k:\,k\le -n}\Biggl\{-\sum_{i=k+1}^{0} I_i + \sum_{i=k}^{0}\omega_i\Biggr\} - \sum_{i=-n+1}^{0}\omega_i \\ &\le \sup_{k:\,k\le -n}\bigl\{-|k|(b-\varepsilon) + |k|(a+\varepsilon)\bigr\} - n(a-\varepsilon) = n(-b+3\varepsilon).\end{aligned}$$

Since $\sum_{i=m+1}^{0} \tilde{I}_i = \widetilde{G}_0 - \widetilde{G}_m$, this proves $\varliminf_{m\to-\infty} |m|^{-1}\sum_{i=m}^{0} \tilde{I}_i \ge b$. □

Lemma A.2. *Fix $0 \le a < b$. Assume given nonnegative real sequences $I = (I_i)_{i\in\mathbb{Z}}$, $\omega = (\omega_i)_{i\in\mathbb{Z}}$, $I^{(h)} = (I_i^{(h)})_{i\in\mathbb{Z}}$ and $\omega^{(h)} = (\omega_i^{(h)})_{i\in\mathbb{Z}}$ where $h \in \mathbb{Z}_{>0}$ is an index. Assume $I_i^{(h)} \to I_i$ and $\omega_i^{(h)} \to \omega_i$ as $h \to \infty$ for all $i \in \mathbb{Z}$, and furthermore,*

$$\varlimsup_{\substack{m\to-\infty\\ h\to\infty}} \Biggl|\frac{1}{|m|}\sum_{i=m}^{0} I_i^{(h)} - b\Biggr| = 0 \quad \textit{and} \quad \varlimsup_{\substack{m\to-\infty\\ h\to\infty}} \Biggl|\frac{1}{|m|}\sum_{i=m}^{0} \omega_i^{(h)} - a\Biggr| = 0. \tag{A-2}$$

Then $\tilde{I} = D(I, \omega)$ and $\widetilde{\omega} = R(I, \omega)$ are well-defined, as are $\tilde{I}^{(h)} = D(I^{(h)}, \omega^{(h)})$ and $\widetilde{\omega}^{(h)} = R(I^{(h)}, \omega^{(h)})$ for large enough h. We have the limits

$$\lim_{h\to\infty} \tilde{I}_i^{(h)} = \tilde{I}_i \quad \textit{and} \quad \lim_{h\to\infty} \widetilde{\omega}_i^{(h)} = \widetilde{\omega}_i \qquad \textit{for all } i \in \mathbb{Z} \tag{A-3}$$

and

$$\varlimsup_{\substack{m\to-\infty\\ h\to\infty}} \Biggl|\frac{1}{|m|}\sum_{i=m}^{0} \tilde{I}_i^{(h)} - b\Biggr| = 0 \quad \textit{and} \quad \varlimsup_{\substack{m\to-\infty\\ h\to\infty}} \Biggl|\frac{1}{|m|}\sum_{i=m}^{0} \widetilde{\omega}_i^{(h)} - a\Biggr| = 0. \tag{A-4}$$

Proof. Assumption (2-17) is satisfied to make $\tilde{I}^{(h)} = D(I^{(h)}, \omega^{(h)})$ well-defined for large enough h.

We can assume $G_0^{(h)} = 0$. Compute $\tilde{I}^{(h)} = D(I^{(h)}, \omega^{(h)})$ as the increments of the function

$$\widetilde{G}_\ell^{(h)} = \sup_{k\le \ell}\Biggl\{G_k^{(h)} + \sum_{i=k}^{\ell}\omega_i^{(h)}\Biggr\}. \tag{A-5}$$

Let k_0 be a maximizer in (2-18) for $\widetilde{G}_\ell$. Then

$$\varliminf_{h\to\infty} \widetilde{G}_\ell^{(h)} \geq \varliminf_{h\to\infty} \Big\{ G_{k_0}^{(h)} + \sum_{i=k_0}^{\ell} \omega_i^{(h)} \Big\} = G_{k_0} + \sum_{i=k_0}^{\ell} \omega_i = \widetilde{G}_\ell. \tag{A-6}$$

Let $k(h)$ be a maximizer in (A-5). If $\varlimsup_{h\to\infty} \widetilde{G}_\ell^{(h)} \leq \widetilde{G}_\ell$ fails then it must be that $k(h) \to -\infty$ along a subsequence. But we can write

$$\begin{aligned}\widetilde{G}_\ell^{(h)} &= G_{k(h)}^{(h)} + \sum_{i=k(h)}^{\ell} \omega_i^{(h)} \\ &= -\sum_{i=k(h)+1}^{0} I_i^{(h)} + \sum_{i=k(h)}^{0} \omega_i^{(h)} + \Big(\mathbf{1}_{\ell>0} \sum_{i=1}^{\ell} \omega_i^{(h)} - \mathbf{1}_{\ell<0} \sum_{i=\ell+1}^{0} \omega_i^{(h)} \Big)\end{aligned} \tag{A-7}$$

which converges to $-\infty$ as $k(h) \to -\infty$ by the assumptions and thereby contradicts (A-6). We have now proved that

$$\lim_{h\to\infty} \widetilde{G}_\ell^{(h)} = \widetilde{G}_\ell \qquad \text{for all } \ell \in \mathbb{Z} \tag{A-8}$$

and thereby verified (A-3) for $\tilde{I}^{(h)}$.

Let $0 < \varepsilon < (b-a)/3$. By assumption (A-2) there exist finite $n_1(\varepsilon)$ and $h_1(\varepsilon)$ such that, when $n \geq n_1(\varepsilon)$ and $h \geq h_1(\varepsilon)$,

$$\begin{aligned}\widetilde{G}_{-n}^{(h)} &= \sup_{k:\,k\leq -n} \Big\{ G_k^{(h)} + \sum_{i=k}^{-n} \omega_i^{(h)} \Big\} = \sup_{k:\,k\leq -n} \Big\{ -\sum_{i=k+1}^{0} I_i^{(h)} + \sum_{i=k}^{0} \omega_i^{(h)} \Big\} - \sum_{i=-n+1}^{0} \omega_i^{(h)} \\ &\leq \sup_{k:\,k\leq -n} \big\{ -|k|(b-\varepsilon) + |k|(a+\varepsilon) \big\} - n(a-\varepsilon) = n(-b+3\varepsilon).\end{aligned}$$

From this,

$$\varlimsup_{m\to-\infty} \sup_{h\geq h_1(\varepsilon)} \frac{\widetilde{G}_m^{(h)}}{|m|} \leq -b+3\varepsilon. \tag{A-9}$$

Since $\sum_{i=m}^{0} \tilde{I}_i^{(h)} = \widetilde{G}_0^{(h)} - \widetilde{G}_{m-1}^{(h)}$ and $\widetilde{G}_0^{(h)} \geq \omega_0^{(h)} \geq 0$, this proves

$$\varliminf_{m\to-\infty} \inf_{h\geq h_1(\varepsilon)} \frac{1}{|m|} \sum_{i=m}^{0} \tilde{I}_i^{(h)} \geq b-3\varepsilon. \tag{A-10}$$

For the complementary upper bound, get a lower bound for $\widetilde{G}_{m-1}^{(h)}$ by taking $k=\ell$ in (A-5).

$$\sum_{i=m}^{0} \tilde{I}_i^{(h)} = \widetilde{G}_0^{(h)} - \widetilde{G}_{m-1}^{(h)} \leq \widetilde{G}_0^{(h)} - G_{m-1}^{(h)} = \widetilde{G}_0^{(h)} + \sum_{i=m}^{0} I_i^{(h)}.$$

Apply limit (A-8) and assumption (A-2). Limit (A-4) has been proved for $\tilde{I}^{(h)}$.

The limits for $\widetilde{\omega}^{(h)}$ follow from the other limits and the generally valid identity

$$\omega_k + I_k = \widetilde{\omega}_k + \tilde{I}_k \tag{A-11}$$

that comes from equations (2-21) and (2-24). □

For reference elsewhere in the paper we state the simple consequence of Lemma A.2 where the sequences are constant functions of h.

Lemma A.3. *Fix $0 \le a < b$. Let $I = (I_k)_{k\in\mathbb{Z}}$ and $\omega = (\omega_j)_{j\in\mathbb{Z}}$ in $\mathbb{R}^{\mathbb{Z}}_{\ge 0}$ satisfy*

$$\lim_{m\to-\infty} \frac{1}{|m|}\sum_{i=m}^{0} I_i = b \quad \textit{and} \quad \lim_{m\to-\infty} \frac{1}{|m|}\sum_{i=m}^{0} \omega_i = a. \tag{A-12}$$

Then $\tilde{I} = D(I,\omega)$ and $\widetilde{\omega} = R(I,\omega)$ are well-defined and satisfy

$$\lim_{m\to-\infty} \frac{1}{|m|}\sum_{i=m}^{0} \tilde{I}_i = b \quad \textit{and} \quad \lim_{m\to-\infty} \frac{1}{|m|}\sum_{i=m}^{0} \widetilde{\omega}_i = a. \tag{A-13}$$

For the purpose of verifying that Busemann functions obey the queueing operation $\tilde{I} = D(I,\omega)$, it is convenient to have a lemma that deduces this from assuming the iterative equations (2-24). The first lemma below makes a statement without randomness.

Lemma A.4. *Let $\{\tilde{I}_k, J_k, I_k, \omega_k\}_{k\in\mathbb{Z}}$ be nonnegative real numbers that satisfy the three assumptions below*:

$$\lim_{m\to-\infty} \sum_{i=m}^{0} (\omega_i - I_{i+1}) = -\infty. \tag{A-14}$$

$$\tilde{I}_k = \omega_k + (I_k - J_{k-1})^+ \quad \textit{and} \quad J_k = \omega_k + (J_{k-1} - I_k)^+ \quad \textit{for all } k \in \mathbb{Z}. \tag{A-15}$$

$$J_k = \omega_k \quad \textit{for infinitely many } k < 0. \tag{A-16}$$

Then $\tilde{I} = D(I,\omega)$ and $J = S(I,\omega)$.

Proof. Rewrite the second equation of (A-15) as follows. Let $W_k = J_k - \omega_k$ and $U_k = \omega_k - I_{k+1}$. Then

$$W_k = (W_{k-1} + U_{k-1})^+. \tag{A-17}$$

This is Lindley's recursion from queueing theory and W_k is the *waiting time* of customer k. Equation (A-17) iterates inductively to give

$$W_k = \biggl\{\biggl(W_\ell + \sum_{i=\ell}^{k-1} U_i\biggr) \bigvee \biggl(\max_{m:\,\ell+1\le m\le k-1} \sum_{i=m}^{k-1} U_i\biggr)\biggr\}^+ \qquad \text{for all } \ell < k. \tag{A-18}$$

We claim that

$$W_k = \biggl(\sup_{m:\,m\le k-1} \sum_{i=m}^{k-1} U_i\biggr)^+ \qquad \text{for all } k \in \mathbb{Z}. \tag{A-19}$$

Dropping the first term on the right in (A-18) and letting $\ell \to -\infty$ gives $\ge$ in (A-19). By assumption (A-16) $W_\ell = 0$ for some $\ell < k$. Then (A-18) gives also $\le$ in (A-19).

The proof is completed by making explicit the content of (A-19). Let G and $\widetilde{G}$ be as defined in the definition of the mappings D and S. Then from (A-19) and (2-23) deduce

$$\begin{aligned} J_k &= \omega_k + \Bigl(\sup_{m:\, m \le k-1} \sum_{i=m}^{k-1} (\omega_i - I_{i+1}) \Bigr)^{+} = \omega_k + \Bigl(\sup_{m:\, m\le k-1} \Bigl\{ G_m - G_k + \sum_{i=m}^{k-1} \omega_i \Bigr\} \Bigr)^{+} \\ &= \omega_k + (\widetilde{G}_{k-1} - G_k)^+ = \omega_k + \widetilde{G}_{k-1} \vee G_k - G_k = \widetilde{G}_k - G_k. \end{aligned}$$

Thus $J = S(I, \omega)$. Finally, from the first equation of (A-15) and $I_k = G_k - G_{k-1}$,

$$\begin{aligned} \tilde{I}_k &= \omega_k + (I_k - J_{k-1})^+ = I_k - J_{k-1} + \omega_k + (J_{k-1} - I_k)^+ = I_k + J_k - J_{k-1} \\ &= I_k + (\widetilde{G}_k - G_k) - (\widetilde{G}_{k-1} - G_{k-1}) = \widetilde{G}_k - \widetilde{G}_{k-1}. \end{aligned}$$

□

Here is a version for a random sequence.

Lemma A.5. *Let $\{\tilde{I}_k, J_k, I_k, \omega_k\}_{k\in\mathbb{Z}}$ be finite nonnegative random variables that satisfy assumptions* (i)–(iii) *below*:

(i) $\lim_{m\to-\infty} \sum_{i=m}^{0} (\omega_i - I_{i+1}) = -\infty$ *almost surely.*

(ii) $\{J_k, \omega_k\}_{k\in\mathbb{Z}}$ *is a stationary process.*

(iii) *Equations*

$$\tilde{I}_k = \omega_k + (I_k - J_{k-1})^+ \quad \textit{and} \quad J_k = \omega_k + (J_{k-1} - I_k)^+ \tag{A-20}$$

are valid for all $k \in \mathbb{Z}$, almost surely.

Then $J_k = \omega_k$ for infinitely many $k < 0$ with probability one, and $\tilde{I} = D(I, \omega)$ and $J = S(I, \omega)$ almost surely.

Proof. Lemma A.4 gives the conclusion once we verify that assumption (A-16) holds almost surely. Using the waiting time notation W from the previous proof, it suffices to show that

$$\mathbb{P}\{W_\ell = 0 \text{ for infinitely many } \ell < 0\} = 1 \tag{A-21}$$

The complementary event is $B = \{\text{there exists } m < 0 \text{ such that } W_k > 0 \text{ for all } k \le m\}$. B is a shift-invariant event. On the event B, the right-hand side of (A-17) is strictly positive for all $k \le m$ (for a random m). This implies, for all $k < m$,

$$0 < W_m = W_{m-1} + U_{m-1} = W_{m-2} + U_{m-2} + U_{m-1} = \cdots = W_k + \sum_{i=k}^{m-1} U_i = W_k + \sum_{i=k}^{m-1} (\omega_i - I_{i+1}).$$

By assumption (i) of the lemma, $W_k \to \infty$ a.s. on the event B as $k \to -\infty$. Let $c < \infty$. By the shift-invariance of B and the stationarity of the process $\{W_k = J_k - \omega_k\}_{k\in\mathbb{Z}}$,

$$\mathbb{P}(W_0 \ge c, B) = \mathbb{P}(W_k \ge c, B) \to \mathbb{P}(B) \quad \text{as } k \to -\infty.$$

We conclude that $W_0 = \infty$ a.s. on the event B, and hence $\mathbb{P}(B) = 0$. Claim (A-21) has been verified. □

Remark A.6 (nonstationary solution to Lindley's recursion). Some result such as Lemma A.5 is needed, for there can be another solution to Lindley's recursion that blows up as $n \to -\infty$. Suppose $\{U_k\}$ is ergodic and $\mathbb{E}U_k < 0$. Pick any random N such that $\sum_{k=m}^{N} U_k < 0$ for all $m \le N$. Set

$$\begin{aligned} W_n &= -\sum_{k=n}^{N} U_k && \text{for } n \le N, \\ W_{N+1} &= 0, \\ W_n &= (W_{n-1} + U_{n-1})^+ && \text{for } n \ge N+2. \end{aligned}$$

One can check that $W_n = (W_{n-1} + U_{n-1})^+$ holds for all $n \in \mathbb{Z}$.

Appendix B. Exponential distributions

The next lemma is elementary. The mapping $(I, J, W) \mapsto (I', J', W')$ in the lemma is an involution, that is, its own inverse.

Lemma B.1. *Let $\alpha, \beta > 0$. Assume given independent variables $W \sim \mathrm{Exp}(\alpha + \beta)$, $I \sim \mathrm{Exp}(\alpha)$, and $J \sim \mathrm{Exp}(\beta)$. Define*

$$\begin{aligned} I' &= W + (I - J)^+, \\ J' &= W + (I - J)^-, \\ W' &= I \wedge J. \end{aligned} \tag{B-1}$$

(i) *$I - J$ and $I \wedge J$ are independent.*

(ii) *$(I - J)^+ \sim \mathrm{Ber}(\beta/(\alpha + \beta)) \cdot \mathrm{Exp}(\alpha)$, that is, the product of a Bernoulli with success probability $\beta/(\alpha + \beta)$ and an independent rate α exponential.*

(iii) *The triple (I', J', W') has the same distribution as (I, J, W).*

We use the previous lemma to establish some facts about the queueing operators. To be consistent with the queueing discussion we parametrize exponentials with their means τ and ρ.

Lemma B.2. *Let $0 < \tau < \rho$. Let $(I_k)_{k \in \mathbb{Z}}$ and $(\omega_j)_{j \in \mathbb{Z}}$ be mutually independent random variables such that $I_k \sim \mathrm{Exp}(\rho^{-1})$ and $\omega_j \sim \mathrm{Exp}(\tau^{-1})$. Let $\tilde{I} = D(I, \omega)$ as defined by (2-18) and (2-19), $\widetilde{\omega} = R(I, \omega)$ as defined by (2-21), and $J_k = \widetilde{G}_k - G_k$ as in (2-20). Let $\Lambda_k = (\{\tilde{I}_j\}_{j \le k}, \ J_k, \ \{\widetilde{\omega}_j\}_{j \le k})$.*

(a) *$\{\Lambda_k\}_{k \in \mathbb{Z}}$ is a stationary, ergodic process. For each $k \in \mathbb{Z}$, the random variables $\{\tilde{I}_j\}_{j \le k}, \ J_k, \ \{\widetilde{\omega}_j\}_{j \le k}$ are mutually independent with marginal distributions*

$$\tilde{I}_j \sim \mathrm{Exp}(\rho^{-1}), \ \ \widetilde{\omega}_j \sim \mathrm{Exp}(\tau^{-1}) \text{ and } J_k \sim \mathrm{Exp}(\tau^{-1} - \rho^{-1}).$$

(b) *$\tilde{I}$ and $\widetilde{\omega}$ are independent sequences of i.i.d. variables.*

Proof. Part (b) follows from part (a) by dropping the J_k coordinate and letting $k \to \infty$. Stationarity and ergodicity of $\{\Lambda_k\}$ follow from its construction as a mapping applied to the independent i.i.d. sequences I and ω.

The distributional claims in part (a) are proved by coupling $(\tilde{I}_k, J_{k-1}, \widetilde{\omega}_k)_{k\in\mathbb{Z}}$ with another sequence whose distribution we know. Construct a process $(\widehat{I}_k, \widehat{J}_{k-1}, \widehat{\omega}_k)_{k\ge 1}$ as follows. First let $\widehat{J}_0$ be an $\mathrm{Exp}(\tau^{-1}-\rho^{-1})$ variable that is independent of (I, ω). Then for $k = 1, 2, 3, \ldots$, iterate the steps

$$\begin{aligned}\widehat{I}_k &= \omega_k + (I_k - \widehat{J}_{k-1})^+,\\ \widehat{J}_k &= \omega_k + (\widehat{J}_{k-1} - I_k)^+,\\ \widehat{\omega}_k &= I_k \wedge \widehat{J}_{k-1}.\end{aligned} \tag{B-2}$$

We prove the following claim by induction for each $m \ge 1$:

$$\begin{aligned}&\text{The variables } \widehat{I}_1, \ldots, \widehat{I}_m, \widehat{J}_m, \widehat{\omega}_1, \ldots, \widehat{\omega}_m \text{ are mutually independent,}\\ &\qquad\text{with marginal distributions } \widehat{I}_k \sim \mathrm{Exp}(\rho^{-1}),\ \widehat{J}_m \sim \mathrm{Exp}(\tau^{-1}-\rho^{-1}) \text{ and } \widehat{\omega}_j \sim \mathrm{Exp}(\tau^{-1}).\end{aligned} \tag{B-3}$$

By construction, the variables $(I_1, \widehat{J}_0, \omega_1)$ are independent with distributions

$$(\mathrm{Exp}(\rho^{-1}),\ \mathrm{Exp}(\tau^{-1}-\rho^{-1}),\ \mathrm{Exp}(\tau^{-1})).$$

The base case $m = 1$ of (B-3) comes by applying Lemma B.1 to the mapping (B-2) with $k = 1$. Now assume (B-3) holds for m. Then $(I_{m+1}, \widehat{J}_m, \omega_{m+1})$ are independent with distributions

$$(\mathrm{Exp}(\rho^{-1}),\ \mathrm{Exp}(\tau^{-1}-\rho^{-1}),\ \mathrm{Exp}(\tau^{-1}))$$

because, by construction, $\widehat{J}_m$ is a function of $(I_1, \ldots, I_m, \widehat{J}_0, \omega_1, \ldots, \omega_m)$ and thereby independent of (I_{m+1}, ω_{m+1}). By Lemma B.1, mapping (B-2) turns the triple $(I_{m+1}, \widehat{J}_m, \omega_{m+1})$ into the triple $(\widehat{I}_{m+1}, \widehat{J}_{m+1}, \widehat{\omega}_{m+1})$ of independent variables, which is also independent of $\widehat{I}_1, \ldots, \widehat{I}_m, \widehat{\omega}_1, \ldots, \widehat{\omega}_m$. Statement (B-3) has been extended to $m + 1$.

Our next claim is as follows:

$$\text{There exists (almost surely a random index) } m_0 \ge 0 \text{ such that } J_{m_0} = \widehat{J}_{m_0}. \tag{B-4}$$

Suppose first that $J_0 \ge \widehat{J}_0$. Then (2-24) and (B-2) imply that $J_k \ge \widehat{J}_k$ for all $k \ge 0$. If (B-4) fails then $J_k > \widehat{J}_k$ for all $k \ge 0$. But then for all $k > 0$,

$$J_k = J_{k-1} + \omega_k - I_k = \cdots = J_0 + \sum_{j=1}^{k} (\omega_j - I_j) \to -\infty \quad \text{almost surely, as } k \to \infty,$$

which contradicts the fact that $J_k \ge 0$ for all k. Thus in this case (B-4) happens. The case $J_0 \le \widehat{J}_0$ is symmetric.

Through equations (2-24) and (B-2), (B-4) implies that $\tilde{I}_k = \widehat{I}_k$, $J_k = \widehat{J}_k$, and $\widetilde{\omega}_k = \widehat{\omega}_k$ for all $k > m_0$. Part (a) follows from (B-3), because for any ℓ, $(\tilde{I}_{\ell-n}, \ldots, \tilde{I}_\ell, J_\ell, \widetilde{\omega}_{\ell-n}, \ldots, \widetilde{\omega}_\ell)$ has the same distribution as $(\tilde{I}_{k-n}, \ldots, \tilde{I}_k, J_k, \widetilde{\omega}_{k-n}, \ldots, \widetilde{\omega}_k)$ which agrees with $(\widehat{I}_{k-n}, \ldots, \widehat{I}_k, \widehat{J}_k, \widehat{\omega}_{k-n}, \ldots, \widehat{\omega}_k)$ with probability tending to one as $k \to \infty$. □

Next we compute a competition probability for two independent homogeneous Poisson processes on $[0, \infty)$ with rates α and β. Let $\{\sigma_i\}_{i\ge 1}$ be the jump times of the rate α Poisson process and $\{\tau_i\}_{i\ge 1}$ the jump times of the rate β Poisson process. For $n \geq 1$ define the events

$$\begin{aligned} A_n &= \{\sigma_i < \tau_i \ \text{ for all } i \in [n]\}, \\ B_n &= \{\sigma_i < \tau_i \ \text{ for all } i \in [n-1],\ \sigma_n > \tau_n\}. \end{aligned}$$

The *Catalan numbers* $\{C_n : n \geq 0\}$ are defined by

$$C_n = \frac{1}{n+1}\binom{2n}{n}. \tag{B-5}$$

The following properties of the Catalan triangle $\{C(n,k) : 0 \leq k \leq n\}$ given in (3-10) can be deduced with elementary arguments. $C(n,0) = 1$, $C(n,k) = \binom{n+k}{k} - \binom{n+k}{k-1}$ for $k > 0$, $C(n,k) = C(n,k-1) + C(n-1,k)$,

$$\sum_{k=0}^{i} C(n,k) = C(n+1,i) \qquad \text{for } 0 \leq i \leq n, \tag{B-6}$$

and

$$\sum_{k=0}^{n} C(n,k) = C(n+1,n) = C(n+1,n+1) = C_{n+1}. \tag{B-7}$$

Lemma B.3. *For $n \geq 1$,*

$$P(A_n) = \sum_{k=0}^{n-1} C(n-1,k)\, \frac{\alpha^n \beta^k}{(\alpha+\beta)^{n+k}}, \tag{B-8}$$

$$P(B_n) = C_{n-1}\, \frac{\alpha^{n-1}\beta^n}{(\alpha+\beta)^{2n-1}}. \tag{B-9}$$

Remark B.4. The generating function of the Catalan numbers is

$$f(x) = \sum_{n \geq 0} C_n x^n = \frac{1 - \sqrt{1-4x}}{2x} \qquad \text{for } |x| \leq \tfrac{1}{4}.$$

Hence from (B-9),

$$\sum_{n=1}^{\infty} P(B_n) = \frac{\beta}{\alpha+\beta}\, f\!\left(\frac{\alpha\beta}{(\alpha+\beta)^2}\right) = \begin{cases} 1 & \text{if } \beta \geq \alpha, \\ \frac{\beta}{\alpha} & \text{if } \beta < \alpha. \end{cases}$$

In other words, the rate α process stays forever ahead of the rate β process with probability $(1-\beta/\alpha)^+$.

Proof. We compute $P(B_n)$ first and then obtain $P(A_n)$ by inclusion-exclusion.

Since $C_0 = 1$, (B-9) holds for $n = 1$. For $n \geq 2$ condition on (σ_n, τ_n):

$$P(B_n) = \int_{a>b>0} P_{(a,b)}\big\{U_i \leq V_i \text{ for } i \in [n-1]\big\}\, P((\sigma_n, \tau_n) \in d(a,b)), \tag{B-10}$$

where under $P_{(a,b)}$, $0 < U_1 < \cdots < U_{n-1}$ are the order statistics of $n-1$ i.i.d. uniform random variables on $[0,a]$ and $0 < V_1 < \cdots < V_{n-1}$ are the same on $[0,b]$, independent of the $\{U_i\}$. We calculate the probability inside the integral.

Below, first use the equal probability of the permutations of $\{x_i\}$ among themselves and $\{y_j\}$ among themselves. Note that $a > b$ and the conditions $x_i < y_i$ force all $\{x_i, y_j\}$ to lie in $[0, b]$. Then use the equal probability of all permutations of $\{x_i, y_j\}$ together. The Catalan number C_k is the number of permutations of $\{x_1, \dots, x_k, y_1, \dots, y_k\}$ such that $x_1 < \cdots < x_k,\ y_1 < \cdots < y_k$ and $x_i < y_i$ for all i (see Corollary 6.2.3 and item **dd** on page 223 of [Stanley 1999]).

$$P_{(a,b)}(U_i \le V_i \text{ for } i \in [n-1]) = \frac{((n-1)!)^2}{(ab)^{n-1}} \int\limits_{\substack{x_1<\cdots<x_{n-1}<a\\ y_1<\cdots<y_{n-1}<b}} \mathbf{1}_{x_i<y_i\ \forall i\in[n-1]}\, d\boldsymbol{x}\, d\boldsymbol{y}$$
$$= C_{n-1}\frac{((n-1)!)^2}{(ab)^{n-1}} \int\limits_{x_1<\cdots<x_{n-1}<y_1<\cdots<y_{n-1}<b} d\boldsymbol{x}\, d\boldsymbol{y} = C_{n-1}\frac{((n-1)!)^2}{(ab)^{n-1}} \cdot \frac{b^{2(n-1)}}{(2n-2)!}.$$

Substitute this back into (B-10). Use the gamma densities of σ_n and τ_n.

$$P(B_n) = C_{n-1}\frac{((n-1)!)^2}{(2n-2)!} \int_{0<b<a<\infty} \frac{b^{n-1}}{a^{n-1}} \cdot \frac{(\alpha a)^{n-1}}{\Gamma(n)} \alpha e^{-\alpha a} \cdot \frac{(\beta b)^{n-1}}{\Gamma(n)} \beta e^{-\beta b}\, da\, db$$
$$= C_{n-1}\frac{\alpha^n \beta^n}{(2n-2)!} \int_{0<b<a<\infty} b^{2n-2} e^{-\alpha a - \beta b}\, da\, db \;=\; C_{n-1}\,\frac{\alpha^{n-1}\,\beta^n}{(\beta+\alpha)^{2n-1}}.$$

We prove (B-8). The case $n = 1$ is elementary. Let $n \ge 2$ and assume (B-8) for $n-1$. Abbreviate $p = \beta/(\alpha+\beta)$ and $q = \alpha/(\alpha+\beta)$. Use (B-6) and (B-7) below.

$$P(A_n) = P(A_{n-1}) - P(B_n) = q^{n-1}\sum_{k=0}^{n-2} C(n-2,k)\, p^k - C_{n-1} q^{n-1}\, p^n$$
$$= q^{n-1}\sum_{k=0}^{n-2} C(n-2,k)\,(p^k - p^n) = q^n \sum_{k=0}^{n-2}\sum_{j=k}^{n-1} C(n-2,k)\, p^j$$
$$= q^n \sum_{j=0}^{n-1}\sum_{k=0}^{j\wedge(n-2)} C(n-2,k)\, p^i = q^n \sum_{j=0}^{n-1} C(n-1, j\wedge(n-2))p^j = q^n \sum_{j=0}^{n-1} C(n-1,j)p^j. \quad \square$$

References

[Anantharam 1993] V. Anantharam, "Uniqueness of stationary ergodic fixed point for a $\cdot/M/K$ node", *Ann. Appl. Probab.* **3**:1 (1993), 154–172. MR

[Baccelli et al. 2000] F. Baccelli, A. Borovkov, and J. Mairesse, "Asymptotic results on infinite tandem queueing networks", *Probab. Theory Related Fields* **118**:3 (2000), 365–405. MR

[Baik et al. 1999] J. Baik, P. Deift, and K. Johansson, "On the distribution of the length of the longest increasing subsequence of random permutations", *J. Amer. Math. Soc.* **12**:4 (1999), 1119–1178. MR

[Bakhtin et al. 2014] Y. Bakhtin, E. Cator, and K. Khanin, "Space-time stationary solutions for the Burgers equation", *J. Amer. Math. Soc.* **27**:1 (2014), 193–238. MR

[Balázs and Seppäläinen 2010] M. Balázs and T. Seppäläinen, "Order of current variance and diffusivity in the asymmetric simple exclusion process", *Ann. of Math.* (2) **171**:2 (2010), 1237–1265. MR

[Balázs et al. 2006] M. Balázs, E. Cator, and T. Seppäläinen, "Cube root fluctuations for the corner growth model associated to the exclusion process", *Electron. J. Probab.* **11** (2006), [paper no. 42]. MR

[Balázs et al. 2011] M. Balázs, J. Quastel, and T. Seppäläinen, "Fluctuation exponent of the KPZ/stochastic Burgers equation", *J. Amer. Math. Soc.* **24**:3 (2011), 683–708. MR

[Balázs et al. 2012] M. Balázs, J. Komjáthy, and T. Seppäläinen, "Microscopic concavity and fluctuation bounds in a class of deposition processes", *Ann. Inst. Henri Poincaré Probab. Stat.* **48**:1 (2012), 151–187. MR

[Balázs et al. 2019a] M. Balázs, O. Busani, and T. Seppäläinen, "Non-existence of bi-infinite geodesics in the exponential corner growth model", preprint, 2019. arXiv

[Balázs et al. 2019b] M. Balázs, F. Rassoul-Agha, and T. Seppäläinen, "Large deviations and wandering exponent for random walk in a dynamic beta environment", *Ann. Probab.* **47**:4 (2019), 2186–2229. MR

[Basu et al. 2019] R. Basu, S. Ganguly, and A. Hammond, "Fractal geometry of Airy_2 processes coupled via the Airy sheet", preprint, 2019. arXiv

[Cator and Groeneboom 2006] E. Cator and P. Groeneboom, "Second class particles and cube root asymptotics for Hammersley's process", *Ann. Probab.* **34**:4 (2006), 1273–1295. MR

[Cator and Pimentel 2012] E. Cator and L. P. R. Pimentel, "Busemann functions and equilibrium measures in last passage percolation models", *Probab. Theory Related Fields* **154**:1-2 (2012), 89–125. MR

[Cator and Pimentel 2013] E. Cator and L. P. R. Pimentel, "Busemann functions and the speed of a second class particle in the rarefaction fan", *Ann. Probab.* **41**:4 (2013), 2401–2425. MR

[Chang 1994] C. S. Chang, "On the input-output map of a $G/G/1$ queue", *J. Appl. Probab.* **31**:4 (1994), 1128–1133. MR

[Corwin 2012] I. Corwin, "The Kardar–Parisi–Zhang equation and universality class", *Random Matrices Theory Appl.* **1**:1 (2012), 1130001, 76. MR

[Corwin 2016] I. Corwin, "Kardar–Parisi–Zhang universality", *Notices Amer. Math. Soc.* **63**:3 (2016), 230–239. MR

[Corwin 2018] I. Corwin, "Commentary on "Longest increasing subsequences: from patience sorting to the Baik–Deift–Johansson theorem" by David Aldous and Persi Diaconis", *Bull. Amer. Math. Soc.* (*N.S.*) **55**:3 (2018), 363–374. MR

[Coupier 2011] D. Coupier, "Multiple geodesics with the same direction", *Electron. Commun. Probab.* **16** (2011), 517–527. MR

[Dauvergne et al. 2018] D. Dauvergne, J. Ortmann, and B. Virág, "The directed landscape", preprint, 2018. arXiv

[Draief et al. 2005] M. Draief, J. Mairesse, and N. O'Connell, "Queues, stores, and tableaux", *J. Appl. Probab.* **42**:4 (2005), 1145–1167. MR

[Ferrari and Kipnis 1995] P. A. Ferrari and C. Kipnis, "Second class particles in the rarefaction fan", *Ann. Inst. H. Poincaré Probab. Statist.* **31**:1 (1995), 143–154. MR

[Ferrari and Martin 2006] P. A. Ferrari and J. B. Martin, "Multi-class processes, dual points and $M/M/1$ queues", *Markov Process. Related Fields* **12**:2 (2006), 175–201. MR

[Ferrari and Martin 2007] P. A. Ferrari and J. B. Martin, "Stationary distributions of multi-type totally asymmetric exclusion processes", *Ann. Probab.* **35**:3 (2007), 807–832. MR

[Ferrari and Martin 2009] P. A. Ferrari and J. B. Martin, "Multiclass Hammersley–Aldous–Diaconis process and multiclass-customer queues", *Ann. Inst. Henri Poincaré Probab. Stat.* **45**:1 (2009), 250–265. MR

[Ferrari and Pimentel 2005] P. A. Ferrari and L. P. R. Pimentel, "Competition interfaces and second class particles", *Ann. Probab.* **33**:4 (2005), 1235–1254. MR

[Ferrari et al. 2009] P. A. Ferrari, J. B. Martin, and L. P. R. Pimentel, "A phase transition for competition interfaces", *Ann. Appl. Probab.* **19**:1 (2009), 281–317. MR

[Georgiou et al. 2015] N. Georgiou, F. Rassoul-Agha, T. Seppäläinen, and A. Yilmaz, "Ratios of partition functions for the log-gamma polymer", *Ann. Probab.* **43**:5 (2015), 2282–2331. MR

[Georgiou et al. 2017a] N. Georgiou, F. Rassoul-Agha, and T. Seppäläinen, "Geodesics and the competition interface for the corner growth model", *Probab. Theory Related Fields* **169**:1-2 (2017), 223–255. MR

[Georgiou et al. 2017b] N. Georgiou, F. Rassoul-Agha, and T. Seppäläinen, "Stationary cocycles and Busemann functions for the corner growth model", *Probab. Theory Related Fields* **169**:1-2 (2017), 177–222. MR

[Glynn and Whitt 1991] P. W. Glynn and W. Whitt, "Departures from many queues in series", *Ann. Appl. Probab.* **1**:4 (1991), 546–572. MR

[Gray 2009] R. M. Gray, *Probability, random processes, and ergodic properties*, 2nd ed., Springer, 2009. MR

[Hoffman 2005] C. Hoffman, "Coexistence for Richardson type competing spatial growth models", *Ann. Appl. Probab.* **15**:1B (2005), 739–747. MR

[Hoffman 2008] C. Hoffman, "Geodesics in first passage percolation", *Ann. Appl. Probab.* **18**:5 (2008), 1944–1969. MR

[Howard and Newman 2001] C. D. Howard and C. M. Newman, "Geodesics and spanning trees for Euclidean first-passage percolation", *Ann. Probab.* **29**:2 (2001), 577–623. MR

[Janjigian et al. 2019] C. Janjigian, F. Rassoul-Agha, and T. Seppäläinen, "Geometry of geodesics through Busemann measures in directed last-passage percolation", preprint, 2019. arXiv

[Johansson 2000] K. Johansson, "Shape fluctuations and random matrices", *Comm. Math. Phys.* **209**:2 (2000), 437–476. MR

[Licea and Newman 1996] C. Licea and C. M. Newman, "Geodesics in two-dimensional first-passage percolation", *Ann. Probab.* **24**:1 (1996), 399–410. MR

[Lindley 1952] D. V. Lindley, "The theory of queues with a single server", *Proc. Cambridge Philos. Soc.* **48** (1952), 277–289. MR

[Loynes 1962] R. M. Loynes, "The stability of a queue with non-independent interarrival and service times", *Proc. Cambridge Philos. Soc.* **58** (1962), 497–520. MR

[Mairesse and Prabhakar 2003] J. Mairesse and B. Prabhakar, "The existence of fixed points for the $\cdot/GI/1$ queue", *Ann. Probab.* **31**:4 (2003), 2216–2236. MR

[Martin 2004] J. B. Martin, "Limiting shape for directed percolation models", *Ann. Probab.* **32**:4 (2004), 2908–2937. MR

[Mountford and Prabhakar 1995] T. Mountford and B. Prabhakar, "On the weak convergence of departures from an infinite series of $\cdot/M/1$ queues", *Ann. Appl. Probab.* **5**:1 (1995), 121–127. MR

[Newman 1995] C. M. Newman, "A surface view of first-passage percolation", pp. 1017–1023 in *Proc. Int. Congress of Math., II* (Zürich, 1994), edited by S. D. Chatterji, Birkhäuser, Basel, 1995. MR

[Pimentel 2016] L. P. R. Pimentel, "Duality between coalescence times and exit points in last-passage percolation models", *Ann. Probab.* **44**:5 (2016), 3187–3206. MR

[Pimentel 2018] L. P. R. Pimentel, "Local behaviour of airy processes", *J. Stat. Phys.* **173**:6 (2018), 1614–1638. MR

[Prabhakar 2003] B. Prabhakar, "The attractiveness of the fixed points of a $\cdot/GI/1$ queue", *Ann. Probab.* **31**:4 (2003), 2237–2269. MR

[Rost 1981] H. Rost, "Nonequilibrium behaviour of a many particle process: density profile and local equilibria", *Z. Wahrsch. Verw. Gebiete* **58**:1 (1981), 41–53. MR

[Seppäläinen 2012] T. Seppäläinen, "Scaling for a one-dimensional directed polymer with boundary conditions", *Ann. Probab.* **40**:1 (2012), 19–73. Corrected version available at http://arxiv.org/abs/0911.2446. MR

[Seppäläinen 2018] T. Seppäläinen, "The corner growth model with exponential weights", pp. 133–201 in *Random growth models*, Proc. Sympos. Appl. Math. **75**, Amer. Math. Soc., Providence, RI, 2018. MR

[Seppäläinen 2020] T. Seppäläinen, "Existence, uniqueness and coalescence of directed planar geodesics: proof via the increment-stationary growth process", *Ann. Inst. Henri Poincaré Probab. Stat.* **56**:3 (2020), 1775–1791. MR

[Stanley 1999] R. P. Stanley, *Enumerative combinatorics, II*, Cambridge Studies in Advanced Mathematics **62**, Cambridge University Press, 1999. MR

Received 30 Sep 2019. Revised 21 Apr 2020. Accepted 11 May 2020.

WAI-TONG LOUIS FAN: waifan@iu.edu
Mathematics, Indiana University, Bloomington, IN, United States

TIMO SEPPÄLÄINEN: seppalai@math.wisc.edu
Department of Mathematics, University of Wisconsin-Madison, United States

PROBABILITY and MATHEMATICAL PHYSICS

Vol. 1, No. 1, 2020
https://doi.org/10.2140/pmp.2020.1.101

OPTIMAL LOWER BOUND ON THE LEAST SINGULAR VALUE OF THE SHIFTED GINIBRE ENSEMBLE

GIORGIO CIPOLLONI, LÁSZLÓ ERDŐS AND DOMINIK SCHRÖDER

We consider the least singular value of a large random matrix with real or complex i.i.d. Gaussian entries shifted by a constant $z \in \mathbb{C}$. We prove an optimal lower tail estimate on this singular value in the critical regime where z is around the spectral edge, thus improving the classical bound of Sankar, Spielman and Teng (*SIAM J. Matrix Anal. Appl.* **28**:2 (2006), 446–476) for the particular shift-perturbation in the edge regime. Lacking Brézin–Hikami formulas in the real case, we rely on the superbosonization formula (*Comm. Math. Phys.* **283**:2 (2008), 343–395).

1. Introduction

The effective numerical solvability of a large system of linear equations $Ax = b$ is determined by the condition number of the matrix A. In many practical applications the norm of A is bounded and thus the condition number critically depends on the smallest singular value $\sigma_1(A)$ of A. When the matrix elements of A come from noisy measured data, then the lower tail probability of $\sigma_1(A)$ tends to exhibit a universal scaling behavior, depending on the variance of the noise. In the simplest case A can be decomposed as

$$A = A_0 + X, \tag{1}$$

where A_0 is a deterministic square matrix and X is drawn from the *Ginibre ensemble*, i.e., X has i.i.d. centered Gaussian matrix elements with variance $\boldsymbol{E}|x_{ij}|^2 = N^{-1}$, where N is the dimension.

The randomness in X smoothens out possible singular behavior of A^{-1}. In particular Sankar, Spielman and Teng [Sankar et al. 2006] showed that the smallest singular value $\sigma_1(A)$ lives on a scale not smaller than N^{-1}, equivalently, the smallest eigenvalue $\lambda_1(AA^*)$ of AA^* lives on a scale $\le N^{-2}$, i.e.,

$$\boldsymbol{P}\big(\lambda_1(AA^*) = [\sigma_1(A)]^2 \le xN^{-2}\big) \lesssim \sqrt{x}, \quad \text{for any} \quad x > 0, \tag{2}$$

up to logarithmic corrections, uniformly in A_0. If X is a complex Ginibre matrix, then the $\sqrt{x}$ bound improves to x.

The authors were partially supported by ERC Advanced Grant No. 338804. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 665385. Schröder is currently affiliated with the ETH Institute for Theoretical Studies.
MSC2010: 15B52, 60B20.
Keywords: supersymmetric formalism, superbosonization, circular law.

The special case $A_0 = 0$ shows that the bound (2) is essentially optimal. Indeed, the tail probability of $\lambda_1(XX^*)$ of real and complex Ginibre ensembles has been explicitly computed by Edelman [1988] as

$$\lim_{N\to\infty} \boldsymbol{P}(\lambda_1(XX^*) \le xN^{-2}) = \begin{cases} 1-e^{-x/2-\sqrt{x}} = \sqrt{x} + \mathcal{O}(x) & \text{in the real case,} \\ 1-e^{-x} = x + \mathcal{O}(x^2) & \text{in the complex case.} \end{cases} \tag{3}$$

The complex Ginibre ensemble has a stronger smoothing effect in (3) due to the additional degrees of freedom. This observation is analogous to the different strength of the level repulsion in real symmetric and complex Hermitian random matrices.

The support of the spectrum of such *information plus noise matrices* AA^* becomes deterministic as $N \to \infty$ and it can be computed from the solution of a certain self-consistent equation [Dozier and Silverstein 2007]. Almost surely no eigenvalues lie outside the support of the limiting measure [Bai and Silverstein 2012]. Thus $\lambda_1(AA^*)$ has a simple N-independent positive lower bound if 0 is away from this support. However, when 0 is well inside the limiting spectrum, the smoothing mechanism becomes important yielding that $\lambda_1(AA^*)$ is of order N^{-2} with a lower tail given in (2). The regime where 0 is near the edge of this support is yet unexplored.

The goal of this paper is to study this transitional regime for $A = X - z$, i.e., for the important special case where $A_0 = -z\mathbb{1}$ is a constant multiple of the identity matrix, as the spectral parameter $z \in \mathbb{C}$ is varied. The limiting density of states of $Y^z := (X-z)(X-z)^*$ is supported in the interval $[0, \mathfrak{e}_+]$ for $|z| \le 1$ and the interval $[\mathfrak{e}_-, \mathfrak{e}_+]$ with $\mathfrak{e}_- > 0$ for $|z| > 1$, where $\mathfrak{e}_\pm$ are explicit functions of $|z|$ given in (18a). As noted above, the problem is relatively simple if $|z| \ge 1+\epsilon$ with some N-independent ϵ as in this case [Bai and Silverstein 2012] implies that almost surely $\lambda_1(Y^z) \ge C(\epsilon) > 0$ is bounded away from zero. In the opposite regime, when $|z| \le 1-\epsilon$, then typically $\lambda_1(Y^z) \sim N^{-2}$, and in fact (2) provides the correct corresponding upper bound (modulo logs).

Our main result on the tail probability of $\lambda_1(Y^z)$ is that for $|z| \le 1 + CN^{-1/2}$

$$\boldsymbol{P}\big(\lambda_1(Y^z) \le x \cdot c(N,z)\big) \lesssim \begin{cases} x + \sqrt{x}e^{-\frac{1}{2}N(\Im z)^2} & \text{in the real case,} \\ x & \text{in the complex case,} \end{cases} \tag{4a}$$

where

$$c(N,z) := \min\left\{\frac{1}{N^{3/2}}, \frac{1}{N^2|1-|z|^2|}\right\}. \tag{5}$$

Our bound is sharp up to logarithmic corrections; see Corollary 2.4 for the precise statement. Notice the transition between the x and $\sqrt{x}$ behavior in the real case of (4a): near the real axis, $|\Im z| \ll N^{-1/2}$, the result is analogous to the real case (3) at $z = 0$, otherwise the complex behavior (3) dominates at the edges even for real X; see Figure 1. These results reveal how the robust bound (2) improves near the spectral edge in the transition regime $-CN^{-1/2} \le 1 - |z| \ll 1$ in both symmetry classes. The transition to the Tracy–Widom scaling in the regime well outside of the spectrum $|z| - 1 \gg N^{-1/2}$ is deferred to our future work.

One motivation for studying $X - z$ is the classical ODE model $\mathrm{d}u/\mathrm{d}t = (X - z)u$ on the stability of large biological networks by May [1972]. For example, the matrix elements x_{ij} may express random

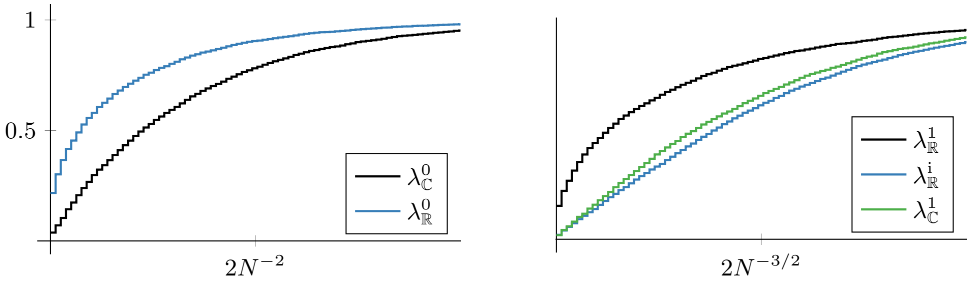


Figure 1. Plots of the cumulative histograms of the smallest eigenvalue $\lambda^z_{\mathbb{R},\mathbb{C}}$ of the matrix $(X-z)(X-z)^*$, where $\mathbb{R},\mathbb{C}$ indicates whether X is distributed according to the real or complex Ginibre ensemble. The data was generated by sampling 5000 matrices of size 200×200. The first plot confirms the difference between the x- and $\sqrt{x}$-scaling close to 0; see (3). The second plot shows that this difference is also observable for shifted Ginibre matrices at the edge $|z|=1$, but only for real spectral parameters $z=\pm 1$. When the complex parameter z is away from the real axis, then the real case behaves similarly to the complex case.

connectivity rates between neurons and z is the overall decay rate of neuron activation [Sompolinsky et al. 1988]. As $\Re z$ crosses 1, there is a fine phase transition in the large time behavior of u that depends on whether X is real or complex Ginibre matrix; see [Chalker and Mehlig 1998] and [Erdős et al. 2018] for the recent mathematical results, as well as for further references. Another important motivation is that an effective lower tail bound on the least singular value of $X-z$ is essential for the proof of the circular law via Girko's formula; see [Bordenave and Chafaï 2012] for a detailed survey. In fact, this is the most delicate ingredient in any proof concerning eigenvalue distribution of large non-Hermitian matrices. In particular, relying on the main result of the current paper, we proved [Cipolloni et al. 2019b] that the local eigenvalue statistics for random matrices with centered i.i.d. entries near the spectral edge asymptotically coincide with those for the corresponding Ginibre ensemble as $N\to\infty$. This is the non-Hermitian analogue of the celebrated Tracy–Widom edge universality for Wigner matrices [Soshnikov 1999; Bourgade et al. 2014]. Similarly, the singular value bound from the present paper is also an important ingredient for the recent CLTs for complex and real i.i.d. matrices [Cipolloni et al. 2019a; 2020].

We now give a brief history of related results. In the $z=0$ case, tail estimates for $\lambda_1(XX^*)$ beyond the Gaussian distribution have been the subject of intensive research [Rudelson 2008; Tao and Vu 2009] eventually obtaining (3) with an additive $\mathcal{O}(e^{-cN})$ error term for any X with i.i.d. entries with subgaussian tails in [Rudelson and Vershynin 2008]. The precise distribution of $\lambda_1(XX^*)$ was shown in [Tao and Vu 2010a] to coincide with the Gaussian case (3) under a bounded high moment condition and with an $\mathcal{O}(N^{-c})$ error term; see also [Che and Lopatto 2019a; 2019b] for more general ensembles. In the case of general A_0, lower bounds on $\lambda_1(AA^*)$ in the non-Gaussian setting have been obtained in [Tao and Vu 2007; 2010b], albeit not uniformly in A_0; see also [Cook 2018; Tikhomirov 2020] beyond the i.i.d. case. We are not aware of any previous results improving (2) in the transitional regime (4a).

Since we consider Ginibre (i.e., purely Gaussian) ensembles, one might think that everything is explicitly computable from the well understood spectrum of X. The eigenvalue density of X converges to the uniform distribution on the unit disk and the spectral radius of X converges to 1 (these results have also been established for the general non-Gaussian case; cf. Girko's circular law [Girko 1984; Bai 1997; Tao and Vu 2008; Geman 1986; Bai and Yin 1986; Bordenave et al. 2018]). Also the joint probability density function of all Ginibre eigenvalues, as well as their local correlation functions are explicitly known; see [Ginibre 1965] and [Mehta 1967] for the relatively simple complex case, and [Lehmann and Sommers 1991; Edelman 1997; Borodin and Sinclair 2009; Forrester and Nagao 2007] for the more involved real case, where the appearance of $\sim N^{1/2}$ real eigenvalues causes a singularity in the local density. However, eigenvalues of X give no direct information on the singular values of $X - z$ and the extensive literature on the Ginibre spectrum is not applicable. Notice that any intuition based upon the eigenvalues of X is misleading: the nearest eigenvalue to z is at a distance of order $N^{-1/2}$ for any $|z| \le 1$. However, $\|(X-z)\|^{-1} \sim \max\{N^{3/4}, N|1-|z|^2|^{1/2}\}$ for $|z| \le 1 + CN^{-1/2}$, as a consequence of our result (4a). This is an indication that typically X is highly nonnormal (another indication is that the largest singular value of X is 2, while its spectral radius is only 1).

Regarding our strategy, in this paper we use supersymmetric methods to express the resolvent of Y^z. In particular, we use a multiple Grassmann integral formula for

$$\varrho_N^z(E) := \frac{1}{N\pi} \Im \boldsymbol{E} \operatorname{Tr} \frac{1}{Y^z - E + \mathrm{i}0}, \tag{6}$$

the averaged density of states (or one-point function) of Y^z at energy $E \in \mathbb{R}$. For $|E| \lesssim c(N, z)$ a sizeable contribution to (6) comes from the lowest eigenvalue $\lambda_1(Y^z)$, hence a good upper estimate on (6) translates into a lower tail bound on $\lambda_1(Y^z)$.

With the help of the superbosonization formula by Littelmann, Sommers and Zirnbauer [Littelmann et al. 2008], we can drastically reduce the number of integration variables: instead of N bosonic and N fermionic variables we will have an explicit expression for (6) involving merely two contour integration variables in complex case and three in the real case. The remaining integrals are still highly oscillatory, but contour deformation allows us to estimate them optimally. In fact, saddle point analysis identifies the leading term as long as $|E| \gg c(N, z)$. However, in the critical regime, $|E| \lesssim c(N, z)$, the saddle point analysis breaks down. The leading term is extracted as a specific rescaling of a universal function given by a double integral. We work out the precise answer for (6) in the complex case and we provide optimal bounds in the real case, deferring the precise asymptotics to further work.

Lower tail estimates require delicate knowledge about individual eigenvalues, i.e., about the density of states below the scale of eigenvalue spacing, and it is crucial to exploit the Gaussianity of X via explicit formulas. There are essentially three methods: (i) orthogonal polynomials, (ii) Brézin–Hikami contour integration formula [1998] and (iii) supersymmetric formalism. We are not aware of any orthogonal polynomial approach to analyze $Y^z = (X - z)(X - z)^*$ in the real case (see [Desrosiers and Forrester 2006] in the complex case and [Mo 2012] for rank-1 perturbation of real X). In the complex case, the ensemble Y^z has also been extensively investigated by the Brézin–Hikami formula in [Ben Arous and Péché 2005], where even the determinantal correlation kernel was computed as a double integral involving the Bessel kernel; see also [Guhr and Wettig 1996; Jackson et al. 1996] for a derivation via the supersymmetric version of the Itzykson–Zuber formula. Although the paper [Ben Arous and Péché 2005] did not analyze the resulting one point function, well known asymptotics for the Bessel function may be used to rederive our bounds and asymptotics on (6), as well as (4a), from [Ben Arous and Péché 2005, Theorem 7.1]; see Appendix C for more details. For the real case, however, there is no analogue of the Brézin–Hikami formula.

Therefore, in this paper we explore the last option, the supersymmetric approach, that is available for both symmetry classes, although the real case is considerably more involved. Our main tool is the powerful superbosonization formula [Littelmann et al. 2008] followed by a delicate multivariable contour integral analysis. We remark that, alternatively, one may also use the Hubbard–Stratonovich transformation; see, e.g., [Afanasiev 2019, Proposition 1] where correlation functions, i.e., expectations of *products* of characteristic polynomials of X were computed in this way. Note, however, that the density of states (6) requires one to analyze *ratios* of determinants, a technically much more demanding task. While explicit formulas can be obtained with both methods (see [Recher et al. 2012] and especially [Kieburg et al. 2009] for an explicit comparison), the subsequent analysis seems to be more feasible with the formula obtained from the superbosonization approach, as our work demonstrates.

Supersymmetry is a compelling method originated in physics [Efetov 1997; Guhr 2011; Verbaarschot et al. 1985] to produce surprising identities related to random matrices, whose potential has not yet been fully exploited in mathematics. It has been especially successful in deriving rigorous results on Gaussian random band matrices [Bao and Erdős 2017; Disertori et al. 2002; 2017; 2018; Shcherbina 2014; 2020; Shcherbina and Shcherbina 2016; 2017; 2019], sometimes even beyond the Gaussian case [Shcherbina 2011; 2013; Spencer 2011], as well as on overlaps of non-Hermitian Ginibre eigenvectors [Fyodorov 2018]. We also mention the recent results in [Fyodorov 2018] and [Fyodorov et al. 2018] as examples of a remarkable interplay between supersymmetric and orthogonal polynomial techniques in the theory of Ginibre and related matrices.

The main object of our work, the Hermitian block random matrix

$$H = H^z := \begin{pmatrix} 0 & X - z \\ X^* - \bar{z} & 0 \end{pmatrix} \tag{7}$$

arose in the physics literature as a chiral random matrix model for massless Dirac operator, introduced by Stephanov [1996]. Typically, instead of z and $\bar{z}$, both shift parameters are chosen equal to z (interpreted as i times the chemical potential) so that the corresponding H is not self-adjoint; this model has been extensively investigated by both supersymmetric and orthogonal polynomial techniques; see, e.g., [Verbaarschot and Zahed 1993; Akemann et al. 2005; Wilke et al. 1998; Guhr and Wettig 1997; Osborn et al. 1999]. However, in the special case when z is real, our H^z as given in (7) coincides with Stephanov's model where z can be interpreted as temperature (or Matsubara frequency); see [Verbaarschot and Wettig 2000, Section 6.1].

2. Model and main results

We consider the model $Y = Y^z = (X - z)(X^* - \overline{z})$ with a fixed complex parameter $z \in \mathbb{C}$ and with a random matrix $X \in \mathbb{C}^{N \times N}$ having independent real or complex Gaussian entries $x_{ab} \sim \mathcal{N}(0, N^{-1})$, where in the complex case we additionally assume $\boldsymbol{E}\, x_{ab}^2 = 0$. Note that Y is related to the block matrix (7) through its resolvent via

$$\frac{\operatorname{Tr}(H - \sqrt{w})^{-1}}{2\sqrt{w}} = \operatorname{Tr}(Y - w)^{-1}, \qquad \Re w > 0,\ \Im w > 0, \tag{8}$$

where the branch of $\sqrt{w}$ is chosen such that $\Im\sqrt{w} > 0$. It is well known that in the large N limit the normalized trace of the resolvent of many random matrix ensembles becomes deterministic and it satisfies an algebraic equation, the matrix Dyson equation (MDE) [Ajanki et al. 2019]. In the current case of i.i.d. entries, the MDE reduces to a simple cubic scalar equation

$$\frac{1}{m_{H^z}} + (w + m_{H^z}) - \frac{|z|^2}{w + m_{H^z}} = 0, \qquad \Im m_{H^z}(w) > 0, \quad \Im w > 0, \tag{9}$$

that has a unique solution, denoted by m_{H^z}. The local law from [Alt et al. 2019] asserts that

$$\frac{1}{2N}\operatorname{Tr}(H^z - w)^{-1} = m_{H^z}(w) + \mathcal{O}_\prec((N\Im w)^{-1}), \tag{10}$$

where $\mathcal{O}_\prec$ denotes a suitable concept of high-probability error term. Together with (8) it follows that the normalized trace of the resolvent $(Y^z - w)^{-1}$ of Y^z is well approximated as

$$\frac{1}{N}\operatorname{Tr}(Y^z - w)^{-1} \approx m^z(w)$$

by the unique solution $m = m^z = m_{Y^z}$ to the equation

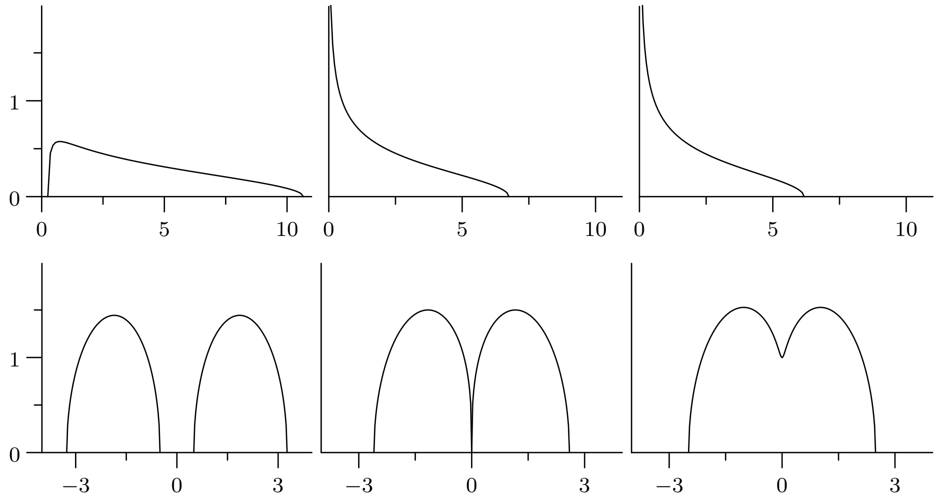
$$\frac{1}{m^z} + w(1 + m^z) - \frac{|z|^2}{1 + m^z} = 0, \qquad \Im m^z(w) > 0, \quad \Re w > 0,\ \Im w > 0, \tag{11}$$

which is given by $m^z(w) = m_{H^z}(\sqrt{w})/\sqrt{w}$. Since m approximates the trace of the resolvent, the density of states is obtained as the imaginary part of the continuous extension of m to the real line, i.e., $\varrho_\#(E) = \pi^{-1}\lim_{\epsilon\to 0+}\Im m_\#(E + \mathrm{i}\epsilon)$ for both choices $\# = H^z, Y^z$. For $\delta := 1 - |z|^2 \approx 0$ the Stieltjes transform m_{H^z} and its density of states exhibit a cusp formation at $w = 0$ as δ crosses the value 0. This cusp formation in H^z implies an analogous transition for m^z; the corresponding density of states are depicted in Figure 2.

Complex case. The main result of this paper in the complex case is an asymptotic double-integral formula for $\boldsymbol{E}\operatorname{Tr}(Y - w)^{-1}$ at $w = E + \mathrm{i}0$, $E \geq 0$. In the transitional regime it is convenient to introduce the rescaled variables

$$\lambda := E/c(N), \qquad \tilde{\delta} := N^{1/2}\delta, \qquad \text{where } \delta := 1 - |z|^2, \tag{12}$$

recalling that $c(N) = c(N, z)$ was defined in (5). For $r \geq 0$, let $\Psi = \Psi(r)$ be the unique solution to the cubic equation $1 + r\Psi + \Psi^3 = 0$ with $\Re\Psi, \Im\Psi > 0$. It is easy to see that $\Psi(r)$ satisfies $\Psi(0) = e^{\mathrm{i}\pi/3}$ and $\Psi(r) \sim \mathrm{i}\sqrt{r}$ for $r \gg 1$. We also introduce the notation $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$ for real numbers a, b.



(a) $\delta = -2$ (b) $\delta = 0$ (c) $\delta = \frac{1}{4}$

Figure 2. Density of states of Y^z and H^z around the cusp formation. The top and bottom figures show a plot of the boundary value of $\Im m^z$ equal to $\Im m_{Y^z}$ and $\Im m_{H^z}$, respectively, on the real line.

Theorem 2.1 (asymptotic 1-point function in the complex case). *Uniformly in $\tilde{\delta} \geq -C$ and $0 \leq \lambda \leq C$ for some fixed large constant $C > 0$, we have*

$$\boldsymbol{E}\,\mathrm{Tr}(Y - \lambda \cdot c(N, \tilde{\delta}) - \mathrm{i}0)^{-1} = \frac{1}{2\pi\mathrm{i}} \frac{N^{3/2}}{\tilde{z}_*} \int \mathrm{d}x \oint \mathrm{d}y\, e^{h(y)-h(x)} \tilde{H}(x, y) + \mathcal{O}(N(1 \vee \tilde{\delta})(1 + |\log \lambda|)), \tag{13a}$$

where

$$\begin{aligned} \tilde{H}(x, y) &:= \frac{1}{x^3} + \frac{1}{x^2 y} + \frac{1}{x y^2} + \frac{\tilde{\delta}\tilde{z}_*}{xy} + \frac{\tilde{\delta}\tilde{z}_*}{x^2}, \\ h(x) &:= -(1 \wedge \tilde{\delta}^{-1})\lambda \tilde{z}_* x + \frac{\tilde{\delta}}{x\tilde{z}_*} + \frac{1}{2x^2\tilde{z}_*^2}, \\ \tilde{z}_* &:= \lambda^{-1/3}(1 \vee \tilde{\delta}^{1/3})|\Psi(\tilde{\delta}\lambda^{-1/3}(1 \vee \tilde{\delta}^{1/3}))|, \\ c(N, \tilde{\delta}) &= N^{-3/2} \cdot (1 \wedge \tilde{\delta}^{-1}), \end{aligned} \tag{13b}$$

and where the x-integration is over any contour from 0 to $e^{3\mathrm{i}\pi/4}\infty$, going out from 0 in the direction of the positive real axis, and the y-integration is over any contour around 0 in a counterclockwise direction. Moreover, in the regime $\lambda \ll 1$, we have the bound

$$\left| \frac{1 \wedge \tilde{\delta}^{-1}}{\tilde{z}_*} \int \mathrm{d}x \oint \mathrm{d}y\, e^{h(y)-h(x)} \tilde{H}(x, y) \right| \lesssim \begin{cases} |\log \lambda|, & \lambda \geq \tilde{\delta}^3, \\ |\log \lambda\tilde{\delta}|, & \lambda < \tilde{\delta}^3. \end{cases} \tag{13c}$$

In (13c), the eigenvalue scaling, i.e., for $\lambda \gg 1$, the analogue of Theorem 2.1 reduces essentially to the local law asymptotics (10), albeit with a better error term due to the presence of the expectation.

Proposition 2.2. *Let $Y^z = (X-z)(X-z)^*$ where X is a complex Ginibre ensemble. Then, uniformly in $\delta := 1-|z|^2$ and $E \in \mathbb{R}$, we have the asymptotic expansion in $E\pm := E - \mathfrak{e}_\pm$,*

$$\begin{aligned} &\boldsymbol{E}\operatorname{Tr}(Y^z - E - \mathrm{i}0)^{-1} \\ &\quad = Nm^z(E+\mathrm{i}0) \times \left(1 + \mathcal{O}\left(\frac{1}{N|E_+|^{3/2}} + \frac{1}{NE^{2/3}} \wedge \left(\frac{\mathbb{1}_{\delta\geq 0}}{NE^{1/2}\delta^{1/2}} + \frac{\mathbb{1}_{\delta<0}}{N|E_-|^{3/2}|\delta|^{5/2}}\right)\right)\right), \end{aligned} \tag{14}$$

where the edges $\mathfrak{e}_\pm$ of $\Im m^z$ are explicit functions of δ given in (18a).

Real case. In the real case our main result is the following optimal bound on $\boldsymbol{E}\operatorname{Tr}(Y+E)^{-1}$ for $E>0$. Recall the notation $\delta := 1-|z|^2$.

Theorem 2.3 (optimal bound on the resolvent trace in the real case). *Let $\rho > 0$ be any small constant. Then uniformly in $E \geq 0$ and $\delta \geq -CN^{-1/2}$ for some fixed large constant $C > 0$ we have that*

$$|\boldsymbol{E}\operatorname{Tr}(Y+E)^{-1}| \lesssim \frac{e^{-\frac{1}{2}N(\Im z)^2}[N^{3/4} \vee N\sqrt{|\delta|}]}{\sqrt{E}} + (N^{3/2} \vee N^2|\delta|)\big[1 + |\log(NE^{2/3})|\big] \tag{15}$$

Finally, we present our bound on the tail asymptotics for both real and complex cases; for most applications, this can be viewed as the main result of this paper. Since a sizeable contribution to $\Im \operatorname{Tr}(Y - E + \mathrm{i}0)^{-1}$ and $\Im \operatorname{Tr}(Y+E)^{-1}$ comes from the smallest eigenvalue $\lambda_1(Y^z)$, by a straightforward Markov inequality we immediately obtain the following corollary on the tail asymptotics of $\lambda_1(Y^z)$ as an easy consequence of Theorems 2.1 and 2.3.

Corollary 2.4 (tail asymptotics of $\lambda_1(Y^z)$). *For any $C>0$, uniformly in $x \in (0, C]$ and $1-|z|^2 > -CN^{-1/2}$, we have the bound*

$$\boldsymbol{P}\big(\lambda_1(Y^z) \leq c(N,z)x\big) \lesssim (1 + |\log x|)x \tag{16}$$

in the complex case, and

$$\boldsymbol{P}\big(\lambda_1(Y^z) \leq c(N,z)x\big) \lesssim e^{-\frac{1}{2}N(\Im z)^2}\sqrt{x} + (1+|\log x|)x \tag{17}$$

in the real case, where we recall the definition of the scaling factor $c(N,z)$ from (5).

Properties of the asymptotic Stieltjes transform m^z. We now record some information on the deterministic Stieltjes transform m^z which will be useful later. The endpoints of the support of the density of states $\pi^{-1}\Im m^z$ are the zeros of the discriminant of the cubic equation (11) since passing through these points with the real parameter $E = \Re w$ creates solutions with nonzero imaginary part. Elementary calculations show that the support of $\Im m_{Y^z}$ is $[0, \mathfrak{e}_+]$ if $0 \leq \delta \leq 1$ and it is $[\mathfrak{e}_-, \mathfrak{e}_+]$ if $\delta < 0$, where

$$\mathfrak{e}_\pm := \frac{8\delta^2 \pm (9-8\delta)^{3/2} - 36\delta + 27}{8(1-\delta)}, \tag{18a}$$

and $\mathfrak{e}_-$ is only considered if $\delta < 0$. Note that while $\mathfrak{e}_+ \sim 1$, the edge $\mathfrak{e}_-$ may be close to 0; more precisely $0 < \mathfrak{e}_- = -4\delta^3/27(1 + \mathcal{O}(|\delta|))$. The slope coefficient of the square-root density at the edge in $\mathfrak{e}_\pm$ is given by

$$\Im m(\mathfrak{e}_\pm \mp \lambda) = \begin{cases} \gamma_\pm \sqrt{\lambda}(1+\mathcal{O}(\sqrt{\lambda})), & \lambda \geq 0, \\ 0, & \lambda \leq 0, \end{cases} \qquad \gamma_\pm := \frac{2\sqrt{2}(\sqrt{9-8\delta}\pm 1)^{3/2}}{(\sqrt{9-8\delta}\pm 3)^{5/2}\sqrt[4]{9-8\delta}}. \tag{18b}$$

Note that while the square-root edge at $\mathfrak{e}_+$ is nonsingular in the sense $\gamma_+ \sim 1$, the square-root edge in $\mathfrak{e}_-$ becomes singular for small $|\delta|$ as

$$\gamma_- = \frac{9}{4|\delta|^{5/2}}(1+\mathcal{O}(|\delta|)).$$

3. Supersymmetric method

Let $\chi_1, \overline{\chi_1}, \dots, \chi_N, \overline{\chi_N}$ denote Grassmannian variables satisfying the commutation rules

$$\chi_i\chi_j = -\chi_j\chi_i, \quad \chi_i\overline{\chi_j} = -\overline{\chi_j}\chi_i, \quad \overline{\chi_i}\,\overline{\chi_j} = -\overline{\chi_j}\,\overline{\chi_i},$$

from which it follows that $\chi_i^2 = \overline{\chi_i}^2 = 0$. As a convention we set $\overline{\overline{\chi_i}} := -\chi_i$. The power series of any function of Grassmannian variables is multilinear and it suffices to define the integral in the sense of Berezin [1987] over Grassmannian variables as the derivatives

$$\partial_{\chi_k}\chi_k = \partial_{\overline{\chi_k}}\overline{\chi_k} = 1, \quad \partial_{\chi_k}1 = \partial_{\overline{\chi_k}}1 = 0, \quad \partial_\chi := \partial_{\chi_1}\partial_{\overline{\chi_1}}\cdots\partial_{\chi_N}\partial_{\overline{\chi_N}}$$

and extend them multilinearly to all finite combinations of monomials in Grassmannians. We denote the column vectors with entries $\chi_1, \dots, \chi_N$ and $\overline{\chi_1}, \dots, \overline{\chi_N}$ by χ and $\overline{\chi}$, respectively. The conjugate transposes of those vectors, i.e., the row vectors with entries $\overline{\chi_1}, \dots, \overline{\chi_N}$ and $-\chi_1, \dots, -\chi_N$ will be denoted by χ^* and $\overline{\chi}^*$, respectively. Note that $(\chi^*)^* = -\chi$, $[\overline{\chi}^*]^* = -\overline{\chi}$. We now define the inner product of Grassmannian vectors χ, ϕ by

$$\langle \chi, \phi \rangle := \sum_i \overline{\chi_i}\phi_i,$$

so that the quadratic form $\sum_{i,j}\overline{\chi_i}A_{ij}\chi_j$ can be written as

$$\langle \chi, A\chi \rangle = \sum_{i,j}\overline{\chi_i}A_{ij}\chi_j,$$

where the matrix-vector product is understood in its usual sense. Similarly, s and $\overline{s}$ denote the column vectors with complex entries $s_1, \dots, s_N$ and their complex conjugates $\overline{s_1}, \dots, \overline{s_N}$, respectively, and for the conjugate transpose we have $(s^*)^* = s$ as usual. We have

$$\langle s, \phi \rangle := \sum_i \overline{s_i}\phi_i, \qquad \langle \chi, s \rangle := \sum_i \overline{\chi_i}s_i,$$

and similarly for quadratic forms. The commutation rules naturally also apply to linear functions of the Grassmannians, and therefore also, for example, $\langle s, \chi \rangle^2 = \langle \chi, s \rangle^2 = 0$ for any vector s of complex numbers. The complex numbers s_i and often called *bosonic* variables, while Grassmannians are called *fermions*, motivated by the basic (anti)commutativity of the bosonic/fermionic field operators in physics.

3A. ***Determinant identities.*** The backbone of the supersymmetric method are the determinant identities

$$\frac{1}{\mathrm{i}^N}\frac{\operatorname{sgn}(\Im w)^N}{\det(H-w)} = \int_{\mathbb{C}^N} \exp\bigl(-\mathrm{i}\operatorname{sgn}(\Im w)\langle s,(H-w)s\rangle\bigr)\,\mathrm{d}s, \qquad \mathrm{d}s := \frac{\mathrm{d}\Re s_1\,\mathrm{d}\Im s_1}{\pi}\cdots\frac{\mathrm{d}\Re s_N\,\mathrm{d}\Im s_N}{\pi},$$
$$\mathrm{i}^N\det(H-w) = \partial_\chi \exp(\mathrm{i}\langle\chi,(H-w)\chi\rangle), \qquad \partial_\chi := \partial_{\chi_1}\partial_{\overline{\chi_1}}\cdots\partial_{\chi_N}\partial_{\overline{\chi_N}},$$

where the exponential is defined by its (terminating) Taylor series. Consequently we can conveniently express the generating function as

$$Z(w,w_1) := \boldsymbol{E}\,\frac{\det(H-w_1)}{\det(H-w)} = \boldsymbol{E}\int_{\mathbb{C}^N}\mathrm{d}s\,\partial_\chi \exp(\mathrm{i}\langle\chi,(H-w_1)\chi\rangle - \mathrm{i}\langle s,(H-w)s\rangle),$$

for $w\in\mathbb{H}$ and $w_1\in\mathbb{C}$, where choice of w with $\Im w>0$ guarantees the convergence of the integral. By taking the w_1 derivative and setting $w=w_1$ it follows that

$$\begin{aligned}\operatorname{Tr}(H-w)^{-1} &= -\frac{\partial}{\partial w_1}\frac{\det(H-w_1)}{\det(H-w)}\bigg|_{w_1=w} = \mathrm{i}\int\langle\chi,\chi\rangle e^{-\mathrm{i}\operatorname{Tr}(H-w)[ss^*+\chi\chi^*]},\\ \int &:= \int_{\mathbb{C}^N}\mathrm{d}s\,\partial_\chi.\end{aligned}\tag{19}$$

3B. ***Superbosonization identity.*** After taking expectations, i.e., performing the Gaussian integration for the entries of $Y=Y^z=(X-z)(X-z)^*$, the resolvent identity (19) will depend on the complex vector s and the Grassmannian vector χ only via certain inner products. More specifically, after defining the $N\times 2$ and $N\times 4$ matrices $\Phi:=(s,\chi)$ and $\Psi:=(s,\overline{s},\chi,\overline{\chi})$, the expectation of the resolvent can be expressed as an integral over the 2×2 or 4×4 supermatrices $\Phi^*\Phi$ or $\Psi^*\Psi$ in the complex and real case, respectively. *Supermatrices* are 2×2 block matrices whose diagonal blocks are commonly referred to as the boson-boson and the fermion-fermion block, while the off-diagonal blocks are the boson-fermion and fermion-boson block. For supermatrices Q the *supertrace* and *superdeterminant*, the natural generalizations of trace and determinant, are given by

$$\operatorname{STr}\begin{pmatrix}x & \sigma\\ \tau & y\end{pmatrix} := \operatorname{Tr}(x)-\operatorname{Tr}(y), \quad \operatorname{SDet}\begin{pmatrix}x & \sigma\\ \tau & y\end{pmatrix} := \frac{\det(x)}{\det(y-\tau x^{-1}\sigma)},\tag{20}$$

and the inverse of a supermatrix is

$$\begin{pmatrix}x & \sigma\\ \tau & y\end{pmatrix}^{-1} = \begin{pmatrix}(x-\tau y^{-1}\sigma)^{-1} & -x^{-1}\sigma(y-\sigma x^{-1}\tau)\\ -y^{-1}\tau(x-\tau y^{-1}\sigma)^{-1} & (y-\sigma x^{-1}\tau)^{-1}\end{pmatrix}.\tag{21}$$

The integral over the remaining degrees of freedom in Φ,Ψ other than the inner products in $\Phi^*\Phi$, $\Psi^*\Psi$ can conveniently be performed using the well known *superbosonization formula* which we now recall. It basically identifies the integration volume of the irrelevant degrees of freedom with the high power of the superdeterminant of the supermatrix containing the relevant inner products (collected in a 2×2 supermatrix Q in the complex case and a 4×4 supermatrix Q in the real case).

3B1. *Complex superbosonization.* For any analytic function F with sufficiently fast decay at $+\infty$ in the boson-boson sector (in the variable x) the complex superbosonization identity from [Littelmann et al. 2008, Eq. (1.10)] implies

$$\int F(\Phi^*\Phi) = \int_Q \mathrm{SDet}^N(Q) F(Q), \quad \int_Q := \frac{1}{2\pi\mathrm{i}} \int \mathrm{d}x \oint \mathrm{d}y \, \partial_\sigma \partial_\tau, \quad Q := \begin{pmatrix} x & \sigma \\ \tau & y \end{pmatrix}, \tag{22}$$

where $\int \mathrm{d}x$ denotes the Lebesgue integral on $[0,\infty)$, $\oint \mathrm{d}y$ denotes the counterclockwise complex line integral on $\{z \in \mathbb{C} \mid |z| = 1\}$ and σ, τ denote independent scalar Grassmannian variables. The key point is that while the integral on the left-hand side is performed over N complex numbers and $2N$ Grassmannians, the integral on the right is simply over a 2×2 supermatrix, i.e., two complex variables and two Grassmannians. Note that the identity in [Littelmann et al. 2008] is more general than (22) in the sense that it allows for bosonic and fermionic sectors of unequal sizes. For the case of equal sizes, which concerns us, the formula gets simplified, the measure DQ in [Littelmann et al. 2008, Eq. (1.8)] becomes the flat Lebesgue measure since two determinants cancel each other as

$$\begin{aligned} \det(1 - x^{-1}\sigma y^{-1}\tau) \det(1 - y^{-1}\tau x^{-1}\sigma) &= e^{\mathrm{Tr}(\log(1 - x^{-1}\sigma y^{-1}\tau) + \log(1 - y^{-1}\tau x^{-1}\sigma))} \\ &= e^{-\sum_{k\geq 1} \frac{1}{k} \mathrm{Tr}((x^{-1}\sigma y^{-1}\tau)^k + (y^{-1}\tau x^{-1}\sigma)^k)} = 1, \end{aligned} \tag{23}$$

where the sum is finite and the last equality followed using the commutation rules.

3B2. *Real superbosonization.* In the real case we similarly have the real superbosonization identity from [Littelmann et al. 2008, Eq. (1.13)]:

$$\begin{aligned} \int F(\Psi^*\Psi) &= \int_Q \mathrm{SDet}^{N/2}(Q) F(Q), \\ \int_Q &:= \frac{1}{(2\pi)^2\mathrm{i}} \int \mathrm{d}x \oint \mathrm{d}y \, \partial_\sigma \left(\frac{\det(y)}{\det(x)} \right)^{1/2} \det\left(1 - \frac{x^{-1}}{y}\sigma\tau\right)^{-1/2}. \end{aligned} \tag{24}$$

The supermatrix Q has 2×2 blocks: x is nonnegative Hermitian, y is a scalar multiple of the identity matrix. The off-diagonal blocks σ, τ are related by

$$\tau := -t_a \sigma^t t_s^{-1}, \quad t_s := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad t_a := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Here the $\int \mathrm{d}x$ integral is the Lebesgue measure on nonnegative Hermitian 2×2 matrices x satisfying $x_{11} = x_{22}$, i.e.,

$$\int \mathrm{d}x := \int_0^\infty \mathrm{d}_{x_{11}} \int_{|x_{12}| \leq x_{11}} \mathrm{d}\Re x_{12} \, \mathrm{d}\Im x_{12},$$

and the fermionic integral is defined as $\partial_\sigma := \partial_{\sigma_{11}} \partial_{\sigma_{22}} \partial_{\sigma_{21}} \partial_{\sigma_{12}}$. Furthermore, under the slight abuse of notation of identifying the 2×2 matrix y, which is a scalar multiple of the identity matrix, with the corresponding scalar, $\oint \mathrm{d}y$ is the complex line integral over $|y| = 1$ in a counterclockwise direction. Unlike in the complex case, the matrix elements of the 4×4 supermatrix Q are not independent; there are only 4 (real) bosonic and 4 fermionic degrees of freedom. These identities among the elements of Q stem from natural relations between the scalar product of the column vectors of Ψ. For example the identity $\langle s,s\rangle=\langle\bar{s},\bar{s}\rangle$ from the first two diagonal elements of $(\Psi^*\Psi)$ corresponds to $x_{11}=x_{22}$, while $\langle\overline{\chi},\overline{\chi}\rangle=0$ is responsible for $y_{12}=0$. The relation

$$\tau=\begin{pmatrix}\sigma_{22} & \sigma_{12}\\ -\sigma_{21} & -\sigma_{11}\end{pmatrix}$$

encoded in the last line of (24) corresponds to relations between scalar products of bosonic and fermionic vectors and their complex conjugates; for example $\tau_{21}=-\sigma_{21}$ comes from the identity $(\Psi^*\Psi)_{41}=\langle-\bar{\chi},s\rangle=-\langle\bar{s},\chi\rangle=-(\Psi^*\Psi)_{23}$, etc.

3C. ***Application to Y^z in the complex case.*** Our goal is to evaluate $\boldsymbol{E}\,\mathrm{Tr}(H-w)^{-1}$ asymptotically on the scale where E is comparable with the eigenvalue spacing. We now use the identity

$$\begin{aligned}\mathrm{Tr}(Y-w)^{-1}&=\mathrm{i}\int\langle\chi,\chi\rangle e^{-\mathrm{i}\,\mathrm{Tr}[(X-z)(X^*-\bar{z})-w](ss^*+\chi\chi^*)}\\ &=\mathrm{i}\int\langle\chi,\chi\rangle e^{\mathrm{i}w\langle s,s\rangle-\mathrm{i}w\langle\chi,\chi\rangle-\mathrm{i}\,\mathrm{Tr}(X-z)(X^*-\bar{z})\Phi\Phi^*}\end{aligned}\tag{25}$$

for $w=E+\mathrm{i}\epsilon$ with $|E|\gg\epsilon>0$. We now compute the ensemble average as

$$\begin{aligned}&\boldsymbol{E}\,e^{-\mathrm{i}\,\mathrm{Tr}(X-z)(X^*-\bar{z})\Phi\Phi^*}\\ &=\Big(\frac{N}{\pi}\Big)^{N^2}\int\exp\Big[-N\,\mathrm{Tr}\,X^*\Big(1+\mathrm{i}\frac{\Phi\Phi^*}{N}\Big)X+\mathrm{i}\bar{z}\,\mathrm{Tr}\,\Phi\Phi^*X+\mathrm{i}z\,\mathrm{Tr}\,X^*\Phi\Phi^*-\mathrm{i}|z|^2\,\mathrm{Tr}\,\Phi\Phi^*\Big]\\ &=\mathrm{SDet}\Big(1+\mathrm{i}\frac{\Phi^*\Phi}{N}\Big)^{-N}\exp\Big(-\mathrm{i}|z|^2\Big[\mathrm{Tr}\,\Phi\Phi^*-\frac{\mathrm{i}}{N}\,\mathrm{Tr}\,\Phi\Phi^*\Big(1+\mathrm{i}\frac{\Phi\Phi^*}{N}\Big)^{-1}\Phi\Phi^*\Big]\Big),\\ &=\mathrm{SDet}\Big(1+\mathrm{i}\frac{\Phi^*\Phi}{N}\Big)^{-N}\exp\Big(-\mathrm{i}|z|^2\,\mathrm{Tr}\,\Phi\Big(1+\frac{\mathrm{i}}{N}\Phi^*\Phi\Big)^{-1}\Phi^*\Big),\\ &=\mathrm{SDet}\Big(1+\mathrm{i}\frac{\Phi^*\Phi}{N}\Big)^{-N}\exp\Big(-N|z|^2\,\mathrm{STr}\Big(1+\frac{\mathrm{i}}{N}\Phi^*\Phi\Big)^{-1}\frac{\mathrm{i}}{N}\Phi^*\Phi\Big).\end{aligned}\tag{26}$$

To perform the $\int=\int_{\mathbb{C}^N}\mathrm{d}s\,\partial_\chi$ integration in (25) we use the superbosonization formula (22) for the function

$$F(\Phi^*\Phi):=\langle\chi,\chi\rangle\,\mathrm{SDet}\Big(1+\mathrm{i}\frac{\Phi^*\Phi}{N}\Big)^{-N}\exp\Big(-N|z|^2\,\mathrm{STr}\Big(1+\frac{\mathrm{i}}{N}\Phi^*\Phi\Big)^{-1}\frac{\mathrm{i}}{N}\Phi^*\Phi+\mathrm{i}w\,\mathrm{STr}\,\Phi^*\Phi\Big).\tag{27}$$

We view F as a function of the four independent variables collected in the entries of the 2×2 matrix $\Phi^*\Phi$. Strictly speaking the function F is only meromorphic but not entire in these four variables, but since the integration regimes on both sides of the superbosonization formula are well separated away from the poles of F, a simple approximation argument outlined in Appendix A justifies its usage. Together with the change of variables $\frac{\mathrm{i}}{N}Q\mapsto Q$ it now implies that

$$\boldsymbol{E}\,\mathrm{Tr}(Y-w)^{-1}=N\int_{Q'}ye^{Nw\,\mathrm{STr}(Q)+N\log\mathrm{SDet}(Q)-N\log\mathrm{SDet}(1+Q)-N|z|^2\,\mathrm{STr}(1+Q)^{-1}Q},$$

where $\int_{Q'}$ indicates the changed integration regime due to the change of variables, more specifically, under Q' the x-integration is over $[0, \mathrm{i}\infty)$ and the y-integration is over a small circle $\{u \in \mathbb{C} \mid |u| = N^{-1}\}$. Note that the change of variables through scaling does not contribute an additional factor, since superdeterminants are scale invariant. It remains to perform the Berezinian integral. To do so we split

$$Q = q + \mu, \quad q = \begin{pmatrix} x & 0 \\ 0 & y \end{pmatrix}, \quad \mu = \begin{pmatrix} 0 & \sigma \\ \tau & 0 \end{pmatrix}$$

and compute

$$\begin{aligned} \exp(-N \log \mathrm{SDet}(1+Q)) &= \exp(-N \operatorname{STr} \log(1+q+\mu)) \\ &= \exp\Big(-N \operatorname{STr} \log(1+q) + \frac{N}{2} \operatorname{STr}(1+q)^{-1}\mu(1+q)^{-1}\mu\Big) \\ &= \exp(-N \log(1+x) + N \log(1+y))\Big(1 + N\frac{\sigma\tau}{(1+x)(1+y)}\Big), \\ \exp(N \log \mathrm{SDet}(Q)) &= \exp(N \log(x) - N \log(y))\Big(1 - N\frac{\sigma\tau}{xy}\Big) \end{aligned}$$

and

$$\begin{aligned} &\exp(-N|z|^2 \operatorname{STr}(1+Q)^{-1}Q) \\ &\qquad = \exp\big(-N|z|^2 \operatorname{STr}[(1+q)^{-1}q - (1+q)^{-1}\mu(1+q)^{-1}\mu(1+q)^{-1}]\big) \\ &\qquad = \exp\Big(-N|z|^2\frac{x}{1+x} + N|z|^2\frac{y}{1+y}\Big)\Big(1 + N|z|^2\frac{\sigma\tau}{(1+x)(1+y)}\Big(\frac{1}{1+x} + \frac{1}{1+y}\Big)\Big). \end{aligned}$$

By combining these identities we arrive at the final result:[1]

$$\begin{aligned} \boldsymbol{E} \operatorname{Tr}(Y-w)^{-1} &= \frac{N^2}{2\pi \mathrm{i}} \int_0^\infty \mathrm{d}x \oint \mathrm{d}y e^{-Nf(x)+Nf(y)} y \cdot G(x,y), \\ G(x,y) &:= \frac{1}{xy} - \frac{1}{(1+x)(1+y)}\Big[1 + \frac{|z|^2}{1+x} + \frac{|z|^2}{1+y}\Big], \\ f(x) &:= \log\frac{1+x}{x} - \frac{|z|^2}{1+x} - wx, \end{aligned} \tag{28}$$

where the x-integration is over $(0, \mathrm{i}\infty)$ and the y-integration is over a circle of radius N^{-1} around the origin.

3D. ***Application to Y^z in the real case.*** We now consider the real case and introduce the $N \times 4$ matrix $\Psi := (s, \bar{s}, \chi, \bar{\chi})$, the 2×2 matrix $Z := \left(\begin{smallmatrix} z & 0 \\ 0 & \bar{z} \end{smallmatrix}\right)$ and the 4×4 matrix $Z_2 := \left(\begin{smallmatrix} Z & 0 \\ 0 & Z \end{smallmatrix}\right)$, and use that

$$\mathrm{i}\bar{z} \operatorname{Tr} \Phi\Phi^* X + \mathrm{i}z \operatorname{Tr} X^t \Phi\Phi^* = \frac{\mathrm{i}}{2} \operatorname{Tr} \Psi Z_2^* \Psi^* X + \frac{\mathrm{i}}{2} \operatorname{Tr} X^t \Psi Z_2 \Psi^*, \quad (\Psi Z_2 \Psi^*)^t = \Psi Z_2^* \Psi^*,$$

[1] Essentially the same formula, obtained by direct computations, was presented by M. Shcherbina in her seminar talk on Jan 11, 2016 in Bonn. Our derivation of the same formula via superbosonization is merely a pedagogical preparation to the much more involved real case in Section 3D.

to compute

$$
\begin{aligned}
&\boldsymbol{E}\, e^{-\mathrm{i}\operatorname{Tr}(X-z)(X^t-\overline{z})\Phi\Phi^*}\\
&=\boldsymbol{E}\Big(\frac{N}{2\pi}\Big)^{\frac{N^2}{2}}\int\exp\Big(-\frac{N}{2}\operatorname{Tr}X^t\Big(1+\frac{2\mathrm{i}}{N}\Phi\Phi^*\Big)X+\mathrm{i}\overline{z}\operatorname{Tr}\Phi\Phi^*X+\mathrm{i}z\operatorname{Tr}X^t\Phi\Phi^*-\mathrm{i}|z|^2\operatorname{Tr}\Phi\Phi^*\Big)\\
&=\boldsymbol{E}\Big(\frac{N}{2\pi}\Big)^{\frac{N^2}{2}}\int\exp\Big(-\frac{N}{2}\operatorname{Tr}X^t\Big(1+\frac{\mathrm{i}}{N}\Psi\Psi^*\Big)X+\frac{\mathrm{i}}{2}\operatorname{Tr}\Psi Z_2^*\Psi^*X+\frac{\mathrm{i}}{2}\operatorname{Tr}X^t\Psi Z_2\Psi^*-\frac{\mathrm{i}|z|^2}{2}\operatorname{Tr}\Psi\Psi^*\Big)\\
&=\operatorname{SDet}\Big(1+\frac{\mathrm{i}}{N}\Psi^*\Psi\Big)^{-N/2}\exp\Big(-\frac{\mathrm{i}}{2}\operatorname{Tr}\Psi Z_2^*\Big(1-\frac{\mathrm{i}}{N}\Psi^*\Big(1+\frac{\mathrm{i}}{N}\Psi\Psi^*\Big)^{-1}\Psi\Big)Z_2\Psi^*\Big)\\
&=\operatorname{SDet}\Big(1+\frac{\mathrm{i}}{N}\Psi^*\Psi\Big)^{-N/2}\exp\Big(-\frac{\mathrm{i}}{2}\operatorname{Tr}\Psi Z_2^*\Big(1+\frac{\mathrm{i}}{N}\Psi^*\Psi\Big)^{-1}Z_2\Psi^*\Big)\\
&=\operatorname{SDet}\Big(1+\frac{\mathrm{i}}{N}\Psi^*\Psi\Big)^{-N/2}\exp\Big(-\frac{N}{2}\operatorname{STr}\Big(1+\frac{\mathrm{i}}{N}\Psi^*\Psi\Big)^{-1}Z_2\frac{\mathrm{i}}{N}\Psi^*\Psi Z_2^*\Big),
\end{aligned}
\tag{29}
$$

where we used that X is real and $\Psi\Psi^*$ is symmetric. The superbosonization formula thus implies

$$
\begin{aligned}
\boldsymbol{E}\operatorname{Tr}(H-w)^{-1}&=\frac{N}{2(2\pi)^2\mathrm{i}}\int\frac{\operatorname{Tr}(y)\det(y)^{1/2}}{\det(x)^{1/2}}\exp\big(-\tfrac{1}{2}\log\det(1-x^{-1}\sigma y^{-1}\tau)\big)\\
&\qquad\times\exp\Big(\frac{N}{2}\big[\operatorname{STr}(wQ)-\log\operatorname{SDet}(1+Q)+\log\operatorname{SDet}(Q)-\operatorname{STr}(1+Q)^{-1}Z_2QZ_2^*\big]\Big)\\
&=\frac{N}{2(2\pi)^2}\int\frac{\operatorname{Tr}(y)\det(y)^{1/2}}{\det(x)^{1/2}}\exp\big(-\tfrac{1}{2}\operatorname{Tr}\log(1-x^{-1}\sigma y^{-1}\tau)\big)\\
&\qquad\times\exp\Big(\frac{N}{2}\big[\operatorname{STr}(wQ)-\operatorname{STr}\log(1+Q)+\operatorname{STr}\log(Q)-\operatorname{STr}(1+Q)^{-1}Z_2QZ_2^*\big]\Big),
\end{aligned}
$$

where

$$
Q=\begin{pmatrix}x&\sigma\\ \tau&y\end{pmatrix}\equiv\frac{\mathrm{i}}{N}\Psi^*\Psi,\qquad \tau=\begin{pmatrix}0&1\\-1&0\end{pmatrix}\sigma^t\begin{pmatrix}0&1\\1&0\end{pmatrix}.
$$

In order to expand the exponential terms to fourth order in σ we introduce the shorthand notation

$$
Q=q+\mu,\quad q=\begin{pmatrix}x&0\\0&y\end{pmatrix},\quad \mu=\begin{pmatrix}0&\sigma\\ \tau&0\end{pmatrix}.
\tag{30}
$$

We compute

$$
\begin{aligned}
\operatorname{STr}(1+Q)^{-1}Z_2QZ_2^*&=\operatorname{STr}(1+q)^{-1}Z_2qZ_2^*\\
&\qquad-\operatorname{STr}(1+q)^{-1}\mu(1+q)^{-1}(Z_2\mu Z_2^*-\mu(1+q)^{-1}Z_2qZ_2^*)\\
&\qquad-((1+q)^{-1})^3(1+q)^{-1}(Z_2\mu Z_2^*-\mu(1+q)^{-1}Z_2qZ_2^*),\\
&=\operatorname{Tr}(1+x)^{-1}ZxZ^*-|z|^2\operatorname{Tr}\frac{y}{1+y}-\operatorname{Tr}(\sigma Z\tau Z^*A+Z\sigma Z^*\tau A)\\
&\qquad+\operatorname{Tr}\sigma\tau AC'-\operatorname{Tr}\sigma\tau A(\sigma Z\tau Z^*A+Z\sigma Z^*\tau A)+\operatorname{Tr}\sigma\tau A\sigma\tau AC',
\end{aligned}
$$

where we introduced matrices A, C' as in

$$\begin{aligned} A &:= \frac{(1+x)^{-1}}{1+y}, \\ B &:= \frac{x^{-1}}{y}, \\ C' &:= ZxZ^*(1+x)^{-1} + |z|^2 \frac{y}{1+y}, \end{aligned}$$

as well as B, which will be used in the sequel. In deriving these formulas we used that q and Z_2 have zero off-diagonal blocks and μ has zero diagonal blocks, to eliminate terms with odd powers of μ after taking the supertrace, and that y is a scalar multiple of the identity. Similarly we find for the logarithmic terms

$$\begin{aligned} &\operatorname{STr}(\log(q+\mu)-\log(1+q+\mu)) \\ &\quad = \operatorname{STr}\log(q(1+q)^{-1}) - \tfrac{1}{2}\operatorname{STr}(q^{-1}\mu)^2 - \tfrac{1}{4}\operatorname{STr}(q^{-1}\mu)^4 + \tfrac{1}{2}\operatorname{STr}((1+q)^{-1}\mu)^2 + \tfrac{1}{4}\operatorname{STr}((1+q)^{-1}\mu)^4 \\ &\quad = \log\frac{\det(x)}{\det(1+x)} - \log\frac{\det(y)}{\det(1+y)} + \operatorname{Tr}\sigma\tau(A-B) + \tfrac{1}{2}\operatorname{Tr}(\sigma\tau A)^2 - \tfrac{1}{2}\operatorname{Tr}(\sigma\tau B)^2 \end{aligned}$$

and

$$-\tfrac{1}{2}\operatorname{Tr}\log(1-x^{-1}\sigma y^{-1}\tau) = \tfrac{1}{2}\operatorname{Tr}\sigma\tau B + \tfrac{1}{4}\operatorname{Tr}(\sigma\tau B)^2.$$

Hence

$$\begin{aligned} \boldsymbol{E}\operatorname{Tr}(H-w)^{-1} &= \frac{N}{2(2\pi)^2\mathrm{i}} \int \mathrm{d}x \oint \mathrm{d}y \frac{\operatorname{Tr}(y)\det(y)^{1/2}}{\det(x)^{1/2}} \exp\Big(-\frac{N}{2}f(x) + \frac{N}{2}f(y)\Big) G(x,y,z), \\ f(x) &:= -w\operatorname{Tr}x + \log\frac{\det(1+x)}{\det(x)} + \operatorname{Tr}ZxZ^*(1+x)^{-1} - 2|z|^2, \\ G(x,y,z) &:= \partial_\sigma \exp\Big[\frac{N}{2}(\operatorname{Tr}\sigma\tau\big(A(1-C') - \big(1-\frac{1}{N}\big)B\big) + \operatorname{Tr}(\sigma Z\tau Z^*A + Z\sigma Z^*\tau A) \\ &\qquad + \operatorname{Tr}\sigma\tau A(\sigma\tau A\big(\tfrac{1}{2}-C'\big) + \sigma Z\tau Z^*A + Z\sigma Z^*\tau A) - \tfrac{1}{2}\Big(1-\frac{1}{N}\Big)\operatorname{Tr}(\sigma\tau B)^2)\Big], \end{aligned} \tag{31}$$

where $\int \mathrm{d}x = \int \mathrm{d}x_{11}\,\mathrm{d}\Re x_{12}\,\mathrm{d}\Im x_{12}$ is the integral over matrices of the form

$$x = \begin{pmatrix} \mathrm{i}x_{11} & \mathrm{i}x_{12} \\ \mathrm{i}\overline{x_{12}} & \mathrm{i}x_{11} \end{pmatrix}$$

with $x_{11} \in [0,\infty)$ and $x_{12} \in \mathbb{C}$ with $|x_{12}| \le x_{11}$. The integral $\oint \mathrm{d}y = \oint \mathrm{d}y_{11}$ is the integral over scalar matrices $y = y_{11}I$ with $\mathrm{d}y_{11}$ being the complex line integral over $|y_{11}| = N^{-1}$ in a counterclockwise direction.

To integrate out the Grassmannians we expand the exponential to second order, use the relation (30) between σ and τ, and use that for 2×2 matrices X and Y which are constant on the diagonal

$(x_{11} = x_{22},\, y_{11} = y_{22})$ we have the identities

$$\begin{aligned}
\frac{N^2}{8}\partial_\sigma \operatorname{Tr}^2(\sigma Z\tau Z^* X + Z\sigma Z^*\tau X) &= -4N^2|z|^2(\Re z)^2\det(X) + N^2|z|^2(\Im z)^2\operatorname{Tr}^2(X),\\
\frac{N^2}{8}\partial_\sigma \operatorname{Tr}^2(\sigma\tau X) &= -N^2\det(X),\\
\frac{N^2}{4}\partial_\sigma \operatorname{Tr}(\sigma\tau X)\operatorname{Tr}(\sigma Z\tau Z^* Y + Z\sigma Z^*\tau Y) &= 2N^2(\Re z)^2\big(\operatorname{Tr}(XY) - \operatorname{Tr}(X)\operatorname{Tr}(Y)\big)\\
&\quad + 2\mathrm{i}N^2(\Im z)(\Re z)\big(X_{12}Y_{21} - X_{21}Y_{12}\big),\\
\frac{N}{2}\partial_\sigma \operatorname{Tr}(\sigma\tau X\sigma\tau Y) &= N(\operatorname{Tr}(X)\operatorname{Tr}(Y) - \operatorname{Tr}(XY)),\\
\frac{N}{2}\partial_\sigma \operatorname{Tr}\sigma\tau X(\sigma Z\tau Z^* X + Z\sigma Z^*\tau X) &= 4N(\Re z)^2\det(X).
\end{aligned}$$

Hence we finally have the expression

$$\begin{aligned}
G = -N^2\bigg[&\det\Big(A(1-C') - \Big(1-\frac{1}{N}\Big)B\Big) + \big(4|z|^2(\Re z)^2 + 2(\Re z)^2(2-\operatorname{Tr}C')\big)\det A\\
&- |z|^2(\Im z)^2\operatorname{Tr}^2 A - 2(\Re z)^2\Big(1-\frac{1}{n}\Big)(\operatorname{Tr}A\operatorname{Tr}B - \operatorname{Tr}AB)\\
&- 2(\Re z)^2(\Im z)^2\det A^2\det(1+y)(4\det(x) - \operatorname{Tr}^2 x)\bigg]\\
&+ N\bigg(\det(A)(1+4(\Re z)^2 - \operatorname{Tr}C') - \Big(1-\frac{1}{N}\Big)\det(B)\bigg).
\end{aligned} \tag{32}$$

We now rewrite (31) by using the parametrizations

$$x = \begin{pmatrix} a & a\sqrt{1-\tau}e^{\mathrm{i}\varphi}\\ a\sqrt{1-\tau}e^{-\mathrm{i}\varphi} & a\end{pmatrix}, \qquad y = \begin{pmatrix}\xi & 0\\ 0 & \xi\end{pmatrix}, \tag{33}$$

with $a \in \mathrm{i}\mathbb{R}_+,\ \tau \in [0,1],\ \varphi \in [0,2\pi]$ and $|\xi| = N^{-1}$. Since the integral over $\varphi \in [0,2\pi]$ is equal to 2π as a consequence of the fact that the functions f, g, G_N defined below do not depend on φ, we have that

$$\boldsymbol{E}\operatorname{Tr}[Y-w]^{-1} = \frac{N}{4\pi\mathrm{i}}\oint \mathrm{d}\xi \int_0^{\mathrm{i}\infty}\mathrm{d}a\int_0^1 \mathrm{d}\tau\,\frac{\xi^2 a}{\tau^{1/2}}e^{N[f(\xi)-g(a,\tau,\eta)]}G_N(a,\tau,\xi,z), \tag{34}$$

where, using the notation $\eta := \Im z$, the functions f and g are defined by

$$f(\xi) := -w\xi + \log(1+\xi) - \log\xi - \frac{|z|^2}{1+\xi}, \tag{35}$$

$$g(a,\tau,\eta) := -wa + \tfrac{1}{2}\log[1+2a+a^2\tau] - \log a - \tfrac{1}{2}\log\tau - \frac{|z|^2(1+a) - 2\eta^2a^2(1-\tau)}{1+2a+a^2\tau}. \tag{36}$$

Note that $g(a,1,\eta) = f(a)$; in particular, we remark that $g(a,1,\eta)$ is independent of η for any $a \in \mathbb{C}$.

Furthermore, using the parametrizations in (33) the function $G_N := G_{1,N} + G_{2,N}$ is given by

$$\begin{aligned} G_{1,N} = \Bigg(& N^2 \frac{p_{2,0,0}}{a^2\xi^2(\xi+1)^2\tau} - N\frac{p_{1,0,0}}{a^2\xi^2(\xi+1)\tau} + \delta N^2 \frac{p_{2,0,1}}{a\xi(\xi+1)^2\tau} - N\delta \frac{p_{1,0,1}}{a\xi(\xi+1)\tau} + N^2\delta^2 \frac{p_{2,0,2}}{(\xi+1)^2}\Bigg) \\ & \times((a^2\tau+2a+1)^2(\xi+1)^2)^{-1}, \\ G_{2,N} = \Bigg(& N^2\eta^2 \frac{p_{2,2,0}}{a\xi(\xi+1)^3\tau} - N\eta^2 \frac{p_{1,2,0}}{a\xi\tau} + N^2\eta^2\delta \frac{p_{2,2,1}}{(\xi+1)}\Bigg) \times((a^2\tau+2a+1)^2(\xi+1)^2)^{-1}, \end{aligned} \tag{37}$$

where $p_{i,j,k} = p_{i,j,k}(a,\tau,\xi)$ are explicit polynomials in a,τ,ξ which we defer to Appendix B, $\eta := \Im z$ and $\delta := 1-|z|^2$. The indices i,j,k in the definition of $p_{i,j,k}$ denote the N, η and δ power, respectively. We split G_N as the sum of $G_{1,N}$ and $G_{2,N}$ since $G_{1,N}$ depends only on $|z|$, whilst $G_{2,N}$ depends explicitly by $\eta = \Im[z]$, hence $G_{2,N} = 0$ if $z \in \mathbb{R}$.

4. Asymptotic analysis in the complex case for the saddle point regime

For the density of states ϱ_{Y^z} on the positive semiaxis $E > 0$ we expect a singular behavior for $E \approx 0$ and a square-root edge for $E \approx \mathfrak{e}_+$. The singularity at $E = 0$ exhibits a phase transition in δ at 0; for $\delta > 0$ the transition is between an $E^{-1/3}$-singularity for $\delta = 0$ and a $\delta^{1/2}E^{-1/2}$-singularity for $0 < \delta \le 1$, while for $\delta < 0$ the transition is between the $E^{-1/3}$-singularity for $\delta = 0$ and square-root edge in $\mathfrak{e}_- \sim |\delta|^3$ of slope $|\delta|^{-5/2}$. We now analyze the location of the critical point(s) x_*, i.e., the solutions to $f'(x_*) = 0$, as well as the asymptotics of the phase function f around them precisely in all of the above regimes. For the saddlepoint approximation the second derivative $f''(x_*)$ is of particular importance and we find that it can only vanish in the vicinity of $E \approx \mathfrak{e}_+$ and $E \approx \mathfrak{e}_- \vee 0$, and otherwise satisfies $|f''(x_*)| \gtrsim 1$.

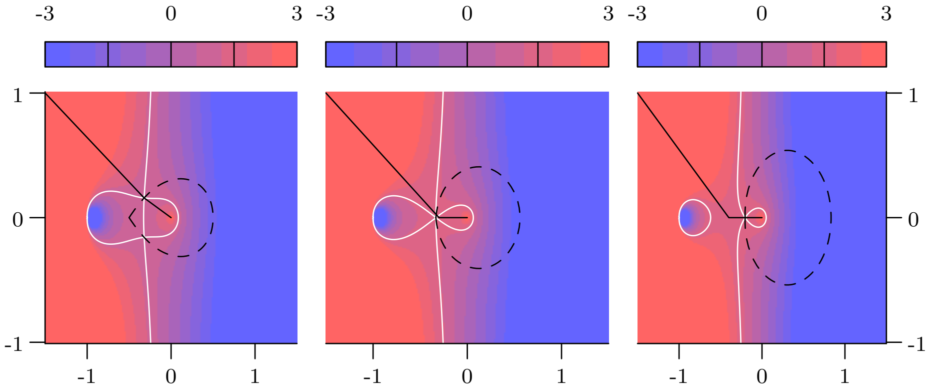
The saddle point equation $f'(x_*) = 0$ leads to the simple cubic equation

$$wx_*^3 + 2wx_*^2 + wx_* + \delta x_* + 1 = 0,$$

which is precisely the MDE equation from (11), whose explicit solution via Cardano's formula reveals that for $E \in (\mathfrak{e}_-, \mathfrak{e}_+)$ there are two relevant critical points $x_*, \overline{x_*}$ with $\Re f(x_*) = \Re f(\overline{x_*})$, while for $E \ge \mathfrak{e}_+$ or $0 \le E \le \mathfrak{e}_-$ there is one relevant critical point x_*, where x_* is given by

$$\begin{gathered} x_* = \begin{cases} e^{-2\mathrm{i}\pi/3}\sqrt[3]{q+\sqrt{q^2+p^3}} + e^{2\mathrm{i}\pi/3}\sqrt[3]{q-\sqrt{q^2+p^3}} - \frac{2}{3}, & E \le \mathfrak{e}_-, \\ e^{2\mathrm{i}\pi/3}\sqrt[3]{q+\sqrt{q^2+p^3}} + e^{-2\mathrm{i}\pi/3}\sqrt[3]{q-\sqrt{q^2+p^3}} - \frac{2}{3}, & \mathfrak{e}_- \le E \le \mathfrak{e}_+, \\ \sqrt[3]{q+\sqrt{q^2+p^3}} + \sqrt[3]{q-\sqrt{q^2+p^3}} - \frac{2}{3}, & E \ge \mathfrak{e}_+, \end{cases} \\ q := \frac{\delta}{3E} + \frac{1}{27} - \frac{1}{2E}, \qquad p := \frac{\delta}{3E} - \frac{1}{9}, \end{gathered} \tag{38}$$

where $q^2 + p^3 > 0$ as long as $E \in (\mathfrak{e}_-, \mathfrak{e}_+)$ and $q^2 + p^3 < 0$ for $E > \mathfrak{e}_+$ or $E < \mathfrak{e}_-$. Here we chose the branch of the cubic root such that $\sqrt[3]{\mathbb{R}} = \mathbb{R}$ and that $\sqrt[3]{z}$ for $z \in \mathbb{C}\setminus\mathbb{R}$ is the cubic root with the maximal real part. Note that the choice of the cubic root implies $\Im x_* \ge 0$ and $x_* = x_*(E) = m^z(E + \mathrm{i}0)$, where m^z has been defined in (11).



(a) $E_+ < 0$ (b) $|E_+| = 0$ (c) $E_+ > 0$

Figure 3. Contour plot of $\Re f(x)$ in the regime $E \approx \mathfrak{e}_+$. The solid white lines represent the level set $\Re f(x) = \Re f(x_*)$, while the solid and dashed black lines represent the chosen contours for the x- and y-integrations, respectively.

Before concluding this section with the proof of Proposition 2.2, we collect certain asymptotics of the critical point x_* and the phase function f in its vicinity, which will also be used in the main estimates of the present paper in Sections 5–6. In the edges $\mathfrak{e}_\pm$ the critical points have the simple expressions

$$x_*(\mathfrak{e}_\pm) = -\frac{2}{3 \pm \sqrt{9-8\delta}}$$

and satisfy $x_*(\mathfrak{e}_+) \sim -1$ and $x_*(\mathfrak{e}_-) = -3/(2\delta)[1 + \mathcal{O}(|\delta|)]$. Elementary expansions of (38) for E near the edges reveal the following asymptotics of x_* in the various regimes.

Regime $E \approx \mathfrak{e}_+$. Close to the spectral edge $E \approx \mathfrak{e}_+$ we have the asymptotic expansion

$$x_* = x_*(\mathfrak{e}_+) + \gamma_+ \sqrt{E_+}(1 + \mathcal{O}(|E_+^{1/2}|)) \tag{39a}$$

in $E_+ := E - \mathfrak{e}_+$, where γ_+ was defined in (18b). The location of saddle point(s) in the regime $E \approx \mathfrak{e}_+$ is depicted in Figure 3. The second derivative of f is asymptotically given by

$$f''(x_*) = \frac{2\sqrt{E_+}}{\gamma_+}(1 + \mathcal{O}(|E_+^{1/2}|)). \tag{39b}$$

Regime $E \approx 0$ *in the case* $\delta \geq 0$. For $E \approx 0$ we have the asymptotic expansions

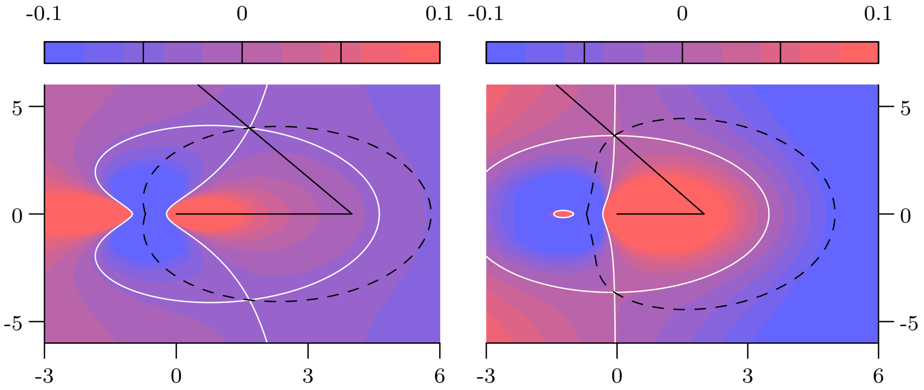
$$x_* = E^{-1/3}\Psi\Big(\frac{\delta}{E^{1/3}}\Big)[1 + \mathcal{O}(\delta + E^{1/3})], \tag{40a}$$

where $\Psi(\lambda)$ is the unique solution to the cubic equation

$$1 + \lambda\Psi(\lambda) + \Psi(\lambda)^3 = 0, \quad \Re\Psi(\lambda) > 0, \quad \Im\Psi(\lambda) > 0, \quad \lambda \geq 0.$$

The explicit function $\Psi(\lambda)$ has the asymptotics

$$\lim_{\lambda \searrow 0} \Psi(\lambda) = \Psi(0) = e^{i\pi/3}, \quad \lim_{\lambda\to\infty} \frac{\Psi(\lambda)}{\sqrt{\lambda}} = i.$$



(a) $0 < \delta \ll E^{1/3} \ll 1$ (b) $0 < E^{1/3} \ll \delta$

Figure 4. Contour plot of $\Re f(x)$ for $\delta > 0$ in the regime $E \approx 0$. The solid white lines represent the level set $\Re f(x) = \Re f(x_*)$, while the solid and dashed black lines represent the chosen contours for the x- and y-integrations, respectively.

Thus it follows that

$$\begin{aligned} x_* &= \Big(\mathrm{i}\sqrt{\frac{\delta}{E}} - 1 + \frac{1}{2\delta}\Big)\Big(1 + \mathcal{O}\Big(\frac{E}{\delta^3}\Big)\Big) \\ &= \Big(\frac{e^{\mathrm{i}\pi/3}}{E^{1/3}} - \frac{2}{3}\Big)\Big(1 + \mathcal{O}\Big(E^{2/3} + \frac{\delta}{E^{1/3}}\Big)\Big), \end{aligned} \tag{40b}$$

where the first expansion is informative in the $E \ll \delta^3$, and the second one in the $E \gg \delta^3$ regime. The location of saddle point(s) in the regime $E \approx 0$ is depicted in Figure 4. For the second derivative we have the expansions

$$\begin{aligned} f''(x_*) &= 3e^{2\mathrm{i}\pi/3} E^{4/3}\Big(1 + \mathcal{O}\Big(E + \frac{\delta}{E^{1/3}}\Big)\Big) \\ &= 2\mathrm{i}\frac{E^{3/2}}{\delta^{1/2}}\Big(1 + \mathcal{O}\Big(\frac{E}{\delta^3}\Big)\Big) \end{aligned} \tag{40c}$$

and similarly for higher derivatives, $|f^{(k)}(x_*)| \sim E^{(2+k)/3} \wedge E^{(k+1)/2}\delta^{-(k-1)/2}$ for $k \geq 3$.

Regime $E \approx \mathfrak{e}_-$ in the case $\delta < 0$. Around the spectral edge $\mathfrak{e}_-$ the critical point admits the asymptotic expansion

$$\begin{aligned} x_* &= x_*(\mathfrak{e}_-) + \gamma_-\Big(1 + \mathcal{O}\Big(\frac{|E|^{1/2}}{\delta^{3/2}}\Big)\Big)\begin{cases} \mathrm{i}\sqrt{|E_-|}, & E_- \geq 0, \\ -\sqrt{|E_-|}, & E_- \leq 0, \end{cases} \\ &= \frac{1}{E^{1/3}}\Big(e^{\mathrm{i}\pi/3} + \frac{\mathrm{i}}{3}e^{\mathrm{i}\pi/3}\frac{\delta}{E^{1/3}}(1 + \mathcal{O}(|E|^{1/3})) + \mathcal{O}\Big(\frac{|\delta|^2}{E^{2/3}}\Big)\Big), \end{aligned} \tag{41a}$$

where $E_- := E - \mathfrak{e}_-$, and these separate expansions are relevant in the $|E| \ll |\delta|^3$ and $|E| \gg |\delta|^3$ regimes, respectively. The location of saddle point(s) in the regime $E \approx \mathfrak{e}_-$ is depicted in Figure 5. The second

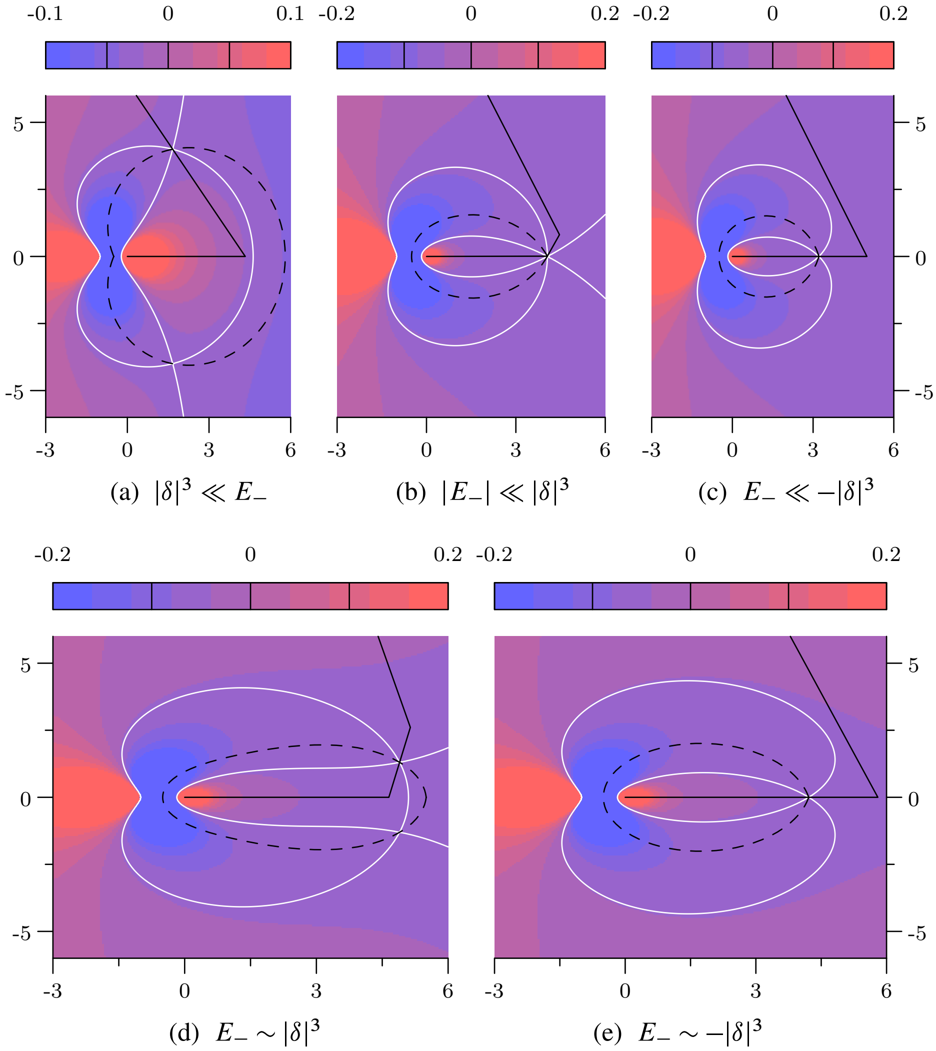


Figure 5. Contour plot of $\Re f(x)$ in the regime $E \approx \mathfrak{e}_-$. The solid white lines represent the level set $\Re f(x) = \Re f(x_*)$, while the solid and dashed black lines represent the chosen contours for the x- and y-integrations, respectively.

derivative around x_* is given by

$$\begin{aligned} f''(x_*) &= \frac{2}{\gamma_-}(1+\mathcal{O}(|E_-|^{1/2}|\delta|^{-3/2})) \times \begin{cases} \sqrt{|E_-|}, & E_- \le 0, \\ -\mathrm{i}\sqrt{|E_-|}, & E_- \ge 0, \end{cases} \\ &= 3e^{2\mathrm{i}\pi/3} E^{4/3}\left(1+\mathcal{O}\left(E+\frac{|\delta|}{E^{1/3}}\right)\right), \end{aligned} \tag{41b}$$

with $\gamma_- \sim |\delta|^{-5/2}$.

Proof of Proposition 2.2. As the functions f and G in (28) are meromorphic we are free to deform the contours for the x- and y-integrals as long as we are not crossing 0 or -1 and the x-contour goes out from 0 in the "right" direction (in the region $\Re[x] < 0, \Im[x] > 0$ in Figure 3, and in the region $\Re[x] > |\Im[x]|$ in Figures 4–5). It is easy to see that the contours can always be deformed in such a way that $\Re f(x) > \Re f(x_*) = \Re f(\overline{x_*})$ and $\Re f(y) < \Re f(x_*)$ for all $x, y \neq x_*, \overline{x_*}$; see Figures 3–5 for an illustration of the chosen contours.

We now compute the integral (28) in the large N limit when E is near the edges. In certain regimes of the parameters N, E and δ a saddle point analysis is applicable after a suitable contour deformation. In most cases, the result is a point evaluation of the integrand at the saddle points. In some transition regimes of the parameters the saddle point analysis only allows us to explicitly scale out some combination of the parameters and leaving an integral depending only on a reduced set of rescaled parameters.

We recall the classical quadratic saddle point approximation for holomorphic functions $f(z), g(z)$ such that $f(z)$ has a unique critical point in some z_*, and that γ can be deformed to go through z_* in such a way that $\Re f(z) < \Re f(z_*)$ for all $\gamma \ni z \neq z_*$. Then for large $\lambda \gg 1$ the saddle point approximation is given by

$$\int_\gamma g(z) e^{\lambda f(z)} \,\mathrm{d}z = \pm g(z_*) e^{\lambda f(z_*)} \sqrt{\frac{2\pi}{\lambda |f''(z_*)|}} \mathrm{i} e^{-\frac{\mathrm{i}}{2} \arg f''(z_*)} \Big(1 + \mathcal{O}\Big(\frac{1}{\lambda}\Big)\Big), \tag{42}$$

where $\pm$ is determined by the direction of γ through z_* with $+$ corresponding to the direction parallel to $\mathrm{i}\exp\big(-\frac{\mathrm{i}}{2}\arg f''(z_*)\big)$. This formula is applicable, i.e., we can use point evaluation in the *saddle point regime*, whenever the lengthscale $\ell_f \sim (N|f''(x_*)|)^{-1/2}$ of the exponential decay from the quadratic approximation of the phase function is much smaller than the scale $\ell_g \sim |g(x_*)|/|\nabla g(x_*)|$ on which g is essentially unchanged. For our integral (28) we thus need to check the condition

$$\frac{1}{\sqrt{N|f''(x_*)|}} \ll \left| \frac{\nabla(yG(x,y))}{yG(x,y)} \Big|_{(x,y)=(x_*,x_*)} \right|^{-1} \tag{43}$$

in all regimes separately.

In the regime $E \approx \mathfrak{e}_+$, using the asymptotics for x_* from (39a) and (39b), the quadratic saddle point approximation is valid if

$$\frac{1}{\sqrt{N|E_+|^{1/2}}} \ll |E_+|^{1/2},$$

i.e., if $|E_+| \gg N^{-2/3}$. Here the length-scale $|E_+|^{1/2}$ represents the length-scale on which $(x,y) \mapsto yG(x,y)$ is essentially constant which can be obtained by explicitly computing the log-derivative

$$\left| \frac{\nabla(yG(x,y))}{yG(x,y)} \Big|_{(x,y)=(x_*,x_*)} \right| \sim |E_+|^{-1/2}.$$

Similar calculations yield that for $E \approx 0$ and $\delta \geq 0$ the quadratic saddle point approximation is valid if

$$\frac{1}{\sqrt{N\big(E^{4/3} \wedge E^{3/2}\delta^{-1/2}\big)}} \ll E^{-1/3} \vee \delta^{1/2} E^{-1/2}, \quad \text{i.e., } E \gg N^{-3/2} \wedge N^{-2}\delta^{-1},$$

while for $E \approx \mathfrak{e}_-$ and $\delta < 0$ the condition (43) reads

$$\frac{1}{\sqrt{N\bigl(E^{4/3} \vee |E_-|^{1/2}|\delta|^{5/2}\bigr)}} \ll E^{-1/3} \wedge |E_-|^{1/2}|\delta|^{-5/2}, \quad \text{i.e., } |E_-| \gg N^{-2/3}|\delta|^{5/3},$$

recalling that $E = E_- + \mathfrak{e}_-$ and $\mathfrak{e}_- \sim \delta^3$ from (18b).

In these regimes we can thus apply (42) to (28) and using that

$$G(x_*, x_*) = f''(x_*), \qquad G(x_*, \overline{x_*}) = 0,$$

as follows from explicit computations, we thus finally conclude (14). Here the error terms in (14) follow from (42) by choosing $\lambda \sim (\ell_g/\ell_f)^2$ according to asymptotics of the second derivatives and log-derivatives above. More precisely, for example in the second case $E \approx 0$ and $E^{1/3} \gg \delta > 0$, the phase function f is approximately given by

$$f(x) \approx E^{2/3}\biggl[\frac{1}{2(E^{1/3}x)} - E^{1/3}x\biggr],$$

while G can asymptotically be written as

$$yG(x, y) \approx x_*G(x_*, x_*) + 3E^{4/3}e^{2\mathrm{i}\pi/3}(2(x - x_*) + (y - x_*)) + \mathcal{O}(E^{5/3}(|x - x_*|^2 + |y - x_*|^2)).$$

Thus we make the change of variables $x = x_* + E^{-1/3}x'$, $y = x_* + E^{-1/3}y'$ to find

$$\begin{aligned}&\frac{N^2}{2\pi\mathrm{i}}E^{-2/3}\int \mathrm{d}x' \int \mathrm{d}y' e^{-NE^{-2/3}f''(x_*)(\frac{x'^2}{2}-\frac{y'^2}{2})-NE^{-1}f'''(x_*)(\frac{x'^2}{6}-\frac{y'^2}{6})+\mathcal{O}(NE^{2/3}(|x'|^4+|y'|^4))} \\ &\qquad\qquad\qquad \times(x_*G(x_*, x_*) + 3Ee^{2\mathrm{i}\pi/3}(y' + 2x') + \mathcal{O}(E(|x'|^2 + |y'|^2))) \\ &\quad = x_*\biggl(1 + \mathcal{O}\biggl(\frac{1}{NE^{2/3}}\biggr)\biggr),\end{aligned}$$

where we used that $G(x_*, x_*) \sim E^{4/3}$. The other cases in (14) can be checked similarly. □

5. Derivation of the 1-point function in the critical regime for the complex case

In this section we prove Theorem 2.1, i.e., we study $\boldsymbol{E}\,\mathrm{Tr}[Y - w]^{-1}$, with $w = E + \mathrm{i}\epsilon$, $1 \gg |E| \gg \epsilon > 0$, for E so close to 0 such that $|E|$ is smaller or comparable with the eigenvalues scaling around 0. We will first consider the case $\delta \geq 0$ and afterwards explain the necessary changes in the regime $-CN^{-1/2} \leq \delta < 0$.

5A. ***Case* $\mathbf{0 \leq \delta \leq 1}$.** In the following of this section we assume that $E > 0$, since we are interested in the computations of (28) for $E = \Re[w]$ inside the spectrum of Y. In order to study the transition between the local law regime, that is considered in Section 4, and the regime when the main contribution to (28) comes from the smallest eigenvalue of Y, we define the parameter

$$c(N) = c(N, \delta) := \frac{1}{N^{3/2}} \wedge \frac{1}{\delta N^2}. \tag{44}$$

In particular, in the regime $E \gg c(N)$ the double integral in (28) is computed by saddle point analysis in (14), i.e., the main contribution comes from the regime around the stationary point x_* of f, with x_*

defined in (40b), whilst for $E \lesssim c(N)$ the main contribution to (28) comes from a larger regime around the stationary point x_*. From now on we assume that $E \lesssim c(N)$. In the following we denote the leading order of the stationary point x_* by

$$z_* = z_*(E,\delta) := E^{-1/3}\Psi(\delta E^{-1/3}), \tag{45}$$

where $\Psi(\lambda)$ was defined in (40a) and has the asymptotics $\Psi(0) = e^{\mathrm{i}\pi/3}$ and $\Psi(\lambda)/\sqrt{\lambda} \to \mathrm{i}$ as $\lambda \to \infty$. Note that $|z_*| \gg 1$ for any $E \ll 1$, $0 \le \delta \le 1$. For this reason, we expect that the main contribution to the double integral in (28) comes from the regime when $|x|$ and $|y|$ are both large, say $|x|, |y| \ge N^\rho$, for some small fixed $0 < \rho < 1/2$. Later on in this section, see Lemma 5.4, we prove that the contribution to (28) in the regime when either $|x|$ or $|y|$ are smaller than N^ρ is exponentially small. In order to get the asymptotics in (13a), is not affordable to estimate the error terms in the Taylor expansion by absolute value. In particular, it is not affordable to estimate the integral of $e^{Nf(y)}$ over Γ by absolute value, hence the improved bound in (54) is needed. To make our writing easier, for any $R \in \mathbb{N}$, $R \ge 2$, we introduce the notation

$$\mathcal{O}^\#(x^{-R}) := \{g \in \mathcal{P}_R\}, \quad \mathcal{O}^\#((x,y)^{-R}) := \{g \in \mathcal{Q}_R\}, \tag{46}$$

where $\mathcal{P}_R$ and $\mathcal{Q}_R$ are defined as Laurent series of order at least R around infinity, i.e.,

$$\mathcal{P}_R := \left\{ g\colon \mathbb{C} \to \mathbb{C} \,\middle|\, g(x) = \sum_{\alpha \ge R} \frac{\tilde{c}_\alpha}{x^\alpha}, \text{ with } |\tilde{c}_\alpha| \le C^\alpha, \text{ if } |x| \ge 2C \right\},$$

$$\mathcal{Q}_R := \left\{ \tilde{g}\colon \mathbb{C}\times\mathbb{C} \to \mathbb{C} \,\middle|\, \tilde{g}(x,y) = \sum_{\alpha,\beta \ge 1, \alpha+\beta \ge R} \frac{c_{\alpha,\beta}}{x^\alpha y^\beta}, \text{ with } |c_{\alpha,\beta}| \le C^{\alpha+\beta}, \text{ if } |x|,|y| \ge 2C \right\},$$

for some constant $C > 0$ that is implicit in the $\mathcal{O}^\#$ notation. Here α, β are integer exponents. Note that

$$\mathcal{O}^\#(|x|^{-R}) = \mathcal{O}(|x|^{-R})$$

for any $x \in \mathbb{C}$. Then, we expand the phase function f for large argument as follows

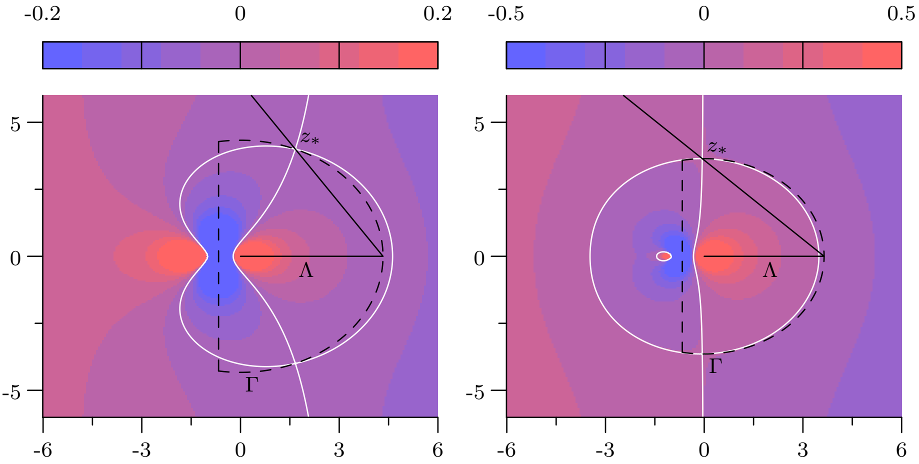
$$f(x) = g(x) + \mathcal{O}^\#(x^{-3} + \delta x^{-2}), \quad g(x) := -(E + \mathrm{i}\epsilon)x + \frac{\delta}{x} + \frac{1}{2x^2}. \tag{47a}$$

and for large x and y we expand G as

$$G(x,y) = H(x,y) + \mathcal{O}^\#((x,y)^{-5} + \delta(x,y)^{-4}), \quad H(x,y) := \frac{1}{x^3 y} + \frac{1}{x^2 y^2} + \frac{1}{x y^3} + \frac{\delta}{x y^2} + \frac{\delta}{x^2 y}. \tag{47b}$$

In order to compute the integral in (28) we deform the contours Λ and Γ through z_*, with z_* defined in (45). In particular, we are allowed to deform the contours as long as the x-contour goes out from zero in region $\Re[x] > |\Im[x]|$, it ends in the region $\Re[x] < 0, \Im[x] > 0$, and it does not cross 0 and -1 along the deformation; the y-contour, instead, can be freely deformed as long as it does not cross 0 and -1. Hence, we can deform the y-contour as $\Gamma = \Gamma_{z_*} := \Gamma_{1,z_*} \cup \Gamma_{2,z_*}$, where

$$\Gamma_{1,z_*} := \left\{ -\tfrac{2}{3} + \mathrm{i}t : 0 \le |t| \le \sqrt{|z_*|^2 - \tfrac{4}{9}} \right\}, \quad \Gamma_{2,z_*} := \{ |z^*| e^{\mathrm{i}\psi} : \psi \in [-\psi_{z_*}, \psi_{z_*}] \}, \tag{48a}$$



(a) $0 < \delta \ll E^{1/3} \ll 1$ (e) $0 < E^{1/3} \ll \delta$

Figure 6. Illustration of the contours (48a)–(48b) together with the phase diagram of $\Re f$, where the white line represents the level set $\Re f(x) = \Re f(z_*)$. Note that the precise choice of the contours is only important close to 0 and for very large $|x|$ as otherwise the phase function is small.

with $\psi_{z_*} = \arccos[-2/(3|z_*|)]$, and the x-contour as $\Lambda = \Lambda_{z_*} := \Lambda_{1,z_*} \cup \Lambda_{2,z_*}$, with

$$\Lambda_{1,z_*} := [0, |z_*|), \quad \Lambda_{2,z_*} := \{|z_*| - qs + \mathrm{i}\, s : s \in [0, +\infty)\}, \tag{48b}$$

where

$$q = q_{z_*} := \Im[z_*]^{-1}(|z_*| - \Re[z_*]). \tag{48c}$$

Note that $q \sim 1$ uniformly in N, E and δ, since $\Re[z_*] \lesssim \Im[z_*]$ for any $E \ll 1$, $0 \le \delta \le 1$. We assume the convention that the orientation of Γ is counter clockwise. See Figure 6 for an illustration of Γ and Λ.

Before proceeding with the computation of the leading term of (28), in the following lemma we state some properties of the function f on the contours Γ, Λ. Using that $\epsilon \ll E$, the proof of the lemma below follows by easy computations.

Lemma 5.1. *Let f be the phase function defined in (28), then the following properties hold true:*

(i) *For any $y = -2/3 + \mathrm{i}t \in \Gamma_{1,z_*}$, we have that*

$$\Re[f(-2/3 + \mathrm{i}t)] = \frac{2E}{3} + \epsilon t - \frac{1}{2t^2} + \mathcal{O}(|t|^{-3} + \delta|t|^{-2}), \tag{49}$$

and

$$\Im[f(-2/3 + \mathrm{i}t)] = \frac{2\epsilon}{3} - Et - \frac{\delta}{t} + \mathcal{O}(|t|^{-3}). \tag{50}$$

(ii) *For $\epsilon = 0$, the function $t \mapsto \Re[f(-2/3 + \mathrm{i}t)]$ on Γ_{1,z_*} is strictly increasing if $t > 0$ and strictly decreasing if $t < 0$.*

(iii) *The function $x \mapsto \Re[f(x)]$ is strictly decreasing on Λ_{1,z_*}.*

(iv) *Let $x\in\Lambda_{2,z_*}$ be parametrized as $x=|z_*|-qs+\mathrm{i}s$, for $s\in[0,+\infty)$, with q defined in (48c), then*

$$\Re[f(x)]=-E(|z_*|-qs)+\epsilon s+\frac{\delta(|z_*|-qs)}{s^2+(qs-|z_*|)^2}+\frac{(1-2\delta)[(|z_*|-qs)^2-s^2]}{(s^2+(qs-|z_*|)^2)^2}+\mathcal{O}([s^2+|z_*|^2]^{-3/2}). \quad (51)$$

Despite the fact that saddle point analysis is not useful anymore in this regime, we expect that the main contribution to (28) comes from the regime in the double integral when both x and y are large, i.e., $|x|,|y|\geq N^\rho$. For this purpose we define

$$\tilde\Lambda:=\{x\in\Lambda_{1,z_*}:|x|\leq N^\rho\},\quad \tilde\Gamma:=\{y\in\Gamma_{1,z_*}:|y|\leq N^\rho\}. \quad (52)$$

In the following part of this section we will first prove that the contribution to (28) in the regime when either $x\in\tilde\Lambda$ or $y\in\tilde\Gamma$ is exponentially small and then we explicitly compute the leading term of (28) in the regime

$$(x,y)\in(\Lambda\setminus\tilde\Lambda)\times(\Gamma\setminus\tilde\Gamma).$$

For this purpose, we first prove a bound for the double integral in the regime $y\in\Gamma\setminus\tilde\Gamma$ or $x\in\Lambda\setminus\tilde\Lambda$ in Lemma 5.2 and Lemma 5.3, respectively, and then we conclude the estimate for $x\in\tilde\Lambda$ or $y\in\tilde\Gamma$ in Lemma 5.4. Finally, in Theorem 2.1 we consider the regime $(x,y)\in(\Lambda\setminus\tilde\Lambda)\times(\Gamma\setminus\tilde\Gamma)$ and compute the leading term of (28).

Lemma 5.2. *Let $c(N)$ be defined in (44), $E\lesssim c(N)$, $b\in\mathbb{N}$, and let f be defined in (28), then*

$$\left|\int_{\Gamma\setminus\tilde\Gamma}\frac{e^{Nf(y)}}{y^b}\,\mathrm{d}y\right|\lesssim|z_*|^{1-b}+\begin{cases}|z_*|, & b=0,\\ 1+|\log(N|z_*|^{-2})|, & b=1,\ \delta<|z_*|^{-1},\\ 1+|\log(N\delta|z_*|^{-1})|, & b=1,\ \delta\geq|z_*|^{-1},\\ N^{\frac{1-b}{2}}\wedge(N\delta)^{1-b}, & b\geq 2,\end{cases} \quad (53)$$

where Γ, $\tilde\Gamma$ are defined in (48a) and (52), respectively. Furthermore, we have that

$$\int_{\Gamma\setminus\tilde\Gamma}e^{Nf(y)}\,\mathrm{d}y=\mathcal{O}(N^{1/2}\vee(N\delta)),\qquad \int_{\Gamma\setminus\tilde\Gamma}\frac{e^{Nf(y)}}{y}\,\mathrm{d}y=\mathcal{O}(1). \quad (54)$$

Proof. First, we notice that if $y\in\Gamma\setminus\tilde\Gamma$ then $|y|\geq N^\rho$, hence we expand f as in (47a), i.e.,

$$f(y)=-(E+\mathrm{i}\epsilon)y+\frac{\delta}{y}+\frac{1}{2y^2}+\mathcal{O}(|y|^{-3}+\delta|y|^{-2}). \quad (55)$$

Moreover, by (45) it follows that $|z_*|\sim E^{-1/3}\vee\sqrt{\delta E^{-1}}$, and so that

$$|Nf(y)|\lesssim NE^{2/3}+N\sqrt{\delta E} \quad (56)$$

for any $|y|\sim|z_*|$, $E\lesssim c(N)$. Note that $\Gamma\setminus\tilde\Gamma=(\Gamma_{1,z_*}\setminus\tilde\Gamma)\cup\Gamma_{2,z_*}$, with Γ_{1,z_*}, Γ_{2,z_*} defined in (48a). By (56) it easily follows that $|Nf(y)|\lesssim 1$ for any $y\in\Gamma_{2,z_*}$, which clearly implies that

$$\int_{\Gamma_{2,z_*}}\left|\frac{e^{Nf(y)}}{y^b}\right||\mathrm{d}y|\lesssim|z_*|^{1-b}. \quad (57)$$

To conclude the proof of (53) we bound the integral on $\Gamma_{1,z_*}\setminus\tilde{\Gamma}$. Let $w = E + \mathrm{i}\epsilon$, then in this regime, by (49)–(50), we have

$$\begin{aligned}&\int_{\Gamma_{1,z_*}\setminus\tilde{\Gamma}} \frac{e^{Nf(y)}}{y^b}\,\mathrm{d}y\\ &= -\mathrm{i}\int_{N^\rho}^{\sqrt{|z_*|^2-4/9}} e^{-N[1/(2t^2)+\mathcal{O}(|t|^{-3}+\delta|t|^{-2})]}\times\left(\frac{e^{-N\left[\mathrm{i}wt+\mathrm{i}\frac{\delta}{t}\right]}}{(-2/3+\mathrm{i}t)^b}+\frac{e^{N\left[\mathrm{i}wt+\mathrm{i}\frac{\delta}{t}\right]}}{(-2/3-\mathrm{i}t)^b}\right)(1+\mathcal{O}(NE))\,\mathrm{d}t. \qquad (58)\end{aligned}$$

For any $b\in\mathbb{N}$, we estimate the integral above as follows

$$\left|\int_{\Gamma_{1,z_*}\setminus\tilde{\Gamma}} \frac{e^{Nf(y)}}{y^b}\,\mathrm{d}y\right| \lesssim \left|\int_{N^\rho}^{|z_*|} \frac{e^{-\frac{N}{2t^2}+\frac{\mathrm{i}\delta N}{t}}}{t^b}\,\mathrm{d}t\right| \lesssim \begin{cases} |z_*|, & b=0,\\ 1+|\log(N|z_*|^{-2})|, & b=1,\ \delta<|z_*|^{-1},\\ 1+|\log(N\delta|z_*|^{-1})|, & b=1,\ \delta\ge|z_*|^{-1},\\ N^{\frac{1-b}{2}}\wedge(N\delta)^{1-b}, & b\ge 2.\end{cases} \qquad (59)$$

Note that in (59) for $b=0,1$ we get the bound in the r.h.s. bringing the absolute value inside the integral, whilst this is not affordable to get the bound for $b\ge 2$, since the term $e^{\mathrm{i}\delta N/t}$ has to be used. Indeed, we would get a bound $N^{(1-b)/2}$ for $b\ge 2$ if we estimate the integral in (59) moving the absolute value inside. In the following part of the proof we compute the integral (58) for $b=0,1$ without estimating it by absolute value.

In particular, for $b=1$, we prove that the leading term of the r.h.s. of (58) is $\mathcal{O}(1)$, instead of the overestimate $1+|\log(N|z_*|^{-2})|$ in (59), as a consequence of the symmetry of Γ_{1,z_*} with respect to 0. For this computation we have to distinguish the cases $\delta\gg N^{-1/2}$ and $\delta\lesssim N^{-1/2}$. If $E\sim c(N)$ and $\delta\lesssim N^{-1/2}$, then $N|z_*|^{-2}\sim 1$, hence the bound in (54) directly follows by (57) and (59). We are left with the cases $E\ll c(N)$ and $E\sim c(N)$, $\delta\gg N^{-1/2}$. For $\delta\gg N^{-1/2}$, we have $|z_*|\sim\sqrt{\delta/E}$ and using $|Nwt|\le NE|z_*|\ll 1$, if $E\ll c(N)$, and $|Nwt|\sim 1$, if $E\sim c(N)$, we conclude

$$\int_{N^\rho}^{\sqrt{|z_*|^2-4/9}} \frac{e^{-N\left[1/(2t^2)\pm\mathrm{i}\frac{\delta}{t}\pm\mathrm{i}tw\right]}}{-2/3+\mathrm{i}t}\,\mathrm{d}t = \int_{N^\rho}^{|z_*|}\frac{e^{-N/(2t^2)\pm\mathrm{i}\frac{\delta}{t}}}{t}\,\mathrm{d}t+\mathcal{O}(1) = |\log(N\delta|z_*|^{-1})|+\mathcal{O}(1).$$

Similarly we prove that the integral in the l.h.s. of the above equalities is equal to $|\log(N|z_*|^{-2})|+\mathcal{O}(1)$ if $\delta\ll N^{-1/2}$. Similar calculation holds if the denominator is $\left(-\frac{2}{3}-\mathrm{i}t\right)$ instead of $\left(-\frac{2}{3}+\mathrm{i}t\right)$, just an overall sign changes. Thus the leading terms from the two parts of the integral in (58) cancel each other. We thus conclude the second bound in (54) combining the above computations with (57) and (58).

Next, we compute the integral of $e^{Nf(y)}$ on $\Gamma\setminus\tilde{\Gamma}$, i.e., we prove the first bound in (54). We consider only the regime $E\ll c(N)$, since in the regime $E\sim c(N)$ the bound in (54) follows directly by (57), (59), and the definition of $c(N)$ in (44), since $|z_*|\sim N^{1/2}\vee N\delta$. On Γ_{2,z_*}, using the parametrization $y=|z_*|e^{\mathrm{i}\psi}$, and that by (56) we have $|Nf(y)|\ll 1$ for $E\ll c(N)$, we Taylor expand $e^{Nf(y)}$ and conclude that

$$\int_{\Gamma_{2,z_*}} e^{Nf(y)}\,\mathrm{d}y = 2|z_*|\mathrm{i}\cdot\left[1+\mathcal{O}(NE|z_*|+\delta N|_*|^{-1}+N|z_*|^{-2})\right]. \qquad (60)$$

Furthermore, by (58) for $b = 0$, using that $E \ll c(N)$ and so that $|Nwt| \ll 1$ on Γ_{1,z_*}, we have

$$\int_{\Gamma_{1,z_*}\setminus\tilde{\Gamma}} e^{Nf(y)}\,\mathrm{d}y = -2\mathrm{i}|z_*| + \mathcal{O}(N^{1/2} \vee N\delta).$$

The minus sign is due to the counter clockwise orientation of Γ, i.e., the vertical line Γ_{1,z_*} is parametrized from the top to the bottom. Combining this computation with (60) and using that

$$NE|z_*|^2 + N\delta + N|z_*|^{-1} \ll N^{1/2} + N\delta,$$

since $|z_*| \sim E^{-1/3} + \sqrt{\delta E^{-1}}$ by (45), we conclude the proof of this lemma. □

Lemma 5.3. *Let $c(N)$ be defined in (44), $E \lesssim c(N)$ and let f be defined in (28) and $a \in \mathbb{R}$, then the following bound holds true*:

$$\int_{\Lambda\setminus\tilde{\Lambda}} \left|\frac{e^{-Nf(x)}}{x^a}\right| |\mathrm{d}x| \lesssim \begin{cases} |z_*|^{1-a} + (NE)^{a-1}, & a < 1, \\ 1 + |\log(N|z_*|^{-2})|, & a = 1, \delta < |z_*|^{-1}, \\ 1 + |\log(N\delta|z_*|^{-1})|, & a = 1, \delta \geq |z_*|^{-1}, \\ N^{\frac{1-a}{2}} \wedge (N\delta)^{1-a}, & a > 1, \end{cases} \tag{61}$$

where Λ, $\tilde{\Lambda}$ are defined in (48b) and (52).

Proof. We split the computation of the integral of $e^{Nf(x)}x^{-a}$ as the sum of the integral over $\Lambda_{1,z_*} \setminus \tilde{\Lambda}$ and Λ_{2,z_*}. Using the parametrization $x = |z_*| - qs + \mathrm{i}s$, with $s \in [0, +\infty)$ and q defined in (48c), by (51), we estimate the integral over Λ_{2,z_*} as follows:

$$\int_{\Lambda_{2,z_*}} \left|\frac{e^{-Nf(x)}}{x^a}\right| |\mathrm{d}x| \lesssim \int_0^{+\infty} \frac{e^{-N\left[-E(|z_*|-qs) + \frac{\delta(|z_*|-qs)}{(|z_*|-qs)^2+s^2} + \frac{(|z_*|-qs)^2-s^2}{2[(|z_*|-qs)^2+s^2]^2}\right]}}{[(|z_*|-qs)^2+s^2]^{a/2}}\,\mathrm{d}s. \tag{62}$$

We split the computation of the integral in the r.h.s. of (62) into two parts:

$$|s| \in [0, |z_*|) \quad \text{and} \quad s \in [|z_*|, +\infty).$$

Since $q \sim 1$ and $NE|z_*| \lesssim 1$, in the regime $|s| \in [0, |z_*|)$ we estimate the integral in the r.h.s. of (62) as

$$e^{NE|z_*|} \int_0^{|z_*|} \frac{e^{-\frac{N\delta}{|z_*|} - \frac{N}{|z_*|^2}}}{|z_*|^a}\,\mathrm{d}s \lesssim |z_*|^{1-a}. \tag{63}$$

In the regime $s \in [|z_*|, +\infty)$, instead, we have

$$e^{NE|z_*|} \int_{|z_*|}^{+\infty} \frac{e^{-NEqs}}{s^a}\,\mathrm{d}s \lesssim \begin{cases} (NE)^{a-1}, & a < 1, \\ 1 + |\log(NE|z_*|)|, & a = 1, \\ |z_*|^{1-a}, & a > 1. \end{cases} \tag{64}$$

We are left with the estimate of the integral over $\Lambda_{1,z_*} \setminus \tilde{\Lambda}$. As for the bound in (59), using that $\Lambda_{1,z_*} \setminus \tilde{\Lambda} = [N^\rho, |z_*|)$ and that $NE|z_*| \lesssim 1$, we have that

$$\int_{\Lambda_{1,z_*} \setminus \tilde{\Lambda}} \left| \frac{e^{-Nf(x)}}{x^a} \right| |\mathrm{d}x| \lesssim \int_{N^\rho}^{|z_*|} \frac{e^{-\frac{N\delta}{s} - \frac{N}{2s^2}}}{s^a} \, \mathrm{d}s \lesssim \begin{cases} |z_*|^{1-a}, & a < 1, \\ 1 + |\log(N|z_*|^{-2})|, & a = 1, \delta < |z_*|^{-1}, \\ 1 + |\log(N\delta|z_*|^{-1})|, & a = 1, \delta \geq |z_*|^{-1}, \\ N^{\frac{1-a}{2}} \wedge (N\delta)^{1-a}, & a > 1. \end{cases} \tag{65}$$

Combining (62)–(65) we conclude the proof of (61). □

Using Lemmas 5.2 and 5.3 in the following lemma we prove that the contribution to (28) in the regime where either $x \in \tilde{\Lambda}$ or $y \in \tilde{\Gamma}$ is exponentially small.

Lemma 5.4. *Let $c(N)$ be defined in (44), and let f, G be defined in (28), then, as $\epsilon \to 0^+$, for any $E \lesssim c(N)$ we have that*

$$\left| \left(\int_\Lambda \mathrm{d}x \int_\Gamma \mathrm{d}y - \int_{\Lambda \setminus \tilde{\Lambda}} \mathrm{d}x \int_{\Gamma \setminus \tilde{\Gamma}} \mathrm{d}y \right) [e^{N[f(y) - f(x)]} y G(x,y)] \right|$$
$$\lesssim N^\rho (N^{1/2} + N\delta + |\log(NE^{2/3})|) e^{-\frac{1}{2}N^{1-2\rho}}. \tag{66}$$

Proof. We split the estimate of the integral over $(\Lambda \times \Gamma) \setminus [(\Lambda \setminus \tilde{\Lambda}) \times (\Gamma \setminus \tilde{\Gamma})]$ into three regimes: $(x,y) \in \tilde{\Lambda} \times \tilde{\Gamma}$, $(x,y) \in \tilde{\Lambda} \times (\Gamma \setminus \tilde{\Gamma})$, $(x,y) \in (\Lambda \setminus \tilde{\Lambda}) \times \tilde{\Gamma}$. By (ii), (iii) of Lemma 5.1, in the regimes $y \in \tilde{\Gamma}$ and $x \in \tilde{\Lambda}$, respectively, it follows that the function f attains its maximum at $y = -2/3 \pm \mathrm{i}N^\rho$ on $\tilde{\Gamma}$ and f attains its minimum at $x = N^\rho$ on $\tilde{\Lambda}$. Hence, by the expansion in (47a) it follows that

$$\sup_{y \in \tilde{\Gamma}} |e^{Nf(y)}| + \sup_{x \in \tilde{\Gamma}} |e^{-Nf(x)}| \lesssim e^{-Nf(N^\rho)}, \tag{67}$$

with

$$f(N^\rho) = \frac{\delta}{N^\rho} + \frac{1}{2N^{2\rho}} + \mathcal{O}(N^{-3\rho} + \delta N^{-2\rho}). \tag{68}$$

Then, by (67) and (68), it follows that the integral over $(x,y) \in \tilde{\Lambda} \times \tilde{\Gamma}$ is bounded by $N^{2\rho} e^{-N^{1-2\rho}}$. Note that in the regimes $(x,y) \in \tilde{\Lambda} \times (\Gamma \setminus \tilde{\Gamma})$ and $(x,y) \in (\Lambda \setminus \tilde{\Lambda}) \times \tilde{\Gamma}$ one among $|x|$ and $|y|$ is bigger than N^ρ. Hence, expanding (28) for large x or y argument, using Lemmas 5.2 and 5.3 to estimate the regimes $x \in \tilde{\Lambda}$ and $y \in \tilde{\Gamma}$, respectively, by (67)–(68), we conclude that the integral over $(x,y) \in \tilde{\Lambda} \times (\Gamma \setminus \tilde{\Gamma})$ is bounded by $N^\rho(N^{1/2} + (N\delta))e^{-N^{1-2\rho}/2}$, and that the one over $(x,y) \in (\Lambda \setminus \tilde{\Lambda}) \times \tilde{\Gamma}$ is bounded by $N^\rho(1 + |\log(NE^{2/3})|)e^{-N^{1-2\rho}/2}$. □

Next, we compute the leading term of (28). We define $\tilde{z}_*$ as

$$\tilde{z}_*(\lambda, \tilde{\delta}) := N^{-1/2} |z_*(\lambda c(N), N^{-1/2}\tilde{\delta})|, \qquad \lambda = \frac{E}{c(N)}, \quad \tilde{\delta} = N^{1/2}\delta, \tag{69}$$

where we also recalled the rescaled parameters λ and $\tilde{\delta}$ from (12). Note that $\tilde{z}_*(\lambda, \tilde{\delta})$ is N-independent, indeed all N factors scale out by using the definition of $z_*(E, \delta)$ from (45). Since $E \ll 1$ and $\delta \in [0,1]$, by (69) it follows that the range of the new parameters is $\tilde{\delta} \leq N^{1/2}$ and $\lambda \ll c(N)^{-1}$.

We are now ready to prove our main result on the leading term of (28), denoted by $q_{\tilde{\delta}}(\lambda)$, in the complex case, Theorem 2.1. Then, the one point function of Y is asymptotically given by $p_{\tilde{\delta}}(\lambda) := \Im[q_{\tilde{\delta}}(\lambda)]$. The main inputs for the proof are the bounds in (53) and (61) that will be used to estimate the error terms in the expansions for large arguments of f and G in (47a) and (47b).

Proof of Theorem 2.1 in the case $\delta \geq 0$. By (28) and Lemma 5.4 it follows that

$$\boldsymbol{E}\,\mathrm{Tr}[Y-w]^{-1} = \frac{N^2}{2\pi \mathrm{i}}\int_{\Lambda\setminus\tilde{\Lambda}} \mathrm{d}x \int_{\Gamma\setminus\tilde{\Gamma}} \mathrm{d}y\, e^{-Nf(x)+Nf(y)} y\cdot G(x,y) \\ + \mathcal{O}\big(N^{\rho}(N^{1/2}+N\delta+|\log(NE^{2/3})|)e^{-\frac{1}{2}N^{1-2\rho}}\big). \quad (70)$$

Note that $|x|, |y| \geq N^{\rho}$ for any $(x,y)\in(\Lambda\setminus\tilde{\Lambda})\times(\Gamma\setminus\tilde{\Gamma})$. In order to prove (13a) we first estimate the error terms in the expansions of f and G in (47a)–(47b) and then in order to get an N-independent double integral we rescale the phase function by $|z_*|$. By Lemma 5.2 and Lemma 5.3, using that $|\log(N|z_*|^{-2})| + |\log(N\delta|z_*|^{-1})| \lesssim |\log(NE^{2/3})|$ by the definition of z_* in (45), it follows that

$$\left|\int_{\Lambda\setminus\tilde{\Lambda}} \mathrm{d}x \int_{\Gamma\setminus\tilde{\Gamma}} \mathrm{d}y \frac{e^{-Nf(x)+Nf(y)}}{x^a y^b}\right| \lesssim \begin{cases} N^{\frac{2-d}{2}}(1\wedge\tilde{\delta}^{1-d})(1\vee\tilde{\delta}), & b=0, \\ N^{\frac{1-b}{2}}(1\wedge\tilde{\delta}^{1-b})(1+|\log(NE^{2/3})|), & a=1,\, b\geq 1, \\ N^{\frac{2-d}{2}}(1\wedge\tilde{\delta}^{2-d}), & a>1,\, b\geq 1, \end{cases} \quad (71)$$

for any $a\geq 1$, $b\in\mathbb{N}$, where $d := a+b$. In order to get the bound in the r.h.s. of (71) we estimated the terms with $b=0$ and $b=1$ using the improved bound in (54), all the other terms are estimated by absolute value. Note that for $\lambda \ll 1$ the bound in (71) and the definition of λ in (69) imply that

$$\lim_{\epsilon\to 0^+} |\boldsymbol{E}\,\mathrm{Tr}[Y-(E+\mathrm{i}\epsilon)]^{-1}| \lesssim N^{3/2}(1\vee\tilde{\delta}) \begin{cases} |\log\lambda|, & \lambda\geq\tilde{\delta}^3, \\ |\log\lambda\tilde{\delta}|, & \lambda<\tilde{\delta}^3, \end{cases}$$

if $\lambda \ll 1$, since the leading term in the expansion of $yG(x,y)$ in (47b) consists of monomials of the form $x^{-a}y^{-b}$, with $a+b=3$, and $\delta x^{-a}y^{-b}$, with $a+b=2$. This concludes the proof of (13c).

Now we prove the more precise asymptotics (13a). We will replace the functions f and G in (70) by their leading order approximations, denoted by g and H from (47a) and (47b). The error of this replacement in the phase function f is estimated by the Taylor expanding the exponent $e^{\mathcal{O}^{\#}(x^{-3}+\delta x^{-2})}$, with $\mathcal{O}^{\#}(x^{-3}+\delta x^{-2})$ defined in (46). Hence, by (47a) and (47b) and the bound in (71), as $\epsilon\to 0^+$, we conclude that

$$\boldsymbol{E}\,\mathrm{Tr}[Y-w]^{-1} \\ = \frac{N^2}{2\pi \mathrm{i}}\int_{\Lambda\setminus\tilde{\Lambda}} \mathrm{d}x \int_{\Gamma\setminus\tilde{\Gamma}} \mathrm{d}y\, e^{-Ng(x)+Ng(y)} H(x,y) + \mathcal{O}\big((N+N^{3/2}\delta)[1+|\log(NE^{2/3})|]\big). \quad (72)$$

The error estimates in (72) come from terms with $d\geq 4$ or terms with $d\geq 3$ multiplied by δ in (71).

We recall that $\lambda = Ec(N)^{-1}$, $\tilde{\delta} = \delta N^{1/2}$, and that $|z_*| = N^{1/2}\tilde{z}_*(\lambda,\tilde{\delta})$. Then, defining the contours

$$\widehat{\Gamma} := |z_*|^{-1}\Gamma, \quad \widehat{\Lambda} := |z_*|^{-1}\Lambda, \quad (73)$$

and using the change of variables $x \to x|z_*|$, $y \to y|z_*|$ in the leading term of (72) we conclude that

$$\boldsymbol{E}\,\mathrm{Tr}[Y-w]^{-1} = \frac{N^{3/2}\tilde{z}_*(\lambda,\tilde{\delta})^{-1}}{2\pi \mathrm{i}} \int_{\hat{\Gamma}} \mathrm{d}y \int_{\hat{\Lambda}} \mathrm{d}x e^{h_{\lambda,\tilde{\delta}}(y)-h_{\lambda,\tilde{\delta}}(x)} \tilde{H}_{\lambda,\tilde{\delta}}(x,y) + \mathcal{O}(N(1\vee\tilde{\delta})[1+|\log\lambda|]) \quad (74)$$

with $h_{\lambda,\tilde{\delta}}(x)$ and $\tilde{H}_{\lambda,\tilde{\delta}}(x,y)$ defined in (13b). Note that in order to get (74) we used that the integral in the regime when either

$$x \in [0, N^{\rho}|z_*|^{-1}] \ \text{ or } \ y \in \left[-2|z_*|^{-1}/3, -2|z_*|^{-1}/3 + \mathrm{i}N^{\rho}|z_*|^{-1}\right]$$

is exponentially small. Moreover, since by holomorphicity we can deform the contour $\widehat{\Lambda}$ to any contour, which does not cross -1, from 0 to $e^{3\mathrm{i}\pi/4}\infty$ and we can deform the contour $\widehat{\Gamma}$ as long as it does not cross 0, (74) concludes the proof of Theorem 2.1. □

5B. ***Case $\delta < 0$, $|\delta| \lesssim N^{-1/2}$.*** We now explain the necessary changes in the case $\delta < 0$. Throughout this section we assume that $E \lesssim N^{-3/2}$. Let x_* be the stationary point of f defined in (41a); that is, the point around where the main contribution to (28) comes from in the saddle point regime for $\delta < 0$. Then, at leading order, x_* is given by $3|\delta|^{-1}/2$ if $E \ll |\delta|^3$, by $e^{\pi\mathrm{i}/3}E^{-1/3}$ if $E \gg |\delta|^3$, and by $\mu(c)e^{\pi\mathrm{i}/3}E^{-1/3}$ if $E = c|\delta|^3$, for some function $\mu(c) > 0$ for any fixed constant $c > 0$ independent of N, E and δ.

This regime can be treated similarly to the regime $0 \le \delta \le 1$, since for $|\delta| \lesssim N^{-1/2}$ the term δx^{-1} in the expansion of f, for $|x| \gg 1$, does not play any role in the bounds of Lemma 5.2, Lemma 5.3. Indeed, instead of the deforming the contours Γ and Λ through the leading term of the stationary point x_*, we deform Γ and Λ through $z_* := e^{\pi\mathrm{i}/3}E^{-1/3}$ as

$$\Gamma = \Gamma_{z_*} := \Gamma_{1,z_*} \cup \Gamma_{2,z_*} \quad \text{and} \quad \Lambda = \Lambda_{z_*} := \Lambda_{1,z_*} \cup \Lambda_{2,z_*},$$

where $\Gamma_{1,z_*}, \Gamma_{2,z_*}$ and $\Lambda_{1,z_*}, \Lambda_{2,z_*}$ are defined in (48a) and (48b), respectively. We could have made the same choice in the case $0 \le \delta \lesssim N^{-1/2}$, but not for the regime $N^{-1/2} \ll \delta \le 1$, hence, to treat both the regimes in the same way, in Section 5A we deformed the contours trough (45). Note that z_* defined here is not the analogue of (45), since in all cases

$$z_* = e^{\pi\mathrm{i}/3E^{-1/3}}.$$

The fact that $E \ll 1$ implies that $|z_*| \gg 1$, hence, like in the case $0 \le \delta \le 1$, we expect that the main contribution to (28) comes from the regime when $|x|, |y| \ge N^{\rho}$, for some small $0 < \rho < 1/2$. Hence, in order to compute the leading term of (28) we expand f and G for large arguments as in (47a) and (47b).

The phase function f defined in (28) satisfies the properties (i), (ii) and (iv) of Lemma 5.1, but (iii) does not hold true for $\delta < 0$ if $E \ll |\delta|^3$. Instead, it is easy to see that the following lemma holds true.

Lemma 5.5. *Let f be the phase function defined in (28), then, as $\epsilon \to 0^+$, the function $x \mapsto \Re[f(x)]$ has a unique global minimum on Λ_{1,z_*} at $x = 3|\delta|^{-1}/2$ if $E \ll |\delta|^3$.*

Note that, since $|\delta| \lesssim N^{-1/2}$, by Lemma 5.5 it follows that the function $x \mapsto \Re[f(x)]$ is strictly decreasing for $0 \le x \ll N^{-1/2}$.

Proof of Theorem 2.1 for $-CN^{-1/2} \le \delta < 0$. Let $\tilde{\Gamma}$, $\tilde{\Lambda}$ be defined in (52), then using that

$$e^{-N\left[\frac{\delta}{s}+\frac{1}{2s^2}\right]} \lesssim e^{-\frac{N}{4s^2}},$$

for $s \in [N^\rho, |z_*|]$ and $|\delta| \lesssim N^{-1/2}$, Lemma 5.3 and the improved bounds in (54) of Lemma 5.2, for $b = 0, 1$, we conclude the bound in the following lemma exactly as in (71) without the improvement involving $\tilde{\delta}$.

Lemma 5.6. *Let $E \lesssim N^{-3/2}$, and let f be defined in (28).Then, the bound*

$$\left|\int_{\Lambda\setminus\tilde{\Lambda}} \mathrm{d}x \int_{\Gamma\setminus\tilde{\Gamma}} \mathrm{d}y \frac{e^{-Nf(x)+Nf(y)}}{x^a y^b}\right| \lesssim \begin{cases} N^{\frac{2-d}{2}}, & b = 0, \\ N^{\frac{1-b}{2}}(1+|\log(NE^{2/3})|), & a = 1,\ b \ge 1, \\ N^{\frac{2-d}{2}}, & a > 1,\ b \ge 1, \end{cases}$$

holds for any $a \ge 1$ and $b \in \mathbb{N}$, where $d := a + b$.

Then, much as in the case $0 \le \delta \le 1$, by Lemma 5.6 we conclude that the contribution to (28) of the regime when either $x \in \tilde{\Lambda}$ or $y \in \tilde{\Gamma}$ is exponentially small, i.e., Lemma 5.4 holds true. Hence, by (28), Lemma 5.6 and the expansion of G in (47b), we easily conclude Theorem 2.1 also in the regime $-CN^{-1/2} \le \delta < 0$. □

6. The real case below the saddle point regime

In this section we prove Theorem 2.3. Throughout this section we always assume that $\Re w < 0$, hence to make our notation easier we define $w = -E + \mathrm{i}\epsilon$ with some $E > 0$ and $\epsilon > 0$. Moreover, we always assume that $E \lesssim c(N)$, with $c(N)$ defined in (44). We are interested in estimating (34) in the transitional regime of $|z|$ around one. For this purpose we introduce the parameter $\delta = \delta_z := 1 - |z|^2$. In order to have an optimal estimate of the leading order term of (34) it is not affordable to estimate the error terms in the expansions, for large a and ξ, of f, $g(\cdot, 1, \eta)$ and $G_{1,N}, G_{2,N}$ by absolute value. For this reason, to compute the error terms in the expansions of f, $g(\cdot, 0, \eta), G_{1,N}, G_{2,N}$ we use a notation $\mathcal{O}^\#(\cdot)$ similar to the one introduced in (46). In order to keep track of the power of τ in the expansion of $G_{1,N}$ and $G_{2,N}$, we define the set of functions

$$\mathcal{O}^\#((a,\tau,\xi)^{-1}) := \left\{ h : \mathbb{C}\times[0,1]\times\mathbb{C} \to \mathbb{C} : h(a,\tau,\xi) = \sum_{\substack{\alpha+\beta\ge 1,\\ 0\le\gamma\le\alpha}} \frac{c_{\alpha,\beta,\gamma}}{a^\alpha \tau^\gamma \xi^\beta} \text{ with } |c_{\alpha,\beta,\gamma}| \le C^{\alpha+\beta} \text{ if } |a|, |a\tau|, |\xi| \ge 2C \right\},$$

for some constant $C > 0$ implicit in the $\mathcal{O}^\#(\cdot)$ notation. The exponents α, β are nonnegative integers.

We expand the functions $f, g(\cdot, 1, \eta), G_{1,N}, G_{2,N}$ for large a and ξ arguments as

$$\begin{aligned} f(\xi) &= (E - \mathrm{i}\epsilon)\xi + \frac{\delta}{\xi} + \frac{1}{2\xi^2} + \mathcal{O}^\#(\xi^{-3} + \delta\xi^{-2}), \\ g(a, 1, \eta) &= (E - \mathrm{i}\epsilon)a + \frac{\delta}{a} + \frac{1}{2a^2} + \mathcal{O}(a^{-3} + \delta a^{-2}), \end{aligned} \tag{75}$$

with $\mathcal{O}^{\#}(\cdot)$ defined as in (46), and

$$
\begin{aligned}
G_{1,N}(a,\tau,\xi,|z|) = \Bigg[& \sum_{\substack{\alpha,\beta\geq 2,\,\alpha+\beta=8,\\ \gamma=\min\{\alpha-1,3\}}} \frac{c_{\alpha,\beta,\gamma}N^2}{a^\alpha \tau^\gamma \xi^\beta} + \sum_{\substack{\alpha,\beta\geq 2,\,\alpha+\beta=7,\\ \gamma=\min\{\alpha-1,3\}}} \frac{c_{\alpha,\beta,\gamma}N^2\delta}{a^\alpha \tau^\gamma \xi^\beta} \\
& + \sum_{\substack{\alpha,\beta\geq 2,\,\alpha+\beta=6,\\ \gamma=\min\{\alpha-1,2\}}} \frac{c_{\alpha,\beta,\gamma}N}{a^\alpha \tau^\gamma \xi^\beta} + \sum_{\substack{\alpha,\beta\geq 2,\alpha+\beta=6\\ \gamma=\min\{\alpha-1,2\}}} \frac{c_{\alpha,\beta,\gamma}N^2\delta^2}{a^\alpha \tau^\gamma \xi^\beta} \Bigg] \\
& \times \big[1+\mathcal{O}^{\#}((a,\tau,\xi)^{-1})\big],
\end{aligned} \tag{76}
$$

$$
\begin{aligned}
G_{2,N}(a,\tau,\xi,z) = \Bigg[& \sum_{\substack{\alpha,\beta\geq 2,\,\alpha+\beta=6,\\ \gamma=\max\{\alpha-1,2\}}} \frac{c_{\alpha,\beta,\gamma}N^2\eta^2}{a^\alpha \tau^\gamma \xi^\beta} \\
& + \sum_{\substack{\alpha,\beta\geq 2,\alpha+\beta=5,\\ \gamma=\max\{\alpha-1,2\}}} \frac{c_{\alpha,\beta,\gamma}N^2\eta^2\delta}{a^\alpha \tau^\gamma \xi^\beta} + \sum_{\substack{\alpha,\beta=2,3,\,\alpha+\beta=5\\ \gamma=\max\{\alpha-1,2\}}} \frac{c_{\alpha,\beta,\gamma}N\eta^2}{a^\alpha \tau^\gamma \xi^\beta} \Bigg] \\
& \times \big[1+\mathcal{O}^{\#}((a,\tau,\xi)^{-1})\big],
\end{aligned} \tag{77}
$$

where $c_{\alpha,\beta,\gamma}\in\mathbb{R}$ is a constant that may change term by term. To make our notation easier in (76)–(77) we used the convention to write a common multiplicative error for all the terms, even if in principle the constants in the series expansion of the error terms differ term by term.

In the following we deform the integration contours in (34) with the following constraints: the ξ contour can be freely deformed as long as it does not cross 0 and -1, the a-contour can be deformed as long as it goes out from zero in the region $\Re[a] > |\Im[a]|$, it ends in the region $\Re[a] > 0$, and it does not cross 0 and -1 along the deformation. The τ-contour will not be deformed.

Specifically, we deform the a-contour in (34) to $\Lambda = [0,+\infty)$, and we can deform the ξ-contour to any contour around 0 not encircling -1 (this contour is denoted by Γ in (78)). Moreover, since for $a\in\mathbb{R}_+$ we have

$$
|e^{-N(E-\mathrm{i}\epsilon)a}| = e^{-NEa}
$$

and since the factor e^{-NEa} makes the integral convergent, we may pass to the limit $\epsilon\to 0^+$. Hence, for any $E>0$ we conclude that

$$
\boldsymbol{E}\,\mathrm{Tr}[Y+E]^{-1} = \frac{N}{4\pi\mathrm{i}}\int_\Gamma \mathrm{d}\xi \int_0^{+\infty}\mathrm{d}a \int_0^1 \mathrm{d}s\, \frac{\xi^2 a}{\tau^{1/2}} e^{N[f(\xi)-g(a,\tau,\eta)]} G_N(a,\tau,\xi,z). \tag{78}
$$

We split the computation of the leading order term of (78) into the cases $-CN^{-1/2}\leq\delta<0$ and $\delta\geq 0$.

6A. ***Case*** $\boldsymbol{0\leq\delta\leq 1}$. In order to estimate the leading term of (78) we compute the ξ-integral and the (a,τ)-integral separately. In particular, we compute the (a,τ)-integral first, performing the τ-integral for any fixed a, and then we compute the a-integral. Note that, since $E'=-E$ is negative, the relevant stationary point of $f(\xi)$ is real and its leading order ξ_* is given by

$$\xi_* := \begin{cases} \sqrt{\frac{\delta}{E}}, & E \ll \delta^3, \\ \mu(c)E^{-1/3}, & E = c\delta^3, \\ E^{-1/3}, & E \gg \delta^3, \end{cases} \tag{79}$$

for some $\mu(c) > 0$ and any fixed constant $c > 0$ independent of N, E and δ. Note that $\xi_* \gg 1$ for any $E \lesssim c(N)$, $0 \le \delta \le 1$. We will show that in the regime $a \in [N^\rho, +\infty)$ the τ-integral is concentrated around 1 as long as $|e^{-N[g(a,\tau,\eta)-g(a,1,\eta)]}|$ is effective, i.e., as long as $N|g(a,\tau,\eta) - g(a,1,\eta)| \gg 1$, and that it is concentrated around 0 if $N|g(a,\tau,\eta) - g(a,1,\eta)| \lesssim 1$. For this purpose, using that $g(a,1,\eta) = f(a)$ for any $a \in \mathbb{C}$, we rewrite (78) as

$$\boldsymbol{E}\,\mathrm{Tr}[Y+E]^{-1} = \frac{N}{4\pi\mathrm{i}}\int_\Gamma \mathrm{d}\xi e^{Nf(\xi)}\xi^2 \int_0^{+\infty} \mathrm{d}a e^{-Nf(a)} a \int_0^1 \mathrm{d}\tau \frac{e^{-N[g(a,\tau,\eta)-g(a,1,\eta)]}}{\tau^{1/2}} G_N, \tag{80}$$

where we used that by holomorphicity, we can deform the contour Γ as $\Gamma = \Gamma_{\xi_*} := \Gamma_{1,\xi_*} \cup \Gamma_{2,\xi_*}$, with $\Gamma_{1,\xi_*}, \Gamma_{2,\xi_*}$ defined in (48a) replacing z_* by ξ_*. We will show that the contribution to (80) of the integrals in the regime when either $|a| \le N^\rho$ or $|\xi| \le N^\rho$, for some small fixed $0 < \rho < 1/2$, is exponentially small. Moreover, we will show that also the τ-integral is exponentially small for τ very close to 0 because of the term $\log\tau$ in the phase function $g(a,\tau,\eta)$. Hence, we define $\tilde{\Gamma}$, $\tilde{\Lambda}$ as in (52), and $I \subset [0,1]$ as $I = I_a := [0, N^{\rho/2}a^{-1}]$, for any $a \in [0, +\infty)$.

In order to compute the leading term of (80) we first bound the integral in the regime

$$(a,\tau,\xi) \in (\Lambda \setminus \tilde{\Lambda}) \times ([0,1] \setminus I) \times (\Gamma \setminus \tilde{\Gamma}),$$

with $\tilde{\Lambda}$ and $\tilde{\Gamma}$ defined in (52), that is the regime where we expect that the main contribution comes from, and then we use these bounds to first prove that the integral in the regime when either $|\xi| \le N^\rho$ or $|a| \le N^\rho$ is exponentially small for any $\tau \in [0,1]$, and then prove that also the τ-integral on I is exponentially small if $|a| \ge N^\rho$. The bounds for the ξ-integral over $\Gamma \setminus \tilde{\Gamma}$ are exactly the same as in Lemma 5.2, since the phase function $f(\xi)$ and the Γ-contour are exactly the same as the complex case. In order to estimate the integral over $(a,\tau) \in (\Lambda \setminus \tilde{\Lambda}) \times ([0,1] \setminus I)$, we start with the estimate of the τ-integral over $[0,1] \setminus I$ in Lemma 6.2 and then we will conclude the computation of the a-integral over $[N^\rho, +\infty)$ in Lemma 6.3.

Before proceeding with the bounds for large $|a|$, $|\xi|$, in the following lemma we state some properties of the functions f and g. The proof of this lemma follows by elementary computations. From now on, for simplicity, we assume that $\eta \ge 0$; the case $\eta < 0$ is completely analogous since the functions g and G_N in (78) depend only on η^2 and $|z|^2$.

Lemma 6.1. *Let f and g be the phase functions defined in* (35) *and* (36), *respectively, then the following properties hold true*:

(i) *For any $\xi = -2/3 + \mathrm{i}t \in \Gamma_{1,\xi_*}$, we have that*

$$\Re[f(-2/3+\mathrm{i}t)] = \frac{2E}{3} + \epsilon t - \frac{1}{2t^2} + \mathcal{O}(|t|^{-3} + \delta|t|^{-2}), \tag{81}$$

and

$$\Im[f(-2/3+\mathrm{i}t)] = \frac{2\epsilon}{3} - Et - \frac{\delta}{t} + \mathcal{O}(|t|^{-3}). \tag{82}$$

(ii) *For $\epsilon = 0$, the function $t \mapsto \Re[f(-2/3+\mathrm{i}t)]$ on Γ_{1,ξ_*} is strictly increasing if $t > 0$ and strictly decreasing if $t < 0$.*

(iii) *For any $a \in [0,+\infty)$, we have that $g(a,\tau,\eta) \geq g(a,\tau,0)$ and the function $\tau \mapsto g(a,\tau,0)$ is strictly decreasing on $[0,1]$.*

(iv) *The function $a \mapsto g(a,1,\eta)$ is strictly decreasing on $[0,\xi_*/2]$.*

Lemma 6.2. *Let $\rho > 0$ be sufficiently small, $I = I_a = [0, N^{\rho/2}a^{-1}]$, $\gamma \in \mathbb{N}$, $\gamma \geq 1$, $c(N)$ be defined in (44), $E \lesssim c(N)$, $0 \leq \delta \leq 1$, and let g be defined as in (36). Then, for any $a \in [N^\rho, +\infty)$, we have*

$$\begin{aligned}\int_{[0,1]\setminus I_a} &\frac{|e^{-N[g(a,\tau,\eta)-g(a,1,\eta)]}|}{\tau^{\gamma+1/2}}\,\mathrm{d}\tau\\ &\lesssim F(a) := \begin{cases} a^2N^{-1}\wedge 1, & N^\rho \leq a \leq \delta^{-1}\wedge\eta^{-1},\\ a(N\delta)^{-1}\wedge 1, & \delta^{-1}\vee N^\rho \leq a \leq \delta\eta^{-2},\\ (N\eta^2)^{-1}\wedge 1, & a \geq (\delta\eta^{-1}\vee 1)\eta^{-1},\end{cases}\\ &\qquad + e^{-\frac{1}{2}N\eta^2} \times \begin{cases} e^{-\frac{N(\delta\vee N^{-1/2})}{a}}, & N^\rho \leq a \leq N(\delta\vee N^{-1/2}),\\ (a(N\delta\vee\sqrt{N})^{-1})^{\gamma-1/2}, & N(\delta\vee N^{-1/2}) \leq a \leq (\delta\vee N^{-1/2})\eta^{-2},\\ (N\eta^2)^{1/2-\gamma}, & a \geq (\delta\vee N^{-1/2})\eta^{-2},\end{cases}\end{aligned} \tag{83}$$

where some regimes in (83) might be empty for certain values of δ and η.

Proof. In order to estimate the integral in the l.h.s. of (83) we first compute the expansion

$$\begin{aligned}g(a,\tau,\eta)-g(a,1,\eta) = \frac{(1-2\delta)(1-\tau)-(1-\tau)^2}{a^2\tau^2} + \frac{2\eta^2(1-\tau)}{\tau} + \frac{(1-\tau)\delta}{a\tau}\\ + \mathcal{O}\left(\frac{1-\tau}{a^2\tau} + \frac{(1-\tau)(\delta+a^{-1})}{a^2\tau^2} + \frac{\eta^2(1-\tau)}{a\tau^2}\right),\end{aligned} \tag{84}$$

which holds true for any $\tau \in [0,1]\setminus I = [N^{\rho/2}a^{-1}, 1]$. Note that by (84) it follows that for any $(a,\tau) \in [N^\rho,+\infty)\times[N^{\rho/2}a^{-1},1]$,

$$g(a,\tau,\eta)-g(a,1,\eta) \geq \frac{1-\tau}{2}\left[\frac{1}{a^2\tau^2} + \frac{\delta}{a\tau} + \frac{\eta^2}{\tau}\right].$$

Then, by (84) it follows that

$$\int_{[0,1]\setminus I} \frac{|e^{-N[g(a,\tau,\eta)-g(a,1,\eta)]}|}{\tau^{\gamma+1/2}}\,\mathrm{d}\tau \lesssim \int_{N^{\rho/2}a^{-1}}^1 \frac{e^{-(1-\tau)\frac{N}{2}\left[\frac{1}{a^2\tau^2}+\frac{\delta}{a\tau}+\frac{\eta^2}{\tau}\right]}}{\tau^{\gamma+1/2}}\,\mathrm{d}\tau. \tag{85}$$

In order to bound the r.h.s. of (85) we split the computations into two cases: $\delta \leq N^{-1/2}$ and $\delta > N^{-1/2}$. We first consider the case $\delta > N^{-1/2}$. In order to prove the bound in the r.h.s. of (83) we further split the computation of the τ-integral into the regimes $\tau \in [N^{\rho/2}a^{-1}, 1/2]$ and $\tau \in [1/2, 1]$. We start estimating the integral over $[1/2, 1]$ as follows:

$$\begin{aligned}\int_{1/2}^{1} \frac{e^{-(1-\tau)\frac{N}{2}\left[\frac{1}{a^2\tau^2}+\frac{\delta}{a\tau}+\frac{\eta^2}{\tau}\right]}}{\tau^{\gamma+1/2}}\,\mathrm{d}\tau &\lesssim \int_{1/2}^{1} e^{-(1-\tau)\frac{N}{2}\left[\frac{1}{a^2}+\frac{\delta}{a}+\eta^2\right]}\,\mathrm{d}\tau \\ &\lesssim \begin{cases} a^2N^{-1}\wedge 1, & N^{\rho}\le a\le \delta^{-1}\wedge\eta^{-1}, \\ a(N\delta)^{-1}\wedge 1, & N^{\rho}\vee\delta^{-1}\le a\le \delta\eta^{-2}, \\ (N\eta^2)^{-1}\wedge 1, & a\ge \delta\eta^{-2}\vee\eta^{-1}. \end{cases} \end{aligned} \tag{86}$$

For the integral over $\tau \in [N^{\rho/2}a^{-1}, 1/2]$, instead, we bound the r.h.s. of (85) as

$$\int_{N^{\rho/2}a^{-1}}^{1/2} \frac{e^{-\frac{N}{2}\left[\frac{\delta}{a\tau}+\frac{\eta^2}{\tau}\right]}}{\tau^{\gamma+1/2}}\,\mathrm{d}\tau \lesssim e^{-\frac{1}{2}N\eta^2}\times \begin{cases} e^{-\frac{1}{2}N\delta a^{-1}}(a(N\delta)^{-1})^{\gamma-1/2}, & N^{\rho}\le a\le\delta\eta^{-2}, \\ (N\eta^2)^{1/2-\gamma}, & a\ge\delta\eta^{-2}. \end{cases} \tag{87}$$

Then, combining (86)–(87) we conclude the bound in (83) for $\delta > N^{-1/2}$. Using similar computations for $\delta \le N^{-1/2}$, we conclude the bound in (83). □

In the following lemma we conclude the bound for the double integral (80) in the regime $(a, \tau) \in ([N^{\rho}, +\infty) \times ([0, 1] \setminus I)$ using the bound in (83) as an input.

Lemma 6.3. *Let $\rho > 0$ be sufficiently small, $I = [0, N^{\rho/2}a^{-1}]$, $0 \le \delta \le 1$, let $c(N)$ be defined in (44), $E \lesssim c(N)$, and let g be defined in (36). Then, for any integers $\alpha \ge 2$, $1 \le \gamma \le \alpha$, we have*

$$\int_{\Lambda\setminus\tilde{\Lambda}}\int_{[0,1]\setminus I}\left|\frac{e^{-Ng(a,\tau,\eta)}}{a^{\alpha-1}\tau^{\gamma+1/2}}\right|\mathrm{d}\tau\,\mathrm{d}a \lesssim C_1+e^{-\frac{1}{2}N\eta^2}(N\eta^2)^{1/2-\gamma}C_2+e^{-\frac{1}{2}N\eta^2}(N\delta\vee\sqrt{N})^{1/2-\gamma}C_3, \tag{88}$$

where

$$C_1 := \begin{cases} 1+|\log[N(\delta\vee N^{-1/2})\xi_*^{-1}]|, & \alpha=2,\ [\delta\eta^{-1}\vee 1]\eta^{-1}>\xi_*, \\ ((N\eta^2)^{-1}\wedge 1)\big(1+|\log[N(\delta\vee N^{-1/2})\xi_*^{-1}]|\big), & \alpha=2,\ [\delta\eta^{-1}\vee 1]\eta^{-1}\le\xi_*, \\ [N(\delta\vee N^{-1/2})]^{2-\alpha}, & \alpha\ge 3,\ [\delta\eta^{-1}\vee 1]\eta^{-1}>\xi_* \\ ((N\eta^2)^{-1}\wedge 1)[N(\delta\vee N^{-1/2})]^{2-\alpha}, & \alpha\ge 3,\ [\delta\eta^{-1}\vee 1]\eta^{-1}\le\xi_* \end{cases}$$

$$C_2 := \begin{cases} \big(1+|\log[N(\delta\vee N^{-1/2})\xi_*^{-1}]|\big)e^{-\frac{NE}{2\eta^2}(\delta\vee N^{-1/2})}, & \alpha=2, \\ [(\delta^{-1}\wedge\sqrt{N})\eta^2]^{\alpha-2}e^{-\frac{NE}{2\eta^2}(\delta\vee N^{-1/2})}, & \alpha\ge 3,\ (\delta^{-1}\wedge\sqrt{N})\eta^2\le NE, \\ [(\delta^{-1}\wedge\sqrt{N})\eta^2]^{\alpha-2}, & \alpha\ge 3,\ (\delta^{-1}\wedge\sqrt{N})\eta^2\ge NE, \end{cases}$$

$$C_3 := \begin{cases} (NE)^{\alpha-\gamma-3/2}, & \gamma=\alpha,\alpha-1,\ (\delta^{-1}\wedge\sqrt{N})\eta^2\le NE, \\ [(\delta^{-1}\wedge N^{1/2})\eta^2]^{\alpha-\gamma-3/2}, & \gamma=\alpha,\alpha-1,\ (\delta^{-1}\wedge\sqrt{N})\eta^2\ge NE, \\ (N\delta\vee\sqrt{N})^{3/2+\gamma-\alpha}, & \gamma\le\alpha-2. \end{cases}$$

Proof. First, we add and subtract $Ng(a,1,\eta)=Nf(a)$ to the phase function in the exponent and conclude, by Lemma 6.2, that

$$\int_{N^\rho}^{+\infty}\int_{[0,1]\setminus I}\left|\frac{e^{-Ng(a,\tau,\eta)}}{a^{\alpha-1}\tau^{\gamma+1/2}}\right|\,\mathrm{d}\tau\,\mathrm{d}a\lesssim\int_{N^\rho}^{+\infty}\frac{|e^{-Ng(a,1,\eta)}|\cdot F(a)}{a^{\alpha-1}}\,\mathrm{d}a, \tag{89}$$

with $F(a)$ defined in (83). In the rest of the proof we often use that $NE\xi_*\lesssim 1$ by the definition of ξ_* in (79), which implies $e^{NE\xi_*}\lesssim 1$. We split the computation of the integral in the r.h.s. of (89) as the sum of the integrals over $[N^\rho,\xi_*]$ and $[\xi_*,+\infty)$. From now on we consider only the case $\delta>N^{-1/2}$, since the case $\delta\le N^{-1/2}$ is completely analogous. In the regime $\delta>N^{-1/2}$ we have

$$E\lesssim c(N)=\delta^{-1}N^{-2}\lesssim\delta^3,$$

therefore

$$\xi_*\sim\sqrt{\delta E^{-1}}$$

from (79).

Then, using the expansion for large a-argument of $g(a,1,\eta)=f(a)$ in (75), we start estimating the integral over $[N^\rho,+\infty)$ as follows:

$$\begin{aligned}
&e^{NE\xi_*}\int_{N^\rho}^{\xi_*}\frac{e^{-N\left[\frac{\delta}{a}+\frac{1}{2a^2}\right]}F(a)}{a^{\alpha-1}}\,\mathrm{d}a\\
&\qquad\lesssim\chi(\delta\eta^{-2}\le\xi_*)e^{-\frac{1}{2}N\eta^2}(N\eta^2)^{1/2-\gamma}\times\begin{cases}1+|\log(N\delta\xi_*^{-1})|, & \alpha=2,\\ (\delta^{-1}\eta^2)^{\alpha-2}, & \alpha\ge 3,\end{cases}\\
&\qquad\qquad+e^{-\frac{1}{2}N\eta^2}\times\begin{cases}N^{1/2-\gamma}\delta^{2-\alpha}\eta^{2\alpha-2\gamma-3}, & \gamma=\alpha,\alpha-1,\delta\eta^{-2}\le\xi_*,\\ (N\delta)^{1/2-\gamma}\xi_*^{3/2+\gamma-\alpha}, & \gamma=\alpha,\alpha-1,\delta\eta^{-2}\ge\xi_*,\\ (N\delta)^{2-\alpha}, & \gamma\le\alpha-2,\end{cases}\\
&\qquad\qquad+\begin{cases}1+|\log(N\delta\xi_*^{-1})|, & \alpha=2,\ [\delta\eta^{-1}\vee 1]\eta^{-1}>\xi_*,\\ ((N\eta^2)^{-1}\wedge 1)(1+|\log(N\delta\xi_*^{-1})|), & \alpha=2,\ [\delta\eta^{-1}\vee 1]\eta^{-1}\le\xi_*,\\ (N\delta)^{2-\alpha}, & \alpha\ge 3\,[\delta\eta^{-1}\vee 1]\eta^{-1}>\xi_*,\\ ((N\eta^2)^{-1}\wedge 1)(N\delta)^{2-\alpha} & \alpha\ge 3\,[\delta\eta^{-1}\vee 1]\eta^{-1}\le\xi_*.\end{cases}
\end{aligned}\tag{90}$$

To conclude the proof we are left with the estimate of the a-integral on $[\xi_*,+\infty)$. In this regime we bound the r.h.s. of (89) as follows:

$$\begin{aligned}
&e^{NE\xi_*}\int_{\xi_*}^{+\infty}\frac{e^{-NEa}F(a)}{a^{\alpha-1}}\,\mathrm{d}a\\
&\quad\lesssim e^{-\frac{1}{2}N\eta^2}(N\eta^2)^{1/2-\gamma}\times\begin{cases}(1+|\log(NE\xi_*)|)\,e^{-NE\delta\eta^{-2}/2}, & \alpha=2,\\ e^{-NE\delta\eta^{-2}}(\delta^{-1}\eta^2)^{\alpha-2}, & \alpha\ge 3,\ \delta^{-1}\eta^2\le NE\wedge\xi_*^{-1},\\ (\delta^{-1}\eta^2)^{\alpha-2}, & \alpha\ge 3,\ NE\le\delta^{-1}\eta^2\le\xi_*^{-1},\\ \xi_*^{2-\alpha}, & \alpha\ge 3,\ \delta^{-1}\eta^2\ge\xi_*^{-1},\end{cases}
\end{aligned}$$

$$
\begin{aligned}
&+\chi(\xi_*\le\delta\eta^{-2})e^{-\frac{1}{2}N\eta^2}(N\delta)^{1/2-\gamma}\times\begin{cases}(NE)^{\alpha-\gamma-3/2}, & \gamma=\alpha,\alpha-1,\ \delta^{-1}\eta^2\le NE,\\ (\delta^{-1}\eta^2)^{\alpha-\gamma-3/2}, & \gamma=\alpha,\alpha-1,\ \delta^{-1}\eta^2\ge NE,\\ \xi_*^{3/2+\gamma-\alpha}, & \gamma\le\alpha-2,\end{cases}\\
&+\begin{cases}1+|\log(N\delta\xi_*^{-1})|, & \alpha=2,\ [\delta\eta^{-1}\vee 1]\eta^{-1}>\xi_*,\\ ((N\eta^2)^{-1}\wedge 1)(1+|\log(N\delta\xi_*^{-1})|), & \alpha=2,\ [\delta\eta^{-1}\vee 1]\eta^{-1}\le\xi_*,\\ (\xi_*)^{2-\alpha}, & \alpha\ge 3\,[\delta\eta^{-1}\vee 1]\eta^{-1}>\xi_*,\\ ((N\eta^2)^{-1}\wedge 1)(\xi_*)^{2-\alpha}, & \alpha\ge 3\,[\delta\eta^{-1}\vee 1]\eta^{-1}>\xi_*.\end{cases}
\end{aligned}
\tag{91}
$$

Finally, combining (90) and (91) we conclude the bound in (88). □

In order to conclude the estimate of the leading order term of (80), in the following lemma, using the bounds in Lemma 5.2 for the ξ-integral and the ones in Lemma 6.3 for the (a,τ)-integral, we prove that the contribution to (80) in the regime when either $a\in[0,N^\rho]$ or $\xi\in\tilde\Gamma$ and in the regime $\tau\in I$ is exponentially small.

Lemma 6.4. *Let $c(N)$ be defined in (44), $0\le\delta\le 1$, $I=I_a=[0,N^{\rho/2}a^{-1}]$, and let f, g and G_N be defined in (35)–(37), then, for any $E\lesssim c(N)$, we have that*

$$
\left|\left(\int_\Gamma \mathrm{d}\xi\int_\Lambda \mathrm{d}a\int_0^1 \mathrm{d}\tau-\int_{\Gamma\setminus\tilde\Gamma}\mathrm{d}\xi\int_{\Lambda\setminus\tilde\Lambda}\mathrm{d}a\int_{[0,1]\setminus I}\mathrm{d}\tau\right)\left[e^{N[f(\xi)-g(a,\tau,\eta)]}\frac{a\xi^2}{\tau^{1/2}}G_N(a,\tau,\xi,z)\right]\right|
\\ \lesssim\frac{N^{5/2+5\rho}(N^{1/2}+N\delta)}{E^{1/2}}e^{-\frac{1}{2}N^{1-2\rho}}. \tag{92}
$$

Proof. We split the proof into three parts, we first prove that the contribution to (80) in the regime $a\in\tilde\Lambda=[0,N^\rho]$ is exponentially small uniformly in $\tau\in[0,1]$ and $\xi\in\Gamma$, then we prove that for $a\ge N^\rho$ the contribution to (80) in the regime $\tau\in I$ is exponentially small uniformly in $\xi\in\Gamma$, and finally we conclude that also the contribution for $\xi\in\tilde\Gamma$ is negligible.

Note that for any $a\in[0,+\infty)$, $\tau\in[0,1]$ we have that the map $\tau\mapsto g(a,\tau,0)$ is strictly decreasing by (iii) of Lemma 6.1, hence, using that $g(a,\tau,\eta)\ge g(a,\tau,0)$ and (ii)–(iv) of Lemma 6.1, it follows that

$$
\sup_{\xi\in\tilde\Gamma}|e^{Nf(\xi)}|+\sup_{a\in\tilde\Lambda}|e^{-Ng(a,\tau,\eta)}|\le\sup_{\xi\in\tilde\Gamma}|e^{Nf(\xi)}|+\sup_{a\in\tilde\Lambda}|e^{-Ng(a,1,0)}|\lesssim e^{-Nf(N^\rho)}, \tag{93}
$$

with

$$
f(N^\rho)=\frac{\delta}{N^\rho}+\frac{1}{2N^{2\rho}}+\mathcal{O}(N^{-3\rho}+\delta N^{-2\rho}). \tag{94}
$$

In order to estimate the regime $a\in\tilde\Lambda$, we split the computation into two cases: $(a,\xi)\in\tilde\Lambda\times\tilde\Gamma$ and $(a,\xi)\in\tilde\Lambda\times(\Gamma\setminus\tilde\Gamma)$. Then, by (93)–(94) it follows that the integral in the regime $(a,\tau,\xi)\in\tilde\Lambda\times[0,1]\times\tilde\Gamma$ is bounded by $N^{2\rho}e^{-N^{1-2\rho}/2}$. Note that in the regime $(a,\tau,\xi)\in\tilde\Lambda\times[0,1]\times(\Gamma\setminus\tilde\Gamma)$ we have $|\xi|\ge N^\rho$. Hence, by the explicit form of $G_{1,N}$, $G_{2,N}$ in (37), using the bound in (93) for $e^{-Ng(a,\tau,\eta)}$, that $|\xi|\ge N^\rho$ and so Lemma 5.2 to bound the regime $\Gamma\setminus\tilde\Gamma$, we conclude that the integral over $(a,\tau,\xi)\in\tilde\Lambda\times[0,1]\times(\Gamma\setminus\tilde\Gamma)$ is bounded by $N^{3+\rho}(N^{1/2}+(N\delta))e^{-N^{1-2\rho}/2}$.

Next, we consider the integral over $(a,\tau,\xi)\in(\Lambda\setminus\tilde{\Lambda})\times[0,N^{\rho/2}a^{-1}]\times\Gamma$. Note that in this regime $a\geq N^{\rho}$. Since $g(a,\tau,\eta)\geq g(a,\tau,0)$ and $\tau\mapsto g(a,\tau,0)$ is strictly decreasing by (iii) of Lemma 6.1, we have that

$$e^{-Ng(a,\tau,\eta)}\leq e^{-Ng(a,N^{\rho/2}a^{-1},0)}, \tag{95}$$

where

$$g(a,N^{\rho/2}a^{-1},\eta)=Ea+\frac{\delta}{N^{\rho/2}}+\frac{1}{N^{\rho}}+\mathcal{O}(N^{-3\rho/2}+\delta N^{-\rho}). \tag{96}$$

Additionally, using the explicit expression of G_N in (37), the bound (93) on $e^{Nf(\xi)}$ for the regime $\xi\in\tilde{\Gamma}$, and Lemma 5.2 for $\xi\in\Gamma\setminus\tilde{\Gamma}$, we get

$$\left|\int_{\Gamma}\mathrm{d}\xi G_N(a,\tau,\xi,z)\xi^2e^{Nf(\xi)}\right|\lesssim C(N,\eta,a,\tau), \tag{97}$$

where

$$C(a,\tau)=C(N,\eta,a,\tau):=(N^{1/2}+N\delta)\left(N^2\eta^2+N^2\tau+\frac{N^2}{a}+\frac{N^2}{a^2\tau}\right).$$

Thus, using (97) and the explicit form of $g(a,\tau,\eta)$ in (36), for $\tau\in[0,N^{\rho/2}a^{-1}]$ we have

$$\begin{aligned}\left|\frac{a}{\tau^{1/2}}e^{-Ng(a,\tau,\eta)}\int_{\Gamma}\mathrm{d}\xi G_N(a,\tau,\xi,z)\xi^2e^{Nf(\xi)}\right|&\lesssim C(a,\tau)e^{-(N-2)g(a,\tau,\eta)}\frac{a^3\tau}{\tau^{1/2}(1+2a+a^2\tau)}\\&\lesssim C(a,\tau)a^2\tau^{1/2}e^{-(N-2)g(a,\tau,0)}e^{-\frac{(N-2)\eta^2a^2}{1+2a+a^2\tau}}\\&\lesssim C(a,\tau)a^2\tau^{1/2}e^{-\frac{1}{2}N^{1-2\rho}}e^{-NEa-\frac{(N-2)\eta^2a^2}{1+2a+a^2\tau}}.\end{aligned}\tag{98}$$

Hence, integrating (98) with respect to (a,τ), and using that in the regime $\tau\in[0,N^{\rho/2}a^{-1}]$,

$$N^{-\rho/2}a\lesssim\frac{a^2}{1+2a+a^2\tau}\lesssim a,$$

we conclude that the integral over $(a,\tau,\xi)\in(\Lambda\setminus\tilde{\Lambda})\times[0,N^{\rho/2}a^{-1}]\times\Gamma$ is bounded by

$$N^{5/2+5\rho}(N^{1/2}+N\delta)E^{-1/2}e^{-N^{1-2\rho}}.$$

Finally, in order to conclude the bound in (92), we are left with the estimate of the integral over $(a,\tau,\xi)\in(\Lambda\setminus\tilde{\Lambda})\times[N^{\rho/2}a^{-1},1]\times\tilde{\Gamma}$. In this regime, using the bound in (93) on $e^{Nf(\xi)}$ for $\xi\in\tilde{\Gamma}$, and Lemma 6.3 to estimate the integral over $(a,\tau)\in(\Lambda\setminus\tilde{\Lambda})\times[N^{\rho/2}a^{-1},1]$, we get the bound $N^{5/2+\rho}E^{-1/2}e^{-N^{1-2\rho}/2}$. This concludes the proof of (92). □

Proof of Theorem 2.3 in the case $\delta\geq 0$. Using Lemma 6.4 we remove the regime $a\leq N^{\rho},\ |\xi|\leq N^{\rho}$ or $\tau\in[0,N^{\rho/2}a^{-1}]$ in (80). Then, using the expansion for $G_{1,N}$ and $G_{2,N}$ in (76)–(77) in the remaining regime of (80), combining Lemmas 5.2 and 6.3 we conclude Theorem 2.3. □

6B. *Case* $-CN^{-1/2} \le \delta < 0$. Now we summarize the necessary changes for the case $\delta < 0$. As in the case $0 \le \delta \le 1$, throughout this section we assume that $E' < 0$ in (35)–(36), i.e., $E' = -E$ with $0 \le E \lesssim N^{-3/2}$.

Let x_* be the real stationary point of f, i.e., x_* at leading order is given by

$$x_* \approx \begin{cases} 3|\delta|^{-1}/2 & \text{if} \quad E \ll |\delta|^3, \\ \mu(c)E^{-1/3} & \text{if} \quad E = c|\delta|^3, \\ E^{-1/3} & \text{if} \quad E \gg |\delta|^3, \end{cases}$$

for some function $\mu(c) > 0$ and any fixed constant $c > 0$ independent of N, E, and δ. As in the complex case, we can treat the regime $0 < -\delta \lesssim N^{-1/2}$ similarly to the regime $0 \le \delta \lesssim N^{-1/2}$, since for $|\delta| \lesssim N^{-1/2}$ the only δ-dependent terms, i.e., the term δa^{-1} in the expansion of $g(a,1,\eta) = f(a)$ in (75) and the term $(1-\tau)\delta(a\tau)^{-1}$ in the expansion of $g(a,\tau,\eta) - g(a,1,\eta)$, do not play any role in the estimates of the (a,τ) integral in the regime $a \ge N^\rho$, $\tau \in [N^{\rho/2}a^{-1}, 1]$. Note that this is also the case for $0 \le \delta \le N^{-1/2}$, when the estimates (83) and (88) were derived. For this reason, unlike the case $0 \le \delta \le 1$, in the present case, $\delta < 0$, $|\delta| \lesssim N^{-1/2}$ and $E \lesssim c(N)$, we do not deform the ξ-contour through the leading order of the saddle x_*, but we always deform it as $\Gamma = \Gamma_{\xi_*} := \Gamma_{1,\xi_*} \cup \Gamma_{2,\xi_*}$, where $\xi_* := E^{-1/3}$, with $\Gamma_{1,\xi_*}, \Gamma_{2,\xi_*}$ defined in (48a) replacing z_* by ξ_*. Note that we could have done the same choice in the case $0 \le \delta \lesssim N^{-1/2}$, but not for the regime $N^{-1/2} \ll \delta \le 1$, hence, in order to treat the regime $0 \le \delta \le 1$ in the same way for any δ, in Section 6A we deformed the contour Γ through (79). For any $E \lesssim c(N)$ and $0 < -\delta \lesssim N^{-1/2}$ we have $\xi_* \gg 1$, hence we prove that the main contribution to (80) comes from the regime when $a, |\xi| \ge N^\rho$. Moreover, similarly to the case $0 \le \delta \le 1$, the contribution to (80) in the regime $(a,\tau,\xi) \in (\Lambda \setminus \tilde{\Lambda}) \times [0, N^{\rho/2}a^{-1}] \times (\Gamma \setminus \tilde{\Gamma})$ will be exponentially small. Hence, in order to estimate the leading order of (80), we expand f, $g(\cdot, 1, \eta)$ and G_N for large a and $|\xi|$ arguments as in (75)–(77).

The phase functions f and g, defined in (35) and (36), respectively, satisfy the properties (i) and (ii) of Lemma 6.1, but not the ones in (iii) and (iv). Instead, it is easy to see that the following lemma holds true.

Lemma 6.5. *Let f and g be the phase functions defined in* (35) *and* (36), *respectively. Then, the following properties hold true*:

(iii') *For any $a \in [0, +\infty)$, we have that $g(a,\tau,\eta) \ge g(a,\tau,0)$ and that*

$$e^{-Ng(a,\tau,0)} \lesssim e^{-Ng(a,\tau_0,0)},$$

for any fixed $\tau_0 \in [0,1]$ and any $\tau \in [0,\tau_0]$.

(iv') *The function $a \mapsto g(a,1,0)$ is strictly decreasing on* $[0, |\delta|^{-1}/2]$.

Since $|\delta| \lesssim N^{-1/2}$, by (iv') of Lemma 6.5 it clearly follows that the function $a \mapsto g(a,1,\eta)$ is strictly decreasing on $[0, N^\rho]$. Note that for $|\delta| \lesssim N^{-1/2}$ we have

$$e^{-N(1-\tau)\left[\frac{1}{a^2\tau^2} + \frac{\delta}{a\tau}\right]} \lesssim e^{-\frac{(1-\tau)N}{2a^2\tau^2}}, \qquad e^{-N\left[\frac{\delta}{a} + \frac{1}{2a^2}\right]} \lesssim e^{-\frac{N}{4a^2}}, \tag{99}$$

for any $a \in [N^\rho, +\infty)$ and $\tau \in [N^{\rho/2}a^{-1}, 1]$. Using (99), inspecting the proofs of (83) and (88) in Lemma 6.2 and Lemma 6.3, respectively, and noticing that in the regime $0 \le \delta \le N^{-1/2}$ the sign of δ did not play any role we conclude that the bounds (83), (88) hold true for the case $\delta < 0$, $|\delta| \lesssim N^{-1/2}$ as well. Then, much as in the case $0 \le \delta \le 1$, by (i)–(ii) of Lemma 6.1 and (iii')–(iv') of Lemma 6.5, using the bound in (88) to estimate the (a, τ)-integral in the regime $(a, \tau) \in [N^{\rho}, +\infty) \times [N^{\rho/2}a^{-1}, 1]$ and the ones in Lemma 5.2 to estimate the ξ-integral in the regime $|\xi| \ge N^{\rho}$ we conclude that the contribution to (80) in the regime when either $a \in [0, N^{\rho}]$ or $|\xi| \le N^{\rho}$ and in the regime $[0, N^{\rho/2}a^{-1}]$ is exponentially small, i.e., Lemma 6.4 holds true once δ is replaced by $|\delta|$ everywhere.

Proof of Theorem 2.3 in the case $-CN^{-1/2} \le \delta < 0$. By combining (88), Lemmas 5.2 and 6.4, using the expansion of G_N in (76)–(77), we conclude the proof of Theorem 2.3 also in the case $-CN^{-1/2} \le \delta < 0$. □

Appendix A. Superbosonization formula for meromorphic functions

The superbosonization formulas [Littelmann et al. 2008, Eq. (1.10) and (1.13)] (see also [Alldridge and Shaikh 2014, Corollary 2.6] for more precise conditions) are stated under the condition that

$$F(\Phi^*\Phi) = F\begin{pmatrix} \langle s, s\rangle & \langle s, \chi\rangle \\ \langle \chi, s\rangle & \langle \chi, \chi\rangle \end{pmatrix},$$

viewed as a function of four independent variables, is holomorphic and decays faster than any inverse power at real $+\infty$ in the $\langle s, s\rangle$ variable (for definiteness, we discuss the complex case; the argument for the real case is analogous). Our function F defined in (26) has a pole at $\langle s, s\rangle = \mathrm{i}N$ and $\langle \chi, \chi\rangle = \mathrm{i}N$ using the definitions (20)–(21) after expanding the inverse of the matrix $1 + \frac{\mathrm{i}}{N}\Phi^*\Phi$ in the Grassmannian variables but this pole is far away from the integration domain on both sides of (22). We now outline a standard approximation procedure to verify the superbosonization formula for such meromorphic functions; for simplicity we consider only our concrete function from (26).

In the first step notice that the integration at infinity on the noncompact domain for the boson-boson variable is absolutely convergent on both sides as guaranteed by the $\exp(\mathrm{i}w \operatorname{STr} \Phi^*\Phi)$ regularization, since $\Im w > 0$.

Second, in the l.h.s. of (22) using Taylor expansions, we expand F into a finite polynomial in the Grassmannian variables with meromorphic coefficient functions in the variable $\langle s, s\rangle$. Algebraically, we perform exactly the same expansion in the r.h.s. of (22). For the fermionic variables σ, τ these expansions naturally terminate after finitely many terms. From (28) it is clear that only the geometric expansion

$$(1 + y)^{-1} = 1 - y + y^2 - \cdots$$

may result in an infinite power series instead of a finite polynomial. However, owing to the contour integral in y and that the integrand has a pole of at most finite order ($\approx N$) at zero, we may replace this power series with its finite truncation without changing the value of the r.h.s. of (22). We choose the order of truncation sufficiently large that the remaining formula contains all nonzero terms on both sides. We denote this new truncated function by $\tilde{F}$.

Now we are in the situation where on both sides of (22), with $\tilde{F}$ replacing F, we have the same finite polynomial in the variables $\langle s, \chi\rangle, \langle \chi, s\rangle$ and $\langle \chi, \chi\rangle$ in the l.h.s. as in the variables σ, τ, y in the r.h.s., with coefficients that are meromorphic in $\langle s, s\rangle$, resp. in x. All coefficient functions $h_k(x)$ are analytic in a neighborhood of the positive real axis (their possible pole is at -1) and they have an exponential decay $\approx \exp(-(\Im w)\langle s, s\rangle)$ in the l.h.s., resp. $\exp(-(\Im w)x)$ in the r.h.s., at infinity from the regularization observed in the first step.

Finally, in the third step, dropping the k index temporarily, we write each coefficient function as $h(x) = g(x)e^{-\alpha x}$ with $\alpha = \frac{1}{2}\Im w$. For any given $\epsilon > 0$ we approximate $g(x)$ via classical (rescaled) Laguerre polynomials $p_n(x)$ of degree n with weight function $e^{-\alpha x}$ such that $\int_0^\infty |g(x) - p_n(x)|^2 e^{-\alpha x}\,\mathrm{d}x \le (\epsilon\alpha)^2$, where n depends on ϵ and $\Im w$. By completeness of the Laguerre polynomials in $L^2(\mathbb{R}_+, e^{-\alpha x}\,\mathrm{d}x)$ and by $\int |g(x)|^2 e^{-\alpha x}\,\mathrm{d}x = \int |h(x)|^2 e^{\alpha x}\,\mathrm{d}x < \infty$ such an approximating polynomial exists. Therefore, with a Schwarz inequality, we have

$$\int_0^\infty |h(x) - p_n(x)e^{-\alpha x}|\,\mathrm{d}x = \int_0^\infty |g(x) - p_n(x)|e^{-\alpha x}\,\mathrm{d}x \le \epsilon.$$

Since there are only finitely many coefficient functions $h(x) = h_k(x)$ in $\tilde{F}$, we can replace each of them with an entire function (namely with a polynomial times $e^{-\alpha x}$) with at most an ϵ error in the r.h.s. of (22). The same estimates hold on the r.h.s. But for these replacements the superbosonization formula [Littelmann et al. 2008, Eq. (1.10)] is applicable since the new functions are entire. The error is at most ϵ on both sides, but this argument is valid for arbitrary $\epsilon > 0$. This proves the superbosonization formula for the function (26).

Appendix B. Explicit formulas for the real symmetric integral representation

Here we collect the explicit formulas for the polynomials of a, ξ, τ in the definition of G_N in (34).

$$\begin{aligned}
p_{2,0,0} &:= a^4\tau^2 + 2a^3\xi\tau + 4a^3\tau - a^2\xi^2\tau + 4a^2\xi^2 + 8a^2\xi + 2a^2\tau \\
&\qquad\qquad + 4a^2 + 2a\xi^3 + 8a\xi^2 + 10a\xi + 4a + \xi^4 + 4\xi^3 + 6\xi^2 + 4\xi + 1, \\
p_{1,0,0} &:= -a^4\xi\tau^2 + a^4\tau^2 - 2a^3\xi^2\tau - 2a^3\xi\tau + 4a^3\tau - a^2\xi^3\tau - 3a^2\xi^2\tau \\
&\qquad\qquad - 2a^2\xi\tau + 4a^2\xi + 2a^2\tau + 4a^2 + 2a\xi^2 + 6a\xi + 4a + \xi^3 + 3\xi^2 + 3\xi + 1, \\
p_{2,2,0} &:= 4(a+1)(a^2\tau + a\xi\tau + 2a\tau + \xi^2 + 2\xi + 1), \\
p_{1,2,0} &:= 4(a+1)(a^2\tau + a\xi\tau + 2a\tau + \xi + 1), \\
p_{2,0,1} &:= 2(a^3\tau^2 + 2a^2\xi\tau + 4a^2\tau + 2a\xi^2 + 2a\xi\tau + 4a\xi + 3a\tau + 2a + \xi^3 + 4\xi^2 + 5\xi + 2), \\
p_{1,0,1} &:= 2(a^3\tau^2 + 2a^2\xi\tau + 4a^2\tau + a\xi^2\tau + 3a\xi\tau + 2a\xi + 3a\tau + 2a + \xi^2 + 3\xi + 2), \\
p_{2,2,1} &:= 4(a+1)(a + \xi + 2), \\
p_{2,0,2} &:= a^2\tau + 2a\xi + 4a + \xi^2 + 4\xi + 4.
\end{aligned}$$

Appendix C. Comparison with the contour-integral derivation

In [Ben Arous and Péché 2005] the correlation kernel of $(X - z)(X - z)^*$ for complex Ginibre matrices X has been derived using contour-integral methods. Earlier, the joint eigenvalue density for the general *Laguerre ensemble* had been obtained in the physics literature [Guhr and Wettig 1996; Jackson et al.

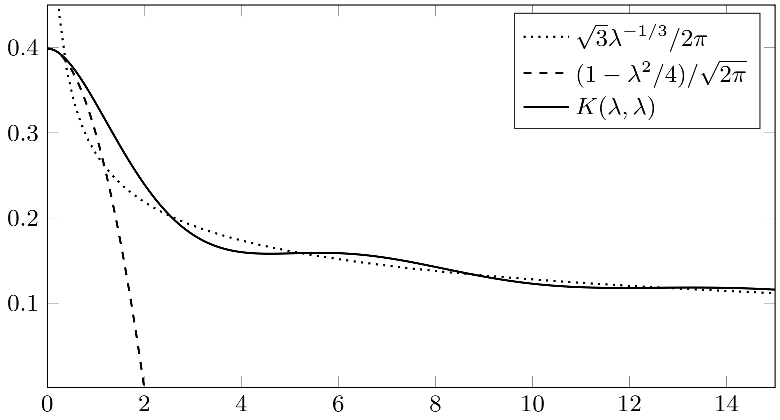


Figure 7. Plot of the 1-point function $K(\lambda,\lambda)=\pi^{-1}\Im q_0(\lambda)$ in the complex case with $|z|=1$. The dotted and dashed lines show the large λ and small λ asymptotes, respectively.

1996] via supersymmetric methods; see also [Verbaarschot and Zahed 1993] with orthogonal polynomials. Adapting [Ben Arous and Péché 2005] to our scaling, and choosing $y_i=\pm 1$, it follows from [Ben Arous and Péché 2005, Theorem 7.1] that for $|z|=1$ the rescaled kernel $K_N(N^{-3/2}\lambda, N^{-3/2}\mu)$ is given by

$$\begin{aligned}&\frac{N^3}{\mathrm{i}\pi}\int_\Gamma \mathrm{d}x\int_\gamma \mathrm{d}y K_B(2N^{1/4}x\sqrt{\lambda},2N^{1/4}y\sqrt{\mu})e^{N(h(y)-h(x))}(1-\frac{1}{1-x^2}\frac{1}{1-y^2})xy,\\ &K_B(x,y):=\frac{xI_1(x)I_0(y)-yI_0(x)I_1(y)}{x^2-y^2},\qquad h(x)=x^2+\log(1-x^2)=-\frac{x^4}{2}+\mathcal{O}(x^5),\end{aligned}\tag{100}$$

where I_0, I_1 are the modified Bessel function of 0-th and 1-st kind. The contour Γ is any contour encircling $[-1,1]$ in a counterclockwise direction (in contradiction to the contours depicted in [Ben Arous and Péché 2005, Figure 8.1]) and the contour γ is composed of two straight half-lines $[0,\mathrm{i}\infty)$ and $[0,-\mathrm{i}\infty)$. The main contribution in (100) comes from the $|x|\lesssim N^{-1/4}$ and $|y|\lesssim N^{-1/4}$ regime which motivates the change of variables $x\mapsto N^{-1/4}x$, $y\mapsto N^{-1/4}y$. Together with the expansion of $1-(1-x^2)^{-1}(1-y^2)^{-1}=-x^2-y^2+\mathcal{O}(x^4+y^4)$ it follows that

$$K_N(N^{-3/2}\lambda,N^{-3/2}\mu)\approx N^{3/2}K(\lambda,\mu),$$

where

$$K(\lambda,\mu):=\frac{\mathrm{i}}{\pi}\int_{\Gamma'}\mathrm{d}x\int_\gamma \mathrm{d}yK_B(2x\sqrt{\lambda},2y\sqrt{\mu})e^{x^4/2-y^4/2}xy(x^2+y^2)$$

and Γ' consists of four straight half-lines $(e^{\mathrm{i}\pi/4}\infty,0]$, $[0,e^{3\mathrm{i}\pi/4}\infty)$, $(e^{5\mathrm{i}\pi/4}\infty,0]$, $[0,e^{7\mathrm{i}\pi/4}\infty)$.

We now compare the limiting 1-point function $K(\lambda,\lambda)$ with the asymptotic expansion we derived in Theorem 2.1, which in the case $|z|=1$, i.e., $\delta=0$, simplifies to

$$q_0(\lambda)=\frac{\lambda^{1/3}}{2\pi\mathrm{i}}\int\mathrm{d}x\oint\mathrm{d}y e^{\lambda^{2/3}(-y+1/(2y^2)+x-1/(2x^2))}\Big(\frac{1}{x^3}+\frac{1}{x^2y}+\frac{1}{xy^2}\Big).$$

The resulting 1-point function, given by $\pi^{-1}\Im q_0(\lambda)$, coincides precisely with $K(\lambda,\lambda)$ and is plotted in Figure 7.

Acknowledgements

The authors are grateful to Nicholas Cook and Patrick Lopatto for pointing out missing references, and to Ievgenii Afanasiev for useful remarks. We would also like to thank the anonymous referees for drawing our attention to additional references in the physics literature.

References

[Afanasiev 2019] I. Afanasiev, "On the correlation functions of the characteristic polynomials of non-Hermitian random matrices with independent entries", *J. Stat. Phys.* **176**:6 (2019), 1561–1582. MR

[Ajanki et al. 2019] O. H. Ajanki, L. Erdős, and T. Krüger, "Stability of the matrix Dyson equation and random matrices with correlations", *Probab. Theory Related Fields* **173**:1-2 (2019), 293–373. MR

[Akemann et al. 2005] G. Akemann, J. C. Osborn, K. Splittorff, and J. J. M. Verbaarschot, "Unquenched QCD Dirac operator spectra at nonzero baryon chemical potential", *Nuclear Phys. B* **712**:1-2 (2005), 287–324. MR

[Alldridge and Shaikh 2014] A. Alldridge and Z. Shaikh, "Superbosonization via Riesz superdistributions", *Forum Math. Sigma* **2** (2014), art. id. e9. MR

[Alt et al. 2019] J. Alt, L. Erdős, and T. Krüger, "Spectral radius of random matrices with independent entries", preprint, 2019. arXiv

[Bai 1997] Z. D. Bai, "Circular law", *Ann. Probab.* **25**:1 (1997), 494–529. MR

[Bai and Silverstein 2012] Z. Bai and J. W. Silverstein, "No eigenvalues outside the support of the limiting spectral distribution of information-plus-noise type matrices", *Random Matrices Theory Appl.* **1**:1 (2012), art.idİ150004. MR

[Bai and Yin 1986] Z. D. Bai and Y. Q. Yin, "Limiting behavior of the norm of products of random matrices and two problems of Geman–Hwang", *Probab. Theory Related Fields* **73**:4 (1986), 555–569. MR

[Bao and Erdős 2017] Z. Bao and L. Erdős, "Delocalization for a class of random block band matrices", *Probab. Theory Related Fields* **167**:3-4 (2017), 673–776. MR

[Ben Arous and Péché 2005] G. Ben Arous and S. Péché, "Universality of local eigenvalue statistics for some sample covariance matrices", *Comm. Pure Appl. Math.* **58**:10 (2005), 1316–1357. MR

[Berezin 1987] F. A. Berezin, *Introduction to superanalysis*, Mathematical Physics and Applied Mathematics **9**, Reidel, Dordrecht, 1987. MR

[Bordenave and Chafaï 2012] C. Bordenave and D. Chafaï, "Around the circular law", *Probab. Surv.* **9** (2012), 1–89. MR

[Bordenave et al. 2018] C. Bordenave, P. Caputo, D. Chafaï, and K. Tikhomirov, "On the spectral radius of a random matrix: an upper bound without fourth moment", *Ann. Probab.* **46**:4 (2018), 2268–2286. MR

[Borodin and Sinclair 2009] A. Borodin and C. D. Sinclair, "The Ginibre ensemble of real random matrices and its scaling limits", *Comm. Math. Phys.* **291**:1 (2009), 177–224. MR

[Bourgade et al. 2014] P. Bourgade, L. Erdös, and H.-T. Yau, "Edge universality of beta ensembles", *Comm. Math. Phys.* **332**:1 (2014), 261–353. MR

[Brézin and Hikami 1998] E. Brézin and S. Hikami, "Level spacing of random matrices in an external source", *Phys. Rev. E* (3) **58**:6, part A (1998), 7176–7185. MR

[Chalker and Mehlig 1998] J. T. Chalker and B. Mehlig, "Eigenvector statistics in non-Hermitian random matrix ensembles", *Phys. Rev. Lett.* **81**:16 (1998), 3367–3370.

[Che and Lopatto 2019a] Z. Che and P. Lopatto, "Universality of the least singular value for sparse random matrices", *Electron. J. Probab.* **24** (2019), art. id. 9. MR

[Che and Lopatto 2019b] Z. Che and P. Lopatto, "Universality of the least singular value for the sum of random matrices", preprint, 2019. arXiv

[Cipolloni et al. 2019a] G. Cipolloni, L. Erdős, and D. Schröder, "Central limit theorem for linear eigenvalue statistics of non-Hermitian random matrices", preprint, 2019. arXiv

[Cipolloni et al. 2019b] G. Cipolloni, L. Erdős, and D. Schröder, "Edge universality for non-Hermitian random matrices", preprint, 2019. arXiv

[Cipolloni et al. 2020] G. Cipolloni, L. Erdős, and D. Schröder, "Fluctuation around the circular law for random matrices with real entries", preprint, 2020. arXiv

[Cook 2018] N. Cook, "Lower bounds for the smallest singular value of structured random matrices", *Ann. Probab.* **46**:6 (2018), 3442–3500. MR

[Desrosiers and Forrester 2006] P. Desrosiers and P. J. Forrester, "Asymptotic correlations for Gaussian and Wishart matrices with external source", *Int. Math. Res. Not.* **2006** (2006), art. id. 27395. MR

[Disertori and Lager 2017] M. Disertori and M. Lager, "Density of states for random band matrices in two dimensions", *Ann. Henri Poincaré* **18**:7 (2017), 2367–2413. MR

[Disertori et al. 2002] M. Disertori, H. Pinson, and T. Spencer, "Density of states for random band matrices", *Comm. Math. Phys.* **232**:1 (2002), 83–124. MR

[Disertori et al. 2018] M. Disertori, M. Lohmann, and S. Sodin, "The density of states of 1D random band matrices via a supersymmetric transfer operator", preprint, 2018. arXiv

[Dozier and Silverstein 2007] R. B. Dozier and J. W. Silverstein, "On the empirical distribution of eigenvalues of large dimensional information-plus-noise-type matrices", *J. Multivariate Anal.* **98**:4 (2007), 678–694. MR

[Edelman 1988] A. Edelman, "Eigenvalues and condition numbers of random matrices", *SIAM J. Matrix Anal. Appl.* **9**:4 (1988), 543–560. MR

[Edelman 1997] A. Edelman, "The probability that a random real Gaussian matrix has k, real eigenvalues, related distributions, and the circular law", *J. Multivariate Anal.* **60**:2 (1997), 203–232. MR

[Efetov 1997] K. Efetov, *Supersymmetry in disorder and chaos*, Cambridge University Press, 1997. MR

[Erdős et al. 2018] L. Erdős, T. Krüger, and D. Renfrew, "Power law decay for systems of randomly coupled differential equations", *SIAM J. Math. Anal.* **50**:3 (2018), 3271–3290. MR

[Forrester and Nagao 2007] P. J. Forrester and T. Nagao, "Eigenvalue statistics of the real Ginibre ensemble", *Phys. Rev. Lett.* **99**:5 (2007), 050603.

[Fyodorov 2018] Y. V. Fyodorov, "On statistics of bi-orthogonal eigenvectors in real and complex Ginibre ensembles: combining partial Schur decomposition with supersymmetry", *Comm. Math. Phys.* **363**:2 (2018), 579–603. MR

[Fyodorov et al. 2018] Y. V. Fyodorov, J. Grela, and E. Strahov, "On characteristic polynomials for a generalized chiral random matrix ensemble with a source", *J. Phys. A* **51**:13 (2018), art. id. 134003.

[Geman 1986] S. Geman, "The spectral radius of large random matrices", *Ann. Probab.* **14**:4 (1986), 1318–1328. MR

[Ginibre 1965] J. Ginibre, "Statistical ensembles of complex, quaternion, and real matrices", *J. Mathematical Phys.* **6** (1965), 440–449. MR

[Girko 1984] V. L. Girko, "The circular law", *Teor. Veroyatnost. i Primenen.* **29**:4 (1984), 669–679. MR

[Guhr 2011] T. Guhr, "Supersymmetry", pp. 135–154 in *The Oxford handbook of random matrix theory*, Oxford Univ. Press, Oxford, 2011. MR

[Guhr and Wettig 1996] T. Guhr and T. Wettig, "An Itzykson–Zuber-like integral and diffusion for complex ordinary and supermatrices", *J. Math. Phys.* **37**:12 (1996), 6395–6413. MR

[Guhr and Wettig 1997] T. Guhr and T. Wettig, "Universal spectral correlations of the Dirac operator at finite temperature", *Nuclear Phys. B* **506**:3 (1997), 589–611. MR

[Jackson et al. 1996] A. D. Jackson, M. K. Şener, and J. J. M. Verbaarschot, "Finite volume partition functions and Itzykson–Zuber integrals", *Phys. Lett. B* **387**:2 (1996), 355–360. MR

[Kieburg et al. 2009] M. Kieburg, H.-J. Sommers, and T. Guhr, "A comparison of the superbosonization formula and the generalized Hubbard–Stratonovich transformation", *J. Phys. A* **42**:27 (2009), artid2̇75206. MR

[Lehmann and Sommers 1991] N. Lehmann and H.-J. Sommers, "Eigenvalue statistics of random real matrices", *Phys. Rev. Lett.* **67**:8 (1991), 941–944. MR

[Littelmann et al. 2008] P. Littelmann, H.-J. Sommers, and M. R. Zirnbauer, "Superbosonization of invariant random matrix ensembles", *Comm. Math. Phys.* **283**:2 (2008), 343–395. MR

[May 1972] R. M. May, "Will a large complex system be stable?", *Nature* **238**:5364 (1972), 413–414.

[Mehta 1967] M. L. Mehta, *Random matrices and the statistical theory of energy levels*, Academic Press, New York, 1967. MR

[Mo 2012] M. Y. Mo, "Rank 1 real Wishart spiked model", *Comm. Pure Appl. Math.* **65**:11 (2012), 1528–1638. MR

[Osborn et al. 1999] J. C. Osborn, D. Toublan, and J. J. M. Verbaarschot, "From chiral random matrix theory to chiral perturbation theory", *Nuclear Physics B* **540**:1 (1999), 317–344.

[Recher et al. 2012] C. Recher, M. Kieburg, T. Guhr, and M. R. Zirnbauer, "Supersymmetry approach to Wishart correlation matrices: Exact results", *J. Stat. Phys.* **148**:6 (2012), 981–998. MR

[Rudelson 2008] M. Rudelson, "Invertibility of random matrices: norm of the inverse", *Ann. of Math.* (2) **168**:2 (2008), 575–600. MR

[Rudelson and Vershynin 2008] M. Rudelson and R. Vershynin, "The Littlewood–Offord problem and invertibility of random matrices", *Adv. Math.* **218**:2 (2008), 600–633. MR

[Sankar et al. 2006] A. Sankar, D. A. Spielman, and S.-H. Teng, "Smoothed analysis of the condition numbers and growth factors of matrices", *SIAM J. Matrix Anal. Appl.* **28**:2 (2006), 446–476. MR

[Shcherbina 2011] T. Shcherbina, "On the correlation function of the characteristic polynomials of the Hermitian Wigner ensemble", *Comm. Math. Phys.* **308**:1 (2011), 1–21. MR

[Shcherbina 2013] T. Shcherbina, "On the correlation functions of the characteristic polynomials of the Hermitian sample covariance matrices", *Probab. Theory Related Fields* **156**:1-2 (2013), 449–482. MR

[Shcherbina 2014] T. Shcherbina, "On the second mixed moment of the characteristic polynomials of 1D band matrices", *Comm. Math. Phys.* **328**:1 (2014), 45–82. MR

[Shcherbina 2020] T. Shcherbina, "Characteristic polynomials for random band matrices near the threshold", *J. Stat. Phys.* **179**:4 (2020), 920–944. MR

[Shcherbina and Shcherbina 2016] M. Shcherbina and T. Shcherbina, "Transfer matrix approach to 1d random band matrices: density of states", *J. Stat. Phys.* **164**:6 (2016), 1233–1260. MR

[Shcherbina and Shcherbina 2017] M. Shcherbina and T. Shcherbina, "Characteristic polynomials for 1D random band matrices from the localization side", *Comm. Math. Phys.* **351**:3 (2017), 1009–1044. MR

[Shcherbina and Shcherbina 2019] M. Shcherbina and T. Shcherbina, "Universality for 1D random band matrices", preprint, 2019. arXiv

[Sompolinsky et al. 1988] H. Sompolinsky, A. Crisanti, and H.-J. Sommers, "Chaos in random neural networks", *Phys. Rev. Lett.* **61**:3 (1988), 259–262. MR

[Soshnikov 1999] A. Soshnikov, "Universality at the edge of the spectrum in Wigner random matrices", *Comm. Math. Phys.* **207**:3 (1999), 697–733. MR

[Spencer 2011] T. Spencer, "Random banded and sparse matrices", pp. 471–488 in *The Oxford handbook of random matrix theory*, Oxford Univ. Press, 2011. MR

[Stephanov 1996] M. A. Stephanov, "Random matrix model of QCD at finite density and the nature of the quenched limit", *Phys. Rev. Lett.* **76**:24 (1996), 4472–4475.

[Tao and Vu 2007] T. Tao and V. Vu, "The condition number of a randomly perturbed matrix", pp. 248–255 in *STOC'07: Proceedings of the 39th Annual ACM Symposium on Theory of Computing*, ACM, New York, 2007. MR

[Tao and Vu 2008] T. Tao and V. Vu, "Random matrices: the circular law", *Commun. Contemp. Math.* **10**:2 (2008), 261–307. MR

[Tao and Vu 2009] T. Tao and V. H. Vu, "Inverse Littlewood–Offord theorems and the condition number of random discrete matrices", *Ann. of Math.* (2) **169**:2 (2009), 595–632. MR

[Tao and Vu 2010a] T. Tao and V. Vu, "Random matrices: the distribution of the smallest singular values", *Geom. Funct. Anal.* **20**:1 (2010), 260–297. MR

[Tao and Vu 2010b] T. Tao and V. Vu, "Smooth analysis of the condition number and the least singular value", *Math. Comp.* **79**:272 (2010), 2333–2352. MR

[Tikhomirov 2020] K. Tikhomirov, "Invertibility via distance for noncentered random matrices with continuous distributions", *Random Structures Algorithms* **57**:2 (2020), 526–562. MR

[Verbaarschot and Wettig 2000] J. J. M. Verbaarschot and T. Wettig, "Random matrix theory and chiral symmetry in QCD", *Annu. Rev. Nucl. Part. Sci.* **50**:1 (2000), 343–410.

[Verbaarschot and Zahed 1993] J. J. Verbaarschot and I. Zahed, "Spectral density of the QCD Dirac operator near zero virtuality", *Phys. Rev. Lett.* **70**:25 (1993), 3852–3855.

[Verbaarschot et al. 1985] J. J. M. Verbaarschot, H. A. Weidenmüller, and M. R. Zirnbauer, "Grassmann integration in stochastic quantum physics: the case of compound-nucleus scattering", *Phys. Rep.* **129**:6 (1985), 367–438. MR

[Wilke et al. 1998] T. Wilke, T. Guhr, and T. Wettig, "Microscopic spectrum of the QCD Dirac operator with finite quark masses", *Phys. Rev. D* **157**:10 (1998), 6486–6495.

Received 22 Oct 2019. Revised 6 Mar 2020. Accepted 30 Mar 2020.

GIORGIO CIPOLLONI: giorgio.cipolloni@ist.ac.at
Institute of Science and Technology Austria, Klosterneuburg, Austria

LÁSZLÓ ERDŐS: lerdos@ist.ac.at
Institute of Science and Technology Austria, Klosterneuburg, Austria

DOMINIK SCHRÖDER: dschroeder@ethz.ch
Institute of Science and Technology Austria, Klosterneuburg, Austria

PROBABILITY and
MATHEMATICAL
PHYSICS

Vol. 1, No. 1, 2020
https://doi.org/10.2140/pmp.2020.1.147

NEW CRITICAL EXPONENT INEQUALITIES FOR PERCOLATION AND THE RANDOM CLUSTER MODEL

TOM HUTCHCROFT

We apply a variation on the methods of Duminil-Copin, Raoufi, and Tassion (*Ann. of Math.* (2) **189**:1 (2019), 75–99) to establish a new differential inequality applying to both Bernoulli percolation and the Fortuin–Kasteleyn random cluster model. This differential inequality has a similar form to that derived for Bernoulli percolation by Menshikov (*Dokl. Akad. Nauk* **288**:6 (1986), 1308–1311) but with the important difference that it describes the distribution of the *volume* of a cluster rather than of its radius. We apply this differential inequality to prove the following:

(1) The critical exponent inequalities $\gamma \leq \delta - 1$ and $\Delta \leq \gamma + 1$ hold for percolation and the random cluster model on any transitive graph. These inequalities are new even in the context of Bernoulli percolation on $\mathbb{Z}^d$, and are saturated in mean-field for Bernoulli percolation and for the random cluster model with $q \in [1, 2)$.

(2) The volume of a cluster has an exponential tail in the entire subcritical phase of the random cluster model on any transitive graph. This proof also applies to infinite-range models, where the result is new even in the Euclidean setting.

1. Introduction

Differential inequalities play a central role in the rigorous study of percolation and other random media. Indeed, one of the most important theorems in the theory of Bernoulli percolation is that the phase transition is *sharp*, meaning (in one precise formulation) that the radius of the cluster of the origin has an exponential tail throughout the entire subcritical phase. This theorem was first proven in independent works of Menshikov [1986] and Aizenman and Barsky [1987]. While these two proofs were rather different, they both relied crucially on differential inequalities: In Menshikov's case this differential inequality was

$$\frac{d}{dp} \log \boldsymbol{P}_p(R \geq n) \geq \frac{1}{p}\left[\frac{n}{\sum_{m=0}^{n} \boldsymbol{P}_p(R \geq m)} - 1\right] \qquad \text{for each } n \geq 1, \tag{1-1}$$

where R denotes the radius of the cluster of the origin, while for Aizenman and Barsky the relevant differential inequalities were

$$M \leq h\frac{\partial M}{\partial h} + M^2 + pM\frac{\partial M}{\partial p} \quad \text{and} \quad \frac{\partial M}{\partial p} \leq dM\frac{\partial M}{\partial h}, \tag{1-2}$$

where we write $|K|$ for the volume of the cluster of the origin, write $M = M_{p,h}$ for the *magnetization*

MSC2010: primary 60K35; secondary 82B27.
Keywords: percolation, random-cluster model, differential inequality, critical exponent, scaling relations, subcritical.

$M = \boldsymbol{E}_p[1-e^{-h|K|}]$, and write d for the degree of the graph. An alternative, simpler proof of sharpness for percolation, which also relies on differential inequalities, was subsequently found by Duminil-Copin and Tassion [2016]. Aside from their use to establish sharpness, the differential inequalities (1-1) and (1-2) also yield further quantitative information about percolation at and near criticality. In particular, both inequalities can be used to derive bounds on *critical exponents* associated to percolation; this is discussed further in Section 1A and reviewed in detail in [Grimmett 1999]. Similar methods have also yielded similar results for the Ising model [Aizenman et al. 1987; Duminil-Copin and Tassion 2016].

Aside from percolation and the Ising model, the class of models that were rigorously proven to undergo sharp phase transitions was, until recently, very limited. In particular, the derivations of both (1-1) and (1-2) rely heavily on the van den Berg–Kesten (BK) inequality [1985], and are therefore rather specific to Bernoulli percolation. This situation has now improved drastically following the breakthrough work of Duminil-Copin, Raoufi, and Tassion [Duminil-Copin et al. 2019b], who showed that the theory of *randomized algorithms* can often be used to prove sharpness of the phase transition in models satisfying the FKG lattice condition. They first applied this new methodology to prove that a differential inequality essentially equivalent to that of Menshikov (1-1) holds for the Fortuin–Kasteleyn random-cluster model (with $q \geq 1$), from which they deduced sharpness of the phase transition for this model and the ferromagnetic Potts model. Variations on their methods have subsequently been used to prove sharpness results for several other models, including Voronoi percolation [Duminil-Copin et al. 2019a], Poisson-Boolean percolation [Duminil-Copin et al. 2018], the Widom-Rowlinson model [Dereudre and Houdebert 2018], level sets of smooth planar Gaussian fields [Muirhead and Vanneuville 2020], and the contact process [Beekenkamp 2018].

The main new technical tool introduced by [Duminil-Copin et al. 2019b] was a generalization of the *OSSS inequality* from product measures to *monotonic measures*. This inequality, introduced by O'Donnell, Saks, Schramm, and Servedio [O'Donnell et al. 2005], can be used to derive differential inequalities for percolation in the following way: Let A be an increasing event depending on at most finitely many edges, and suppose that we have an algorithm for computing whether or not A occurs. This algorithm decides sequentially which edges to reveal the status of, with decisions depending on what it has previously seen and possibly also some external randomness, stopping when it has determined whether or not A has occurred. For each edge e, let δ_e be the *revealment* of e, defined to be the probability that the status of the edge e is ever queried by the algorithm. Then the OSSS inequality implies that

$$\frac{d}{dp}\log \boldsymbol{P}_p(A) \geq \frac{1-\boldsymbol{P}_p(A)}{p(1-p)\max_{e\in E}\delta_e}. \tag{1-3}$$

In particular, if $\boldsymbol{P}_p(A)$ is not too large and there exists a randomized algorithm determining whether or not A holds with low maximum revealment, then the logarithmic derivative of $\boldsymbol{P}_p(A)$ is large. This yields an extremely flexible methodology for deriving differential inequalities for percolation. Even greater flexibility is provided by the *two-function* version of the OSSS inequality, which implies in particular that if A and B are events, where A is increasing and we have some randomized algorithm that determines whether or not B occurs, then

$$\frac{d}{dp}\log \boldsymbol{P}_p(A) \geq \frac{\boldsymbol{P}_p(B \mid A)-\boldsymbol{P}_p(B)}{p(1-p)\max_{e\in E}\delta_e}. \tag{1-4}$$

The new differential inequality. In this article, we apply the OSSS inequality to establish a new differential inequality for percolation and the random cluster model. Once we establish this inequality, we use it to prove several other new results for these models which are detailed in the following subsections. Our new inequality is similar to Menshikov's inequality (1-1) but describes the distribution of the *volume* of a cluster rather than of its radius. In the case of percolation on a transitive graph, we obtain in particular that

$$\frac{d}{dp}\log \boldsymbol{P}_p(|K|\geq n)\geq \frac{1}{2p(1-p)}\left[\frac{(1-e^{-\lambda})n}{\lambda\sum_{m=1}^{\lceil n/\lambda\rceil}\boldsymbol{P}_p(|K|\geq m)}-1\right] \tag{1-5}$$

for each $n\geq 0$, $\lambda>0$, and $0\leq p<1$, where K is the cluster of some vertex v and $|K|$ is the number of vertices it contains. We will typically apply this inequality with $\lambda=1$, but the freedom to change λ is sometimes useful for optimizing constants.

We derive (1-5) by introducing a *ghost field* as in [Aizenman and Barsky 1987], i.e., an independent Bernoulli process $\mathcal{G}$ on the vertices of G such that $\mathcal{G}(v)=1$ with probability $1-e^{-\lambda/n}$ for each vertex v of G. We call vertices with $\mathcal{G}(v)=1$ *green*. We then apply the two-function OSSS inequality where A is the event that $|K|\geq n$ and B is the event that K includes a green vertex, and our algorithm simply examines the ghost field at every site and then explores the cluster of each green vertex it discovers. In particular, this algorithm has the property that *the revealment of an edge is equal to the magnetization* up to a factor of 2; see (3-2).

In the remainder of the introduction we describe consequences of the differential inequality (1-5) and of its generalization to the random cluster model.

1A. ***Critical exponent inequalities for percolation.*** In this section we discuss the applications of our differential inequality (1-5) to rigorously establish inequalities between critical exponents in percolation. We first recall the definition of Bernoulli bond percolation, referring the reader to, e.g., [Grimmett 1999] for further background. Let $G=(V,E)$ be a connected, locally finite, transitive graph, such as the hypercubic lattice $\mathbb{Z}^d$. Here, *locally finite* means that every vertex has finite degree, and *transitive* means that for any two vertices x and y of G, there is an automorphism of G mapping x to y. In *Bernoulli bond percolation*, each edge of G is either deleted (closed) or retained (open) independently at random with retention probability $p\in[0,1]$ to obtain a random subgraph ω_p of G. The connected components of ω_p are referred to as *clusters*. We write $\boldsymbol{P}_p$ and $\boldsymbol{E}_p$ for probabilities and expectations taken with respect to the law of ω_p.

It is expected that the behaviour of various quantities describing percolation at and near the critical parameter

$$p_c=\inf\{p\in[0,1]:\omega_p \text{ has an infinite cluster a.s.}\}$$

are described by *critical exponents*. For example, it is predicted that for each $d\geq 2$ there exist exponents β,γ,δ, and Δ such that percolation on $\mathbb{Z}^d$ satisfies

$$\boldsymbol{P}_p(|K|=\infty)\approx(p-p_c)^{\beta}\quad\text{as } p\downarrow p_c,\qquad \boldsymbol{E}_p[|K|]\approx(p_c-p)^{-\gamma}\quad\text{as } p\uparrow p_c,$$
$$\boldsymbol{P}_{p_c}(|K|\geq n)\approx n^{-1/\delta}\quad\text{as } n\uparrow\infty,\qquad \boldsymbol{E}_p[|K|^k]\approx(p_c-p)^{-(k-1)\Delta+\gamma}\quad\text{as } p\uparrow p_c,$$

where K is the cluster of the origin and $\approx$ means that the ratio of the logarithms of the two sides tends to 1 in the appropriate limit. Proving the existence of and computing these critical exponents is considered to be a central problem in mathematical physics. While important progress has been made in two dimensions [Kesten 1987; Smirnov and Werner 2001; Smirnov 2001; Lawler et al. 2002], in high dimensions [Hara and Slade 1990; Aizenman and Newman 1984; Barsky and Aizenman 1991; Nguyen 1987; Fitzner and van der Hofstad 2017], and in various classes of infinite-dimensional graphs [Schonmann 2001; 2002; Hutchcroft 2017; 2019], the entire picture remains completely open in dimensions $3 \le d \le 6$.

A further central prediction of the nonrigorous theory is that, if they exist, these exponents should satisfy the *scaling relations*

$$\gamma = \beta(\delta - 1) \quad \text{and} \quad \beta\delta = \Delta \tag{1-6}$$

in every dimension. (There are also two further scaling relations involving the exponents α, ν, and η, which we have not introduced.) See [Grimmett 1999] for a heuristic derivation of these exponents for mathematicians and, e.g., [Cardy 1996] for more physical derivations. The heuristic derivations of (1-6) do not rely on any special features of percolation, and the scaling relations (1-6) are expected to hold for any natural model of random media undergoing a continuous phase transition.

A rigorous proof of (1-6) remains elusive. Special cases in which progress has been made include the two-dimensional case, where the scaling relations (1-6) were proven by Kesten [1987], and the high-dimensional case, where it has been proven rigorously [Aizenman and Newman 1984; Hara and Slade 1990; Nguyen 1987; Fitzner and van der Hofstad 2017] that the exponents take their *mean-field* values of $\beta = 1,\ \gamma = 1,\ \delta = 2$, and $\Delta = 2$, from which it follows that (1-6) holds. (See, e.g., [Fitzner and van der Hofstad 2017; Slade 2006] for a detailed overview of what is known in high-dimensional percolation.) See also [Vanneuville 2019] for related results on two-dimensional Voronoi percolation. Aside from this, progress on the rigorous understanding of (1-6) has been limited to proving *inequalities* between critical exponents. In particular, it is known that

$$1 \le \beta(\delta - 1), \quad \beta\delta \ge 2, \quad \frac{\gamma\delta}{\delta - 1} \ge 2, \quad \frac{\gamma\delta}{\delta - 1} \le \Delta, \quad \text{and} \quad 2\gamma \ge \Delta, \tag{1-7}$$

whenever these exponents are well-defined: The first of these inequalities is due to Aizenman and Barsky [1987], the second, third, and fourth are due to Newman [1986; 1987a; 1987b], and the fifth is due to Aizenman and Newman [1984]. All of these inequalities are saturated when the exponents take their mean-field values, and the fourth is expected to be an equality in every dimension. These inequalities are complemented by the *mean-field bounds*

$$\beta \le 1, \qquad \gamma \ge 1, \qquad \delta \ge 2, \quad \text{and} \quad \Delta \ge 2, \tag{1-8}$$

which were first proven to hold in [Chayes and Chayes 1987], [Aizenman and Newman 1984], [Aizenman and Barsky 1987], and [Durrett and Nguyen 1985], respectively. See [Grimmett 1999, Chapters 9 and 10] for further details, and [Menshikov 1986; Newman 1986; 1987a; Duminil-Copin and Tassion 2016; Duminil-Copin et al. 2019b] for alternative proofs of some of these inequalities.

Our first application of the differential inequality (1-5) is to rigorously prove two new critical exponent inequalities, namely that

$$\gamma \le \delta - 1 \quad \text{and} \quad \Delta \le \gamma + 1. \tag{1-9}$$

Note that the inequalities of (1-9) are consistent with the conjectural scaling relations (1-6) due to the mean-field bound $\beta \leq 1$, and are saturated when the relevant exponents take their mean-field values. The first of these inequalities is particularly interesting as it points in a different direction to the previously known inequalities given in (1-7).

We will deduce (1-9) as a corollary of the following two theorems, which are derived from (1-5) and which give more precise quantitative versions of these critical exponent inequalities. The first of these theorems relates the distribution of the volume of a critical cluster to the distribution of the volume of a subcritical cluster. It implies the critical exponent inequalities $\gamma \leq \delta - 1$ and $\Delta \leq \delta$. Recall that we write K for the cluster of some arbitrarily chosen vertex.

Theorem 1.1. *Let G be an infinite, connected, locally finite transitive graph, and suppose that there exist constants $C > 0$ and $\delta > 1$ such that*

$$\boldsymbol{P}_{p_c}(|K| \geq n) \leq C n^{-1/\delta}$$

for every $n \geq 1$. Then the following hold:

(1) *There exist positive constants c and C' such that*

$$\boldsymbol{P}_p(|K| \geq n) \leq C' n^{-1/\delta} \exp[-c(p_c - p)^{\delta} n]$$

for every $0 \leq p < p_c$ and $n \geq 1$.

(2) *There exists a constant C'' such that*

$$\boldsymbol{E}_p[|K|^k] \leq k! \left[\frac{C''}{p_c - p} \right]^{(\delta-1)+(k-1)\delta}$$

for every $0 \leq p < p_c$ and $k \geq 1$.

The next theorem bounds the growth of the k-th moment of the cluster volume as $p \uparrow p_c$ in terms of the growth of the first moment as $p \uparrow p_c$. It implies the critical exponent inequality $\Delta \leq \gamma + 1$.

Theorem 1.2. *Let G be an infinite, connected, locally finite transitive graph, and suppose that there exist constants $C > 0$ and $\gamma \geq 0$ such that*

$$\boldsymbol{E}_p[|K|] \leq C(p_c - p)^{-\gamma}$$

for every $n \geq 1$ and $0 \leq p < p_c$. Then there exists a constant C' such that

$$\boldsymbol{E}_p[|K|^k] \leq k! \left[\frac{C'}{p_c - p} \right]^{\gamma+(k-1)(\gamma+1)}$$

for every $0 \leq p < p_c$ and $k \geq 1$.

In light of the results of [Hutchcroft 2020], Theorem 1.1 also has the following consequence for percolation on unimodular transitive graphs of exponential growth. Here, the *growth* of a transitive graph G is defined to be $\operatorname{gr}(G) = \lim_{n\to\infty} |B(v,n)|^{1/n}$ where v is a vertex of G and $|B(v,n)|$ is the ball of radius n around v. See [Hutchcroft 2019; 2020] for more on what is known concerning percolation on such transitive graphs.

Corollary 1.3. *For every $g > 1$ and $M < \infty$ there exist constants $C = C(g, M)$ and $A = A(g, M)$ such that for every unimodular transitive graph G with degree at most M and* $\mathrm{gr}(G) \geq g$, *the bound*

$$\boldsymbol{E}_p[|K|^k] \leq C(p_c - p)^{-Ak} \tag{1-10}$$

holds for every $0 \leq p < p_c$ and $k \geq 1$.

1B. ***The random cluster model.*** In this section we discuss generalizations and applications of (1-5) to the random cluster model (a.k.a. FK-percolation). Since its introduction by Fortuin and Kasteleyn [1972], the random cluster model has become recognized as the archetypal example of a dependent percolation model, and is closely connected to the Ising and Potts models. We refer the reader to [Grimmett 2006] for further background on the model. We expect that the results in this section will also generalize to other models for which sharpness has been proven via the methods of [Duminil-Copin et al. 2019b].

We begin by defining the random cluster model, which we do at the natural generality of *weighted graphs*. We will take a slightly unconventional approach to allow for a unified treatment of short- and long-range models. In this paper, a *weighted graph* $G = (G, J)$ is defined to be a countable graph $G = (V, E)$ together with an assignment of positive *coupling constants* $\{J_e : e \in E\}$ such that for each vertex of G, the sum of the coupling constants J_e over all e adjacent to v is finite. A graph automorphism of G is a weighted graph automorphism of (G, J) if it preserves the coupling constants, and a weighted graph is said to be *transitive* if for every $x, y \in V$ there is an automorphism sending x to y. Note that our weighted graphs are *not* required to be locally finite.

Let $(G = (V, E), J)$ be a weighted graph with V finite, so that $\sum_{e \in E} J_e < \infty$. (Since we did not assume that G is simple, it is possible for the edge set to be infinite.) For each $q > 0$ and $\beta \geq 0$, we let the *random cluster measure* $\phi_{G,\beta,q}$ be the purely atomic probability measure on $\{0, 1\}^E$ defined by

$$\phi_{G,\beta,q}(\{\omega\}) = \frac{1}{Z_{G,\beta,q}} q^{\#\text{clusters}(\omega)} \prod_{e \in E} (e^{\beta J_e} - 1)^{\omega(e)},$$

where $Z_{G,\beta,q}$ is a normalizing constant. In particular, $\phi_{G,\beta,q}$ is supported on configurations containing at most finitely many edges. It is easily verified that this measure is well-defined under the above hypotheses, that is, that $Z_{G,\beta,q} < \infty$. If $q = 1$ and $J_e \equiv 1$, the measure $\phi_{G,\beta,q}$ is simply the law of Bernoulli bond percolation with retention probability $\beta = -\log(1 - p)$. Similarly, if $q = 1$ and the coupling constants are nonconstant then the measure $\phi_{G,\beta,q}$ is the law of *inhomogeneous Bernoulli bond percolation*.

Now suppose that G is an infinite weighted graph. For each $q \geq 1$, we define the *free* and *wired* random cluster measures $\phi^{\mathrm{f}}_{\beta,q}$ and $\phi^{\mathrm{w}}_{\beta,q}$ on G by taking limits along finite subgraphs of G with either free or wired boundary conditions. Let $(V_n)_{n \geq 1}$ be an increasing sequence of finite subsets of V with $\bigcup_{n \geq 1} V_n = V$. For each $n \geq 1$, we define G_n to be the subgraph of G induced by V_n and let G_n^* be the graph obtained by identifying all vertices in $V \setminus V_n$ and deleting all self-loops that are created. Both G_n and G_n^* inherit the coupling constants of G in the natural way. It is shown in [Grimmett 2006, Chapter 4] that if $q \geq 1$ and $\beta \geq 0$ then the weak limits

$$\phi^{\mathrm{f}}_{G,\beta,q} := \operatorname*{w-lim}_{n \to \infty} \phi_{G_n,\beta,q} \quad \text{and} \quad \phi^{\mathrm{w}}_{G,\beta,q} := \operatorname*{w-lim}_{n \to \infty} \phi_{G_n^*,\beta,q}$$

are well-defined and do not depend on the choice of exhaustion $(V_n)_{n \geq 1}$ for every $q \geq 1$ and $n \geq 1$. (It is not known whether these infinite volume limits are well-defined when $q < 1$.) From now on we will drop the G from our notation and write simply $\phi^{\mathrm{f}}_{\beta,q}$ and $\phi^{\mathrm{w}}_{\beta,q}$. Note that $\phi^{\mathrm{w}}_{\beta,q}$ stochastically dominates $\phi^{\mathrm{f}}_{\beta,q}$ for each fixed $\beta \geq 0$ and $q \geq 1$, and that for each $q \geq 1$, $\# \in \{\mathrm{w}, \mathrm{f}\}$, and $0 \leq \beta_1 \leq \beta_2$, the measure $\phi^{\#}_{\beta_2,q}$ stochastically dominates $\phi^{\#}_{\beta_1,q}$.

The generalization of the differential inequality (1-5) to the random cluster model may be stated as follows. Here $\left(\frac{d}{d\beta}\right)_+$ denotes the *lower-right Dini derivative*, which we introduce properly in Section 2B.

Proposition 1.4. *Let (G, J) be an infinite transitive weighted graph, and let $q \geq 1$ and $\# \in \{\mathrm{f}, \mathrm{w}\}$. Then*

$$\max_{e\in E}\left[\frac{e^{\beta J_e}-1}{J_e}\right]\left(\frac{d}{d\beta}\right)_+ \log \phi^{\#}_{\beta,q}(|K| \geq n) \geq \frac{1}{2}\left[\frac{(1-e^{-\lambda})n}{\lambda \sum_{m=1}^{\lceil n/\lambda\rceil} \phi^{\#}_{\beta,q}(|K|\geq m)} - 1\right] \tag{1-11}$$

for every $\beta \geq 0$, $\lambda > 0$, and $n \geq 1$.

Our main application of Proposition 1.4 is to establish the following sharpness result for the random cluster model. For each $\# \in \{\mathrm{f}, \mathrm{w}\}$ the *critical parameter* $\beta_c^{\#}$ is defined to be

$$\beta_c^{\#} = \beta_c^{\#}(G, q) = \inf\left\{\beta \geq 0 : \phi^{\#}_{G,\beta,q}(|K_v| = \infty) > 0 \text{ for some } v \in V\right\}.$$

We always have that $\beta_c^{\mathrm{w}} \leq \beta_c^{\mathrm{f}}$ by stochastic domination. It is known that $\beta_c^{\mathrm{w}} = \beta_c^{\mathrm{f}}$ for the random cluster model on $\mathbb{Z}^d$ and other transitive amenable graphs, while it is believed that strict inequality should hold for $q > 2$ in the nonamenable case [Grimmett 2006, Chapter 10]. It was shown in [Duminil-Copin et al. 2019b] that the following holds for every connected, locally finite, transitive graph, every $q \geq 1$, and every $\# \in \{\mathrm{f}, \mathrm{w}\}$:

(1) If $\beta < \beta_c^{\#}$ then there exist positive constants C_β, c_β such that

$$\phi^{\#}_{\beta,q}(R \geq n) \leq C_\beta e^{-c_\beta n}$$

for every $n \geq 1$, where R is the radius of the cluster of some fixed vertex v as measured by the graph metric on G.

(2) There exists a constant c such that

$$\phi^{\#}_{\beta,q}(|K| = \infty) \geq c(\beta - \beta_c^{\#})$$

for every $\beta > \beta_c^{\#}$ with $\beta - \beta_c^{\#}$ sufficiently small.

The following theorem improves this result by establishing an exponential tail for the *volume* rather than the radius and also by applying to long-range models, which were not treated by [Duminil-Copin et al. 2019b]. (Note that in the case of finite-range models on $\mathbb{Z}^d$, the results of [Duminil-Copin et al. 2019b] were known to imply an exponential tail on the volume by earlier conditional results [Grimmett 2006, Section 5.6].)

Theorem 1.5. *Let (G, J) be an infinite transitive weighted graph. Let $q \geq 1$ and $\# \in \{\mathrm{f}, \mathrm{w}\}$. Then the following hold.*

(1) *For every* $0 \le \beta < \beta_c^\#$ *there exist positive constants* C_β *and* c_β *such that*

$$\phi^\#_{\beta,q}(|K| \ge n) \le C_\beta e^{-c_\beta n}$$

for every $n \ge 1$.

(2) *The inequality*

$$\phi^\#_{\beta,q}(|K| = \infty) \ge \frac{\beta - \beta_c^\#}{2 \max_{e \in E}\left[\frac{e^{\beta J_e} - 1}{J_e}\right] + \beta - \beta_c^\#}$$

holds for every $\beta > \beta_c^\#$.

An immediate corollary of Theorem 1.5 is that $\phi^\#_{\beta,q}[|K|] < \infty$ for every $\beta < \beta_c^\#$ under the same hypotheses, which did not follow from the results of [Duminil-Copin et al. 2019b] in the case that the graph G has exponential volume growth. This allows us to apply the method of [Hutchcroft 2016] and the fact that $\phi^{\mathrm{f}}_{\beta,q}$ is weakly left-continuous in β for each $q \ge 1$ [Grimmett 2006, Proposition 4.28c] to deduce the following corollary for the random cluster model on transitive graphs of exponential growth. This adaptation has already been carried out in the case $q = 2$ (the FK-Ising model) by Raoufi [2018].

Corollary 1.6. *Let* G *be a connected*, *locally finite*, *transitive graph of exponential growth and let* $q \ge 1$. *Then* $\phi^{\mathrm{f}}_{\beta_c^{\mathrm{f}},q}(|K| = \infty) = 0$.

Finally, we generalize Theorems 1.1 and 1.2 to the random cluster model.

Theorem 1.7. *Let* (G, J) *be an infinite transitive weighted graph. Let* $\beta_0 > 0,\ q \ge 1$, *and* $\# \in \{\mathrm{f}, \mathrm{w}\}$, *and suppose that there exist constants* $C > 0$ *and* $\delta > 1$ *such that*

$$\phi^\#_{\beta_0,q}(|K| \ge n) \le C n^{-1/\delta}$$

for every $n \ge 1$. *Then the following hold*:

(1) *There exist positive constants* c_1 *and* C_1 *such that*

$$\phi^\#_{\beta,q}(|K| \ge n) \le C_1 n^{-1/\delta} \exp[-c_1(\beta_0 - \beta)^\delta n]$$

for every $n \ge 1$ *and* $0 \le \beta < \beta_0$.

(2) *There exists a positive constant* c_2 *such that*

$$\phi^\#_{\beta,q}[|K|^k] \le k![c_2(\beta_0 - \beta)]^{-\delta k + 1}$$

for every $k \ge 1$ *and* $0 \le \beta < \beta_0$.

Theorem 1.8. *Let* (G, J) *be an infinite transitive weighted graph. Let* $\beta_0 > 0,\ q \ge 1$, *and* $\# \in \{\mathrm{f}, \mathrm{w}\}$, *and suppose that there exist constants* $C > 0$ *and* $\gamma \ge 0$ *such that*

$$\phi^\#_{\beta,q}[|K|] \le C(\beta_0 - \beta)^{-\gamma}$$

for every $n \ge 1$. *Then there exists a positive constant* c *such that*

$$\phi^\#_{\beta,q}[|K|^k] \le k![c(\beta_0 - \beta)]^{-(k-1)(\gamma+1)-\gamma}$$

for every $n, k \ge 1$ *and* $0 \le \beta < \beta_0$.

It follows from these theorems that the critical exponent inequalities $\gamma \le \delta - 1$ and $\Delta \le \gamma + 1$ hold for the random cluster model whenever these exponents are well-defined, $\delta > 1$ and $q \ge 1$.

We remark that the previous literature on critical exponents for the random cluster model with $q \notin \{1, 2\}$ seems rather limited, although the sharpness results of [Duminil-Copin et al. 2019b] imply the mean-field bound $\beta \le 1$. It is also known that the exponent inequalities we derive here are sharp in mean-field for $q \in [1, 2)$, where the exponents are the same as for percolation [Bollobás et al. 1996]. Note that it is expected that when q is large the random cluster model undergoes a *discontinuous* (first-order) phase transition; see [Bollobás et al. 1996; Laanait et al. 1991; Duminil-Copin et al. 2016; Ray and Spinka 2019].

Remark 1.9. Note that our methods still yield useful bounds in the case $\delta \le 1$, but the exact forms of the inequalities will be different. For example, in Theorem 1.7 if $\delta = 1$ one would obtain that there exist constants C' and c such that

$$\phi^{\#}_{\beta,q}(|K| \ge n) \le C' n^{-1} \exp\left[-\frac{c(\beta_0 - \beta)n}{-\log(\beta_0 - \beta)}\right],$$

while if $\delta < 1$ then one would simply obtain that there exist constants C' and c such that

$$\phi^{\#}_{\beta,q}(|K| \ge n) \le C' n^{-1/\delta} \exp[-c(\beta_0 - \beta)n].$$

Such inequalities are not relevant for percolation due to the mean-field lower bound $\delta \ge 2$, but may be useful for the random-cluster model. Our method also applies unproblematically to more complicated bounds, so that one could convert, say, a critical bound of the form $\phi^{\#}_{\beta_c,q}(|K| \ge n) \le C n^{-1/\delta} \log^{\alpha}(n+1)$ with $\delta > 1$ and $\alpha \in \mathbb{R}$ into a subcritical bound of the form

$$\phi^{\#}_{\beta,q}(|K| \ge n) \le C' n^{-1/\delta} \log^{\alpha}(n+1) \exp\left[-\frac{c(\beta_c - \beta)^{\delta} n}{(-\log(\beta_c - \beta))^{\alpha\delta}}\right].$$

2. Background

2A. *Monotonic measures.* Let A be a countable set. A probability measure μ on $\{0, 1\}^A$ is said to be *positively associated* if

$$\mu(f(\omega)g(\omega)) \ge \mu(f(\omega))\mu(g(\omega))$$

for every pair of increasing functions $f, g : \{0, 1\}^A \to \mathbb{R}$, and is said to be *monotonic* if

$$\mu(\omega(e) = 1 \mid \omega|_F = \xi) \ge \mu(\omega(e) = 1 \mid \omega|_F = \zeta)$$

whenever $F \subset A$, $e \in A$, and $\xi, \zeta \in \{0, 1\}^F$ are such that $\xi \ge \zeta$. It follows immediately from this definition that if μ is a monotonic measure on $\{0, 1\}^A$ and ν is a monotonic measure on $\{0, 1\}^B$, then the product measure $\mu \otimes \nu$ is monotonic on $\{0, 1\}^{A \sqcup B}$.

Monotonic measures are positively associated, but positively associated measures need not be monotonic; see [Grimmett 2006, Chapter 2]. Indeed, it is proven in [Grimmett 2006, Theorem 2.24] that if

A is finite and μ gives positive mass to every element of $\{0,1\}^A$ then it is monotonic if and only if it satisfies the *FKG lattice condition*, which states that

$$\mu(\omega_1 \vee \omega_2)\mu(\omega_1 \wedge \omega_2) \geq \mu(\omega_1)\mu(\omega_2)$$

for every $\omega_1, \omega_2 \in \{0,1\}^A$. In particular, it follows readily from this that the random cluster measures with $q \geq 1$ on any (finite or countably infinite) weighted graph (G, J) are monotonic.

2B. ***Derivative formulae.*** Let $G = (G, J)$ be a *finite* weighted graph. Then for every function $F : \{0,1\}^E \to \mathbb{R}$, we have the derivative formula [Grimmett 2006, Theorem 3.12]

$$\frac{d}{d\beta}\phi_{\beta,q}[F(\omega)] = \sum_{e \in E} \frac{J_e}{e^{\beta J_e} - 1} \operatorname{Cov}_{\phi_{\beta,q}}[F(\omega), \omega(e)], \tag{2-1}$$

where we write $\operatorname{Cov}_\mu[X, Y] = \mu(XY) - \mu(X)\mu(Y)$ for the covariance of two random variables X and Y under the measure μ.

To discuss the generalization of this derivative formula to the infinite volume case, we must first introduce *Dini derivatives*, referring the reader to [Kannan and Krueger 1996] for further background. The *lower-right Dini derivative* of a function $f : [a, b] \to \mathbb{R}$ is defined to be

$$\left(\frac{d}{dx}\right)_+ f(x) = \liminf_{\varepsilon \downarrow 0} \frac{f(x+\varepsilon) - f(x)}{\varepsilon}$$

for each $x \in [a, b)$. Note that if $f : [a, b] \to \mathbb{R}$ is increasing then we have that

$$f(b) - f(a) \geq \int_a^b \left(\frac{d}{dx}\right)_+ f(x)\, \mathrm{d}x,$$

so that we may use differential inequalities involving Dini derivatives in essentially the same way that we use standard differential inequalities. (It is a theorem of Banach [Kannan and Krueger 1996, Theorem 3.6.5] that measurable functions have measurable Dini derivatives, so that the above integral is well-defined.) We also have the validity of the usual logarithmic derivative formula

$$\left(\frac{d}{dx}\right)_+ \log f(x) = \frac{1}{f(x)} \left(\frac{d}{dx}\right)_+ f(x).$$

The following proposition yields a version of (2-1) valid in the infinite-volume setting.

Proposition 2.1. *Let $G = (G, J)$ be a weighted graph and let $F : \{0,1\}^E \to \mathbb{R}$ be an increasing function, and let $\# \in \{\mathrm{f}, \mathrm{w}\}$. Then*

$$\left(\frac{d}{d\beta}\right)_+ \phi^\#_{\beta,q}[F(\omega)] \geq \sum_{e \in E} \frac{J_e}{e^{\beta J_e} - 1} \operatorname{Cov}_{\phi^\#_{\beta,q}}[F(\omega), \omega(e)] \tag{2-2}$$

for every $\beta \geq 0$.

Proof. We prove the claim in the case $\# = \mathrm{f}$, the case $\# = \mathrm{w}$ being similar. Fix $\beta_0 \geq 0$ and let A be a finite set of edges. Let $(V_n)_{n \geq 1}$ be an exhaustion of V, let G_n be the subgraph of G induced by V_n and let E_n

be the edge set of G_n. For each $n \geq 1$, $\beta \geq 0$ and $q \geq 1$ we define $\phi_{G_n,\beta,\beta_0,q,A}(\{\omega\})$ by

$$\phi_{G_n,\beta,\beta_0,q,A}(\{\omega\}) = \frac{1}{Z} q^{\#\text{clusters}(\omega)} \prod_{e \in A \cap E_n} (e^{\beta J_e} - 1)^{\omega(e)} \prod_{e \in E_n \setminus A} (e^{\beta_0 J_e} - 1)^{\omega(e)}$$

for an appropriate normalizing constant $Z = Z(n, \beta, q, A)$. The usual proof of the existence of the infinite volume random cluster measures yields that the measures $\phi_{G_n,\beta,\beta_0,q,A}$ converge weakly to a limiting measure $\phi^{\mathrm{f}}_{\beta,\beta_0,q,A}$ as $n \to \infty$. Using the assumption that A is finite, it is straightforward to adapt the usual proof of (2-1) to show that $\phi^{\mathrm{f}}_{\beta,\beta_0,q,A}[F(\omega)]$ is differentiable and that

$$\frac{d}{d\beta} \phi^{\mathrm{f}}_{\beta,\beta_0,q,A}[F(\omega)] = \sum_{e \in A} \frac{J_e}{e^{\beta J_e} - 1} \operatorname{Cov}_{\phi^{\mathrm{f}}_{\beta,\beta_0,q,A}}[F(\omega), \omega(e)]$$

for every function $F(\omega) \to \mathbb{R}$ with $\phi^{\mathrm{f}}_{\beta_0,q}[F(\omega)] < \infty$. On the other hand, the FKG property implies that $\phi^{\mathrm{f}}_{\beta,q}$ stochastically dominates $\phi^{\mathrm{f}}_{\beta,\beta_0,q,A}$ for every $\beta \geq \beta_0$, and we deduce that if F is increasing then

$$\begin{aligned}
\liminf_{\beta \downarrow \beta_0} \frac{\phi^{\mathrm{f}}_{\beta,q}[F(\omega)] - \phi^{\mathrm{f}}_{\beta_0,q}[F(\omega)]}{\beta - \beta_0} &\geq \liminf_{\beta \downarrow \beta_0} \sup_A \frac{1}{\beta - \beta_0} \big[\phi^{\mathrm{f}}_{\beta,\beta_0,q,A}[F(\omega)] - \phi^{\mathrm{f}}_{\beta_0,\beta_0,q,A}[F(\omega)]\big] \\
&\geq \sup_A \liminf_{\beta \downarrow \beta_0} \frac{1}{\beta - \beta_0} \big[\phi^{\mathrm{f}}_{\beta,\beta_0,q,A}[F(\omega)] - \phi^{\mathrm{f}}_{\beta_0,\beta_0,q,A}[F(\omega)]\big] \\
&= \sup_A \sum_{e \in A} \frac{J_e}{e^{\beta J_e} - 1} \operatorname{Cov}_{\phi^{\mathrm{f}}_{\beta_0,\beta_0,q,A}}[F(\omega), \omega(e)] \\
&= \sum_{e \in E} \frac{J_e}{e^{\beta_0 J_e} - 1} \operatorname{Cov}_{\phi^{\mathrm{f}}_{\beta_0,q}}[F(\omega), \omega(e)],
\end{aligned}$$

where the final equality follows by positive association. The claim follows since $\beta_0 \geq 0$ was arbitrary. □

2C. ***Decision trees and the OSSS inequality.*** Let $\mathbb{N} = \{1, 2, \ldots\}$, and let E be a countable set. A *decision tree* is a function $T : \{0, 1\}^E \to E^{\mathbb{N}}$ from subsets of E to infinite E-valued sequences with the property that $T_1(\omega) = e_1$ for some fixed $e_1 \in E$, and for each $n \geq 2$ there exists a function $S_n : (E \times \{0, 1\})^{n-1} \to E$ such that

$$T_n(\omega) = S_n[(T_i, \omega(T_i))_{i=1}^{n-1}].$$

In other words, T is a deterministic procedure for querying the values of $\omega \in \{0, 1\}^E$, that starts by querying the value of $\omega(e_1)$ and chooses which values to query at each subsequent step as a function of the values it has already observed.

Now let μ be a probability measure on $\{0, 1\}^E$ and let ω be a random variable with law μ. Given a decision tree T and $n \geq 1$ we let $\mathcal{F}_n(T)$ be the σ-algebra generated by the random variables $\{T_i(\omega) : 1 \leq i \leq n\}$ and let $\mathcal{F}(T) = \bigcup \mathcal{F}_n(T)$. For each measurable function $f : \{0, 1\}^E \to [-1, 1]$, we say that T *computes* f if $f(\omega)$ is measurable with respect to the μ-completion of $\mathcal{F}(T)$. By the martingale convergence theorem, if f is μ-integrable this is equivalent to the statement that

$$\mu[f(\omega) \mid \mathcal{F}_n(T)] \xrightarrow[n \to \infty]{} f(\omega) \quad \mu\text{-a.s.}$$

For each $e \in E$, we define the *revealment probability*

$$\delta_e(T,\mu) = \mu(\text{there exists } n \geq 1 \text{ such that } T_n(\omega) = e).$$

Finally, following [O'Donnell et al. 2005], we define for each probability measure μ on $\{0,1\}^E$ and each pair of measurable functions $f, g : \{0,1\}^E \to \mathbb{R}$ the quantity

$$\mathrm{CoVr}_\mu[f,g] = \mu \otimes \mu[|f(\omega_1) - g(\omega_2)|] - \mu[|f(\omega_1) - g(\omega_1)|],$$

where ω_1, ω_2 are drawn independently from the measure μ, so that if f and g are $\{0,1\}$-valued then

$$\mathrm{CoVr}_\mu[f,g] = 2\,\mathrm{Cov}_\mu[f,g] = 2\mu\bigl(f(\omega) = g(\omega) = 1\bigr) - 2\mu\bigl(f(\omega) = 1\bigr)\mu\bigl(g(\omega) = 1\bigr). \tag{2-3}$$

We are now ready to state Duminil-Copin, Raoufi, and Tassion's generalization of the OSSS inequality to monotonic measures [Duminil-Copin et al. 2019b].

Theorem 2.2. *Let E be a finite or countably infinite set and let μ be a monotonic measure on $\{0,1\}^E$. Then for every pair of measurable, μ-integrable functions $f, g : \{0,1\}^E \to \mathbb{R}$ with f increasing and every decision tree T computing g we have that*

$$\tfrac{1}{2}|\mathrm{CoVr}_\mu[f,g]| \leq \sum_{e \in E} \delta_e(T,\mu)\,\mathrm{Cov}_\mu[f,\omega(e)].$$

Remark 2.3. In [Duminil-Copin et al. 2019b], only the special case of Theorem 2.2 in which E is finite and $f = g$ is stated. The version with E finite but f not necessarily equal to g follows by an easy modification of their proof, identical to that carried out in [O'Donnell et al. 2005, Section 3.3] — note in particular that when running this modified proof only f is required to be increasing. The restriction that E is finite can be removed via a straightforward Martingale argument [Duminil-Copin et al. 2019b, Remark 2.4].

The statement above will be somewhat inconvenient in our analysis as the algorithm we use is naturally described as a *parallel algorithm* rather than a serial algorithm. To allow for such parallelization, we define a *decision forest* to be a collection of decision trees $F = \{T^i : i \in I\}$ indexed by a countable set I. Given a decision forest $F = \{T^i : i \in I\}$ we let $\mathcal{F}(F)$ be the smallest σ-algebra containing all of the σ-algebras $\mathcal{F}(T^i)$. Given a measure μ on $\{0,1\}^E$, a function $f : \{0,1\}^E \to \mathbb{R}$ and a decision forest F, we say that F *computes* f if f is measurable with respect to the μ-completion of the σ-algebra $\mathcal{F}(F)$. We also define the revealment probability $\delta_e(F,\mu)$ to be the probability under μ that there exist $i \in I$ and $n \geq 1$ such that $T^i_n(\omega) = e$.

Corollary 2.4. *Let E be a finite or countably infinite set and let μ be a monotonic measure on $\{0,1\}^E$. Then for every pair of measurable, μ-integrable functions $f, g : \{0,1\}^E \to \mathbb{R}$ with f increasing and every decision forest F computing g we have that*

$$\tfrac{1}{2}|\mathrm{CoVr}_\mu[f,g]| \leq \sum_{e \in E} \delta_e(F,\mu)\,\mathrm{Cov}_\mu[f,\omega(e)].$$

Proof. We may assume that $I = \{1, 2, \ldots\}$. The claim may be deduced from Theorem 2.2 by "serializing" the decision forest F into a decision tree T. This can be done, for example, by executing the j-th step of the decision tree T^i at the time p_i^j where p_i is the i-th prime, and re-querying the first input queried by T^1 at all times that are not prime powers. This decision tree T clearly computes the same functions as the decision forest F and has $\delta_e(T,\mu)=\delta_e(F,\mu)$ for every $e\in E$, so the claim follows from Theorem 2.2. □

3. Derivation of the differential inequality

Given a graph $G=(V,E)$, a vertex v and a configuration $\omega\in\{0,1\}^E$, we write $K_v=K_v(\omega)$ for the cluster of v in ω.

Proposition 3.1. *Let $G=(V,E)$ be a countable graph and μ be a monotonic measure on $\{0,1\}^E$. Then*

$$\sum_{e\in E}\operatorname{Cov}_\mu[\mathbb{1}(|K_v|\geq n),\omega(e)]\geq\left[\frac{(1-e^{-\lambda})-\mu\big[1-e^{-\lambda|K_v|/n}\big]}{2\sup_{u\in V}\mu\big[1-e^{-\lambda|K_u|/n}\big]}\right]\mu(|K_v|\geq n) \tag{3-1}$$

for every $v\in V$, $n\geq 1$ and $\lambda>0$.

Proof. Let $\omega\in\{0,1\}^E$ be a random variable with law μ. Independently of ω, let $\eta\in\{0,1\}^V$ be a random subset of V where vertices are included independently at random with inclusion probability $h=1-e^{-\lambda/n}\leq\lambda/n$. We refer to η as the *ghost field* and call vertices with $\eta(v)=1$ *green*. Let $\mathbb{P}$ and $\mathbb{E}$ denote probabilities and expectations taken with respect to the joint law of ω and η, which is monotonic. Fix a vertex v, and let $f,g:\{0,1\}^{E\cup V}\to\{0,1\}$ be the increasing functions defined by

$$f(\omega,\eta)=\mathbb{1}(|K_v(\omega)|\geq n)\quad\text{and}\quad g(\omega,\eta)=\mathbb{1}(\eta(u)=1\text{ for some }u\in K_v(\omega)).$$

For each $u\in V$, we define T^u to be a decision tree that first queries the status of $\eta(u)$, halts if it discovers that $\eta(u)=0$, and otherwise explores the cluster of u in ω. We now define this decision tree more formally; if the reader is satisfied with the informal description they may safely skip the rest of this paragraph. Fix an enumeration of E and a vertex $u\in V$. Set $T_1^u(\omega,\eta)=u$. If $\eta(u)=0$, set $T_n^u=u$ for every $n\geq 2$ (i.e., halt). If $\eta(u)=1$, we define $T_n^u(\omega,\eta)$ for $n\geq 2$ as follows. At each step of the decision tree, we will have a set of vertices U_n^u, a set of revealed open edges O_n^u, and a set of revealed closed edges C_n^u. We initialize by setting $U_1^u=u$ and $O_n^u=C_n^u=\varnothing$. Suppose that $n\geq 1$ and that we have computed $(U_k^u,O_k^u,C_k^u,T_k^u)$ for $k\leq n$. If every edge with at least one endpoint in U_n^u is either in O_n^u or C_n^u then we set $(U_{n+1}^u,O_{n+1}^u,C_{n+1}^u,T_{n+1}^u)=(U_n^u,O_n^u,C_n^u,T_n^u)$ (i.e., we halt). Otherwise, we set T_{n+1}^u to be the element of the set of edges that touch U_n^u but are not in O_n^u or C_n^u that is minimal with respect to the fixed enumeration of E. If $\omega(T_{n+1}^u)=1$ we set U_{n+1}^u to be the union of U_n^u with the endpoints of T_{n+1}^u, set $O_{n+1}^u=O_n^u\cup\{T_{n+1}^u\}$ and set $C_{n+1}^u=C_n^u$. Otherwise, $\omega(T_{n+1}^u)=0$ and we set $U_{n+1}^u=U_n^u$, set $O_{n+1}^u=O_n^u$ and set $C_{n+1}^u=C_n^u\cup\{T_{n+1}^u\}$.

It is easily verified that this decision tree T^u satisfies

$$\{x\in V\cup E:T_n^u(\omega,\eta)=x\text{ for some }n\geq 1\}=\begin{cases}\{u\} & \eta(u)=0,\\ \{u\}\cup E(K_u(\omega)) & \eta(u)=1,\end{cases}$$

where $E(K_u(\omega))$ is the set of edges with at least one endpoint in $K_u(\omega)$. In particular, we clearly have that the decision forest $F=\{T^u:u\in V\}$ computes g.

Since f and g are increasing and $\{0,1\}$-valued, we may apply Corollary 2.4 and (2-3) to deduce that

$$\operatorname{Cov}_{\mathbb{P}}[f,g] \le \sum_{e\in E}\delta_e(F,\mu)\operatorname{Cov}_{\mathbb{P}}[f,\omega(e)] + \sum_{u\in V}\delta_u(F,\mu)\operatorname{Cov}_{\mathbb{P}}[f,\eta(v)] = \sum_{e\in E}\delta_e(F,\mu)\operatorname{Cov}_{\mu}[f,\omega(e)],$$

where the equality on the right follows since $f(\omega,\eta)=\mathbb{1}(|K_v(\omega)|\ge n)$ is independent of η. Now, an edge e is revealed by $F(\omega,\eta)$ if and only if the cluster of at least one endpoint of e contains a green vertex, and, writing $\eta(A)=\sum_{u\in A}\eta(u)$ for each set $A\subseteq V$, it follows that

$$\delta_e(F,\mu)\le 2\sup_{u\in V}\mathbb{P}(\eta(K_u)\ge 1) = 2\sup_{u\in V}\mu[1-e^{-\lambda|K_u|/n}] \tag{3-2}$$

for every $e\in E$ and hence that

$$\operatorname{Cov}_{\mathbb{P}}[f,g]\le 2\sup_{u\in V}\mu[1-e^{-\lambda|K_u|/n}]\sum_{e\in E}\operatorname{Cov}_{\mu}[f,\omega(e)]. \tag{3-3}$$

To conclude, simply note that

$$\begin{aligned}\operatorname{Cov}_{\mathbb{P}}[f,g] &= \mathbb{P}(|K_v|\ge n,\ \eta(K_v)\ge 1) - \mathbb{P}(\eta(K_v)\ge 1)\mu(|K_v|\ge n)\\ &= \mu\big[(1-e^{-\lambda|K_v|/n})\mathbb{1}(|K_v|\ge n)\big] - \mu[1-e^{-\lambda|K_v|/n}]\mu(|K_v|\ge n)\\ &\ge (1-e^{-\lambda})\mu(|K_v|\ge n) - \mu[1-e^{-\lambda|K_v|/n}]\mu(|K_v|\ge n). \end{aligned}\tag{3-4}$$

Combining (3-3) and (3-4) and rearranging yields the desired inequality. □

Proof of Proposition 1.4. This is immediate from Propositions 2.1 and 3.1 together with the inequality $1-e^{-\lambda|K_v|/n}\le 1\wedge\lambda|K_v|/n$, which yields the bound

$$\phi^{\#}_{\beta,q}[1-e^{-\lambda|K_v|/n}] \le \frac{\lambda}{n}\phi^{\#}_{\beta,q}\left[\frac{n}{\lambda}\wedge|K_v|\right] \le \frac{\lambda}{n}\sum_{m=1}^{\lceil n/\lambda\rceil}\phi^{\#}_{\beta,q}(|K_v|\ge m).$$ □

Taking the limit as $\lambda\downarrow 0$ in Proposition 1.4 yields the following corollary.

Corollary 3.2. *Let (G,J) be an infinite transitive weighted graph, and let $q\ge 1$ and $\#\in\{\mathrm{f},\mathrm{w}\}$. Then*

$$\max_{e\in E}\left[\frac{e^{\beta J_e}-1}{J_e}\right]\left(\frac{d}{d\beta}\right)_{+}\log\phi^{\#}_{\beta,q}(|K|\ge n) \ge \frac{1}{2}\left[\frac{n}{\phi^{\#}_{\beta,q}[|K|]}-1\right] \tag{3-5}$$

for every $n\ge 1$ and $\beta\ge 0$.

Since $\phi^{\#}_{\beta,q}(|K|\ge n)$ is increasing in β, the following inequalities may be obtained by integrating the differential inequalities of Proposition 1.4 and Corollary 3.2: Letting $C(\beta)=\max_{e\in E}[(e^{\beta J_e}-1)/J_e]$ for each $\beta\ge 0$, and noting that $C(\beta)$ is an increasing function of β, we have that

$$\phi^{\#}_{\beta,q}(|K|\ge n)\le\phi^{\#}_{\beta_0,q}(|K|\ge n)\exp\left[-\frac{(1-e^{-1})(\beta_0-\beta)n}{2C(\beta_0)\sum_{m=1}^{n}\phi^{\#}_{\beta_0,q}(|K|\ge m)}+\frac{\beta_0-\beta}{2C(\beta_0)}\right] \tag{3-6}$$

and

$$\phi^{\#}_{\beta,q}(|K|\ge n)\le\phi^{\#}_{\beta_0,q}(|K|\ge n)\exp\left[-\frac{(\beta_0-\beta)n}{2C(\beta_0)\phi^{\#}_{\beta_0,q}[|K|]}+\frac{\beta_0-\beta}{2C(\beta_0)}\right] \tag{3-7}$$

for every $n\ge 1$, $0\le\beta\le\beta_0$, and $q\ge 1$.

4. Analysis of the differential inequality

4A. ***Critical exponent inequalities.*** In this section we apply Proposition 1.4 to prove Theorems 1.1, 1.2, 1.7 and 1.8.

Proof of Theorems 1.1 and 1.7. It suffices to prove Theorem 1.7, of which Theorem 1.1 is a special case. Fix $\beta_0 > 0$, and suppose that there exist constants $C > 0$ and $\delta > 1$ such that $\phi^{\#}_{\beta_0,q}(|K| \geq n) \leq Cn^{-1/\delta}$ for every $n \geq 1$. In this proof, we use $\preceq$ and $\succeq$ to denote inequalities that hold up to positive multiplicative constants depending only on (G, J), δ, C, and β_0. Since $\delta > 1$ we have

$$\sum_{m=1}^{n} \phi^{\#}_{\beta_0,q}(|K| \geq m) \preceq n^{1-1/\delta}$$

for every $n \geq 1$ and hence by (3-6) that

$$\phi^{\#}_{\beta_1,q}(|K| \geq n) \preceq n^{-1/\delta} \exp[-c_1(\beta_0 - \beta_1)n^{1/\delta}]$$

for every $0 \leq \beta_1 < \beta_0$ and $n \geq 1$. This inequality may be summed over n to obtain that

$$\phi^{\#}_{\beta_1,q}[|K|] = \sum_{n\geq 1} \phi^{\#}_{\beta_1,q}(|K| \geq n) \preceq \sum_{n \geq 1} n^{-1/\delta} \exp[-c_1(\beta_0 - \beta_1)n^{1/\delta}] \preceq (\beta_0 - \beta_1)^{-\delta+1} \tag{4-1}$$

for every $0 \leq \beta_1 < \beta_0$ and $n \geq 1$. Thus, we have by (3-7) that

$$\phi^{\#}_{\beta,q}(|K| \geq n) \preceq n^{-1/\delta} \exp\left[-c_2 \frac{(\beta_1 - \beta)n}{(\beta_0 - \beta_1)^{-\delta+1}}\right]$$

for every $0 \leq \beta < \beta_1 < \beta_0$ and $n \geq 1$. Item (1) of the theorem follows by taking $\beta_1 = (\beta_0 + \beta)/2$.

Item (2) of the theorem is a simple analytic consequence of item (1) since we have that

$$\phi^{\#}_{\beta,q}(|K| \geq x) = \phi^{\#}_{\beta,q}(|K| \geq \lceil x \rceil) \preceq x^{-1/\delta} \exp[-c_2(\beta_0 - \beta)^{\delta} x]$$

for every $x > 0$, and consequently, letting $\varepsilon = c_2(\beta_0 - \beta)^{\delta}$ and $\alpha = k - 1 - 1/\delta$, we have that

$$\begin{aligned} \phi^{\#}_{\beta,q}[|K|^k] = k \int_0^\infty x^{k-1} \phi^{\#}_{\beta,q}(|K| \geq x)\, \mathrm{d}x &\preceq k \int_{x=0}^\infty x^{\alpha} e^{-\varepsilon x}\, \mathrm{d}x \\ &= k\varepsilon^{-\alpha-1} \int_{y=0}^\infty y^{\alpha} e^{-y}\, \mathrm{d}y \\ &= k\Gamma(\alpha+1)\varepsilon^{-\alpha-1} \leq k!\varepsilon^{-\alpha-1} \\ &= k![c_2(\beta_0 - \beta)]^{-\delta(k-1)-(\delta-1)} \end{aligned} \tag{4-2}$$

for every $k \geq 1$ and $0 \leq \beta < \beta_0$, where we used the change of variables $y = \varepsilon x$ in the final equality on the first line. □

Proof of Theorems 1.2 and 1.8. This proof is similar to that of Theorem 1.7. Fix $\beta_0 > 0$, and suppose that there exist constants $C > 0$ and $\gamma \geq 0$ such that $\phi^{\#}_{\beta,q}(|K| \geq n) \leq C(\beta_0 - \beta)^{-\gamma}$ for every $0 \leq \beta < \beta_0$. In this proof, we use $\preceq$ and $\succeq$ to denote inequalities that hold up to positive multiplicative constants depending only on (G, J), γ, C, and β_0. By (3-7) and Markov's inequality there exist positive constants c_1 and c_2 such that

$$\phi^{\#}_{\beta,q}(|K| \geq n) \preceq \frac{\phi^{\#}_{\beta_1,q}[|K|]}{n} \exp\left[-c_1 \frac{n(\beta_1-\beta_0)}{\phi^{\#}_{\beta_1,q}[|K|]} n\right] \preceq \frac{1}{(\beta_0-\beta)^{\gamma} n} \exp[-c_2(\beta_1-\beta_0)^{\gamma+1} n]$$

for every $0 \leq \beta < \beta_1 < \beta_0$ and $n \geq 1$. The proof may be concluded by taking $\beta_1 = (\beta_0+\beta)/2$ and performing essentially the same calculation as in (4-2). □

4B. ***Sharpness of the phase transition.*** We next apply Proposition 1.4 to prove Theorem 1.5. The proof is very similar to that given in [Duminil-Copin et al. 2019b]. (Note that the analysis presented there substantially simplified the original analysis of Menshikov [1986].) We include it for completeness since our differential inequality is slightly different, and so that we can optimize the constants appearing in item (2) of Theorem 1.5.

Proof of Theorem 1.5. We begin by defining

$$\begin{aligned}\tilde{\beta}^{\#}_c &= \sup\left\{\beta \geq 0 : \text{there exist } c, C > 0 \text{ such that } \phi^{\#}_{\beta,q}(|K| \geq n) \leq Cn^{-c} \text{ for every } n \geq 1\right\}\\ &= \inf\left\{\beta \geq 0 : \limsup_{n\to\infty} \frac{\log \phi^{\#}_{\beta,q}(|K| \geq n)}{\log n} \geq 0\right\}.\end{aligned}$$

We trivially have that $\tilde{\beta}^{\#}_c \leq \beta^{\#}_c$. Moreover, it is an immediate consequence of Theorem 1.7 that for every $0 \leq \beta < \tilde{\beta}^{\#}_c$ there exist $C_\beta, c_\beta > 0$ such that

$$\phi^{\#}_{\beta,q}(|K| \geq n) \leq C_\beta e^{-c_\beta n} \quad \text{for every } n \geq 1. \tag{4-3}$$

We next claim that $\phi^{\#}_{\beta,q}(|K| = \infty) > 0$ for every $\beta > \tilde{\beta}^{\#}_c$. To this end, write $P_n(\beta) = \phi^{\#}_{\beta,q}(|K| \geq n)$ and $\Sigma_n(\beta) = \sum_{m=0}^{n-1} P_m(\beta)$ and let

$$T_k(\beta) = \frac{1}{\log k} \sum_{n=1}^{k} \frac{1}{n} \phi^{\#}_{\beta,q}(|K| \geq n)$$

for each $\beta \geq 0$ and $k \geq 2$, so that $\lim_{k\to\infty} T_k(\beta) = \phi^{\#}_{\beta,q}(|K| = \infty)$ for each $\beta \geq 0$. Let

$$C(\beta) = \max_{e\in E}[(e^{\beta J_e} - 1)/J_e]$$

as above. Applying Proposition 1.4 with $\lambda = 1$, we obtain that

$$\left(\frac{d}{d\beta}\right)_+ T_k(\beta) \geq \frac{1}{2C(\beta)\log k} \sum_{n=1}^{k} \left[\frac{(1-e^{-1})P_n(\beta)}{\Sigma_n(\beta)} - \frac{P_n(\beta)}{n}\right]$$

for every $\beta \geq 0$ and $k \geq 2$. Using the inequality

$$\frac{P_n}{\Sigma_n} \geq \int_{\Sigma_n}^{\Sigma_{n+1}} \frac{1}{x}\, dx = \log \Sigma_{n+1} - \log \Sigma_n,$$

we deduce that

$$\left(\frac{d}{d\beta}\right)_+ T_k(\beta) \geq \frac{(1-e^{-1})\log \Sigma_{k+1}(\beta)}{2C(\beta)\log k} - \frac{T_k(\beta)}{2C(\beta)}$$

for every $\beta \geq 0$ and $k \geq 1$. Fixing $\tilde{\beta}_c^\# < \beta_1 < \beta_2$ and using that $C(\beta)$ is increasing in β, we deduce that

$$\left(\frac{d}{d\beta}\right)_+ T_k(\beta) \geq \frac{(1-e^{-1})\log \Sigma_{k+1}(\beta_1)}{2C(\beta_2)\log k} - \frac{T_k(\beta_2)}{2C(\beta_1)}$$

for every $\beta_1 \leq \beta \leq \beta_2$ and hence by definition of $\tilde{\beta}_c^\#$ that

$$\limsup_{k\to\infty} \inf_{\beta_1\leq\beta\leq\beta_2} \left(\frac{d}{d\beta}\right)_+ T_k(\beta) \geq \frac{(1-e^{-1})}{2C(\beta_2)} - \frac{\phi^\#_{\beta_2,q}(|K|=\infty)}{2C(\beta_1)}.$$

Integrating this inequality yields that

$$\phi^\#_{\beta_2,q}(|K|=\infty) \geq \limsup_{k\to\infty} \int_{\beta_1}^{\beta_2} \left(\frac{d}{d\beta}\right)_+ T_k(\beta)\,\mathrm{d}\beta \geq \frac{(1-e^{-1})(\beta_2-\beta_1)}{2C(\beta_2)} - \frac{(\beta_2-\beta_1)}{2C(\beta_1)}\phi^\#_{\beta_2,q}(|K|=\infty)$$

which rearranges to give that

$$\phi^\#_{\beta_2,q}(|K|=\infty) \geq \left(\frac{C(\beta_1)}{C(\beta_2)}\right)\frac{(1-e^{-1})(\beta_2-\beta_1)}{2C(\beta_1)+\beta_2-\beta_1} > 0. \tag{4-4}$$

The claim now follows since $\tilde{\beta}_c^\# < \beta_1 < \beta_2$ were arbitrary. We deduce that $\tilde{\beta}_c^\# \geq \beta_c^\#$ and hence that $\tilde{\beta}_c^\# = \beta_c^\#$, so that in particular item (1) of the theorem follows from (4-3).

It remains only to prove item (2), which amounts to improving the constants in (4-4). We apply Proposition 1.4 again to obtain that

$$\left(\frac{d}{d\beta}\right)_+ T_k(\beta) \geq \frac{1}{2C(\beta)\log k}\sum_{n=1}^{k}\left[\frac{(1-e^{-\lambda})P_n(\beta)}{\lambda\Sigma_{\lfloor n/\lambda\rfloor}(\beta)} - \frac{P_n(\beta)}{n}\right]$$

for each $k \geq 2$, $\beta \geq 0$, and $\lambda > 0$. Arguing in a similar way to before, we obtain that

$$\limsup_{k\to\infty} \inf_{\beta_1\leq\beta\leq\beta_2} \left(\frac{d}{d\beta}\right)_+ T_k(\beta) \geq \frac{1-e^{-\lambda}}{2C(\beta_2)} - \frac{\phi^\#_{\beta_2,q}(|K|=\infty)}{2C(\beta_1)}$$

for every $\beta_c^\# < \beta_1 \leq \beta_2$. Sending $\lambda \to \infty$, it follows by elementary analysis that

$$\phi^\#_{\beta',q}(|K|=\infty) \geq \int_{\beta_c^\#}^{\beta'} \frac{1-\phi^\#_{\beta,q}(|K|=\infty)}{2C(\beta)}\,\mathrm{d}\beta \geq \frac{(\beta'-\beta_c^\#)(1-\phi^\#_{\beta',q}(|K|=\infty))}{2C(\beta')}$$

for every $\beta' \geq \beta_c^\#$, which rearranges to give the claim. □

Acknowledgments. We thank Hugo Duminil-Copin and Geoffrey Grimmett for helpful discussions.

References

[Aizenman and Barsky 1987] M. Aizenman and D. J. Barsky, "Sharpness of the phase transition in percolation models", *Comm. Math. Phys.* **108**:3 (1987), 489–526. MR

[Aizenman and Newman 1984] M. Aizenman and C. M. Newman, "Tree graph inequalities and critical behavior in percolation models", *J. Statist. Phys.* **36**:1-2 (1984), 107–143. MR

[Aizenman et al. 1987] M. Aizenman, D. J. Barsky, and R. Fernández, "The phase transition in a general class of Ising-type models is sharp", *J. Statist. Phys.* **47**:3-4 (1987), 343–374. MR

[Barsky and Aizenman 1991] D. J. Barsky and M. Aizenman, "Percolation critical exponents under the triangle condition", *Ann. Probab.* **19**:4 (1991), 1520–1536. MR

[Beekenkamp 2018] T. Beekenkamp, "Sharpness of the percolation phase transition for the contact process on $\mathbb{Z}^d$", preprint, 2018. arXiv

[van den Berg and Kesten 1985] J. van den Berg and H. Kesten, "Inequalities with applications to percolation and reliability", *J. Appl. Probab.* **22**:3 (1985), 556–569. MR

[Bollobás et al. 1996] B. Bollobás, G. Grimmett, and S. Janson, "The random-cluster model on the complete graph", *Probab. Theory Related Fields* **104**:3 (1996), 283–317. MR

[Cardy 1996] J. Cardy, *Scaling and renormalization in statistical physics*, Cambridge Lecture Notes in Physics **5**, Cambridge University Press, 1996. MR

[Chayes and Chayes 1987] J. T. Chayes and L. Chayes, "The mean field bound for the order parameter of Bernoulli percolation", pp. 49–71 in *Percolation theory and ergodic theory of infinite particle systems* (Minneapolis, 1984–1985), edited by H. Kesten, IMA Vol. Math. Appl. **8**, Springer, 1987. MR

[Dereudre and Houdebert 2018] D. Dereudre and P. Houdebert, "Sharp phase transition for the continuum Widom–Rowlinson model", preprint, 2018. arXiv

[Duminil-Copin and Tassion 2016] H. Duminil-Copin and V. Tassion, "A new proof of the sharpness of the phase transition for Bernoulli percolation and the Ising model", *Comm. Math. Phys.* **343**:2 (2016), 725–745. MR

[Duminil-Copin et al. 2016] H. Duminil-Copin, M. Gagnebin, M. Harel, I. Manolescu, and V. Tassion, "Discontinuity of the phase transition for the planar random-cluster and Potts models with $q > 4$", preprint, 2016. arXiv

[Duminil-Copin et al. 2018] H. Duminil-Copin, A. Raoufi, and V. Tassion, "Subcritical phase of d-dimensional Poisson–Boolean percolation and its vacant set", preprint, 2018. arXiv

[Duminil-Copin et al. 2019a] H. Duminil-Copin, A. Raoufi, and V. Tassion, "Exponential decay of connection probabilities for subcritical Voronoi percolation in $\mathbb{R}^d$", *Probab. Theory Related Fields* **173**:1-2 (2019), 479–490. MR

[Duminil-Copin et al. 2019b] H. Duminil-Copin, A. Raoufi, and V. Tassion, "Sharp phase transition for the random-cluster and Potts models via decision trees", *Ann. of Math.* (2) **189**:1 (2019), 75–99. MR

[Durrett and Nguyen 1985] R. Durrett and B. Nguyen, "Thermodynamic inequalities for percolation", *Comm. Math. Phys.* **99**:2 (1985), 253–269. MR

[Fitzner and van der Hofstad 2017] R. Fitzner and R. van der Hofstad, "Mean-field behavior for nearest-neighbor percolation in $d > 10$", *Electron. J. Probab.* **22** (2017), paper no. 43. MR

[Fortuin and Kasteleyn 1972] C. M. Fortuin and P. W. Kasteleyn, "On the random-cluster model, I: Introduction and relation to other models", *Physica* **57** (1972), 536–564. MR

[Grimmett 1999] G. Grimmett, *Percolation*, 2nd ed., Grundlehren der Mathematischen Wissenschaften **321**, Springer, 1999. MR

[Grimmett 2006] G. Grimmett, *The random-cluster model*, Grundlehren der Mathematischen Wissenschaften **333**, Springer, 2006. MR

[Hara and Slade 1990] T. Hara and G. Slade, "Mean-field critical behaviour for percolation in high dimensions", *Comm. Math. Phys.* **128**:2 (1990), 333–391. MR

[Hutchcroft 2016] T. Hutchcroft, "Critical percolation on any quasi-transitive graph of exponential growth has no infinite clusters", *C. R. Math. Acad. Sci. Paris* **354**:9 (2016), 944–947. MR

[Hutchcroft 2017] T. Hutchcroft, "Non-uniqueness and mean-field criticality for percolation on nonunimodular transitive graphs", preprint, 2017. arXiv

[Hutchcroft 2019] T. Hutchcroft, "Percolation on hyperbolic graphs", *Geom. Funct. Anal.* **29**:3 (2019), 766–810. MR

[Hutchcroft 2020] T. Hutchcroft, "Locality of the critical probability for transitive graphs of exponential growth", *Ann. Probab.* **48**:3 (2020), 1352–1371. MR

[Kannan and Krueger 1996] R. Kannan and C. K. Krueger, *Advanced analysis on the real line*, Springer, 1996. MR

[Kesten 1987] H. Kesten, "Scaling relations for 2D-percolation", *Comm. Math. Phys.* **109**:1 (1987), 109–156. MR

[Laanait et al. 1991] L. Laanait, A. Messager, S. Miracle-Solé, J. Ruiz, and S. Shlosman, "Interfaces in the Potts model, I: Pirogov–Sinai theory of the Fortuin–Kasteleyn representation", *Comm. Math. Phys.* **140**:1 (1991), 81–91. MR

[Lawler et al. 2002] G. F. Lawler, O. Schramm, and W. Werner, "One-arm exponent for critical 2D percolation", *Electron. J. Probab.* **7** (2002), paper no. 2. MR

[Menshikov 1986] M. V. Menshikov, "Coincidence of critical points in percolation problems", *Dokl. Akad. Nauk SSSR* **288**:6 (1986), 1308–1311. In Russian; translated in *Theory Probab. Appl.* **32**:3, 547–550. MR

[Muirhead and Vanneuville 2020] S. Muirhead and H. Vanneuville, "The sharp phase transition for level set percolation of smooth planar Gaussian fields", *Ann. Inst. Henri Poincaré Probab. Stat.* **56**:2 (2020), 1358–1390. MR

[Newman 1986] C. M. Newman, "Some critical exponent inequalities for percolation", *J. Statist. Phys.* **45**:3-4 (1986), 359–368. MR

[Newman 1987a] C. M. Newman, "Another critical exponent inequality for percolation: $\beta \geq 2/\delta$", pp. 695–699 in *Proceedings of the symposium on statistical mechanics of phase transitions—mathematical and physical aspects* (Trebon, 1986), vol. 47, edited by R. Kotecký, 1987. MR

[Newman 1987b] C. M. Newman, "Inequalities for γ and related critical exponents in short and long range percolation", pp. 229–244 in *Percolation theory and ergodic theory of infinite particle systems* (Minneapolis, 1984–1985), IMA Vol. Math. Appl. **8**, Springer, 1987. MR

[Nguyen 1987] B. G. Nguyen, "Gap exponents for percolation processes with triangle condition", *J. Statist. Phys.* **49**:1-2 (1987), 235–243. MR

[O'Donnell et al. 2005] R. O'Donnell, M. Saks, O. Schramm, and R. A. Servedio, "Every decision tree has an influential variable", pp. 31–39 in *46th Annual IEEE Symposium on Foundations of Computer Science*, 2005.

[Raoufi 2018] A. Raoufi, "The critical Ising model on amenable graphs of exponential growth", *Enseign. Math.* **64**:3-4 (2018), 305–314. MR

[Ray and Spinka 2019] G. Ray and Y. Spinka, "A short proof of the discontinuity of phase transition in the planar random-cluster model with $q > 4$", 2019. arXiv

[Schonmann 2001] R. H. Schonmann, "Multiplicity of phase transitions and mean-field criticality on highly non-amenable graphs", *Comm. Math. Phys.* **219**:2 (2001), 271–322. MR

[Schonmann 2002] R. H. Schonmann, "Mean-field criticality for percolation on planar non-amenable graphs", *Comm. Math. Phys.* **225**:3 (2002), 453–463. MR

[Slade 2006] G. Slade, *The lace expansion and its applications*, edited by J. Picard, Lecture Notes in Mathematics **1879**, Springer, 2006. MR

[Smirnov 2001] S. Smirnov, "Critical percolation in the plane: conformal invariance, Cardy's formula, scaling limits", *C. R. Acad. Sci. Paris Sér. I Math.* **333**:3 (2001), 239–244. MR

[Smirnov and Werner 2001] S. Smirnov and W. Werner, "Critical exponents for two-dimensional percolation", *Math. Res. Lett.* **8**:5-6 (2001), 729–744. MR

[Vanneuville 2019] H. Vanneuville, "Annealed scaling relations for Voronoi percolation", *Electron. J. Probab.* **24** (2019), paper no. 39. MR

Received 24 Oct 2019. Revised 20 Feb 2020. Accepted 24 Feb 2020.

TOM HUTCHCROFT: t.hutchcroft@maths.cam.ac.uk
Statslab, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge, United Kingdom

PROBABILITY and MATHEMATICAL PHYSICS
Vol. 1, No. 1, 2020
https://doi.org/10.2140/pmp.2020.1.167

MEAN-FIELD TRICRITICAL POLYMERS

ROLAND BAUERSCHMIDT AND GORDON SLADE

We provide an introductory account of a tricritical phase diagram, in the setting of a mean-field random walk model of a polymer density transition, and clarify the nature of the density transition in this context. We consider a continuous-time random walk model on the complete graph, in the limit as the number of vertices N in the graph grows to infinity. The walk has a repulsive self-interaction, as well as a competing attractive self-interaction whose strength is controlled by a parameter g. A chemical potential ν controls the walk length. We determine the phase diagram in the (g, ν) plane, as a model of a density transition for a single linear polymer chain. A dilute phase (walk of bounded length) is separated from a dense phase (walk of length of order N) by a phase boundary curve. The phase boundary is divided into two parts, corresponding to first-order and second-order phase transitions, with the division occurring at a tricritical point. The proof uses a supersymmetric representation for the random walk model, followed by a single block-spin renormalisation group step to reduce the problem to a 1-dimensional integral, followed by application of the Laplace method for an integral with a large parameter.

1. The model and results

1.1. ***Introduction.*** Models of critical phenomena such as the Ising model and percolation continue to be of central interest in the probability literature. In such models, a single parameter (temperature for the Ising model or occupation density for percolation) is tuned to a critical value in order to observe universal critical behaviour. In tricritical models, it is instead necessary to tune two parameters simultaneously to observe tricritical behaviour. Despite their importance for physical applications, tricritical phenomena have received much less attention in the mathematical literature than critical phenomena. Our purpose in this paper is to provide an introductory account of a tricritical phase diagram, in the setting of a mean-field random walk model of a polymer density transition, and to clarify the nature of the density transition in this context.

The self-avoiding walk is a starting point for the mathematical modelling of the chemical physics of a single linear polymer chain in a solvent [de Gennes 1979]. The theory of the self-avoiding walk has primarily been developed in the setting of an infinite lattice, often $\mathbb{Z}^d$. So far, this theory has failed to provide theorems capturing the critical behaviour in dimensions $d = 2, 3$, such as a precise description of the typical end-to-end distance, and such problems are rightly considered to be both highly important and notoriously difficult. On $\mathbb{Z}^d$, basic quantities such as the susceptibility — the generating function $\sum_{n=0}^{\infty} c_n z^n$ for the number of n-step self-avoiding walks started from the origin — can be used to model a polymer chain in the *dilute phase*. The susceptibility is undefined when $|z|$ exceeds the reciprocal of

MSC2010: primary 82B27, 82B41; secondary 60K35.
Keywords: polymer model, complete graph, mean field, phase transition, tricritical point, theta point.

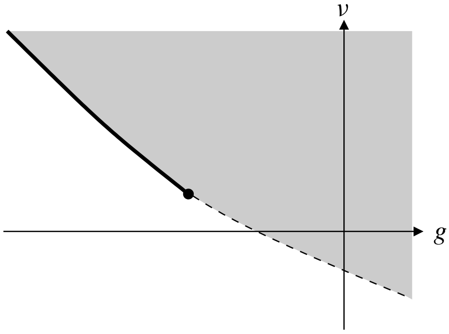


Figure 1. A typical tricritical phase diagram. The second-order curve (dashed line) and the first-order curve (solid line) meet at the tricritical point. The shaded region is the dilute phase (bounded susceptibility) and the unshaded region is the dense phase.

the connective constant $\mu = \lim_{n\to\infty} c_n^{1/n}$. It is however large values of z that are required to model the *dense phase*, as in [Bousquet-Mélou et al. 2005; Duminil-Copin et al. 2014; Madras 1995], and some finite-volume approximation is needed for this. Much remains to be learned about the phase transition from the dilute to the dense phase, including its tricritical nature.

We study a mean-field model based on a continuous-time random walk on the complete graph on N vertices, in the limit $N \to \infty$. The walk has a repulsive self-interaction which models the excluded-volume effect of a linear polymer, as well as a competing attractive self-interaction which models the tendency of the polymer to avoid contact with the solvent. The strength of the self-attraction is controlled by a parameter g, with attraction increasing as g becomes more negative. A chemical potential ν controls the walk length. We investigate the phase diagram in the (g, ν) plane $\mathbb{R}^2$ (positive and negative values), as a model of a density transition for a single linear polymer chain.

In the physics literature, the nature of the phase diagram is well understood. The dilute and dense phases are separated by a phase boundary curve $\nu = \nu_c(g)$ as in Figure 1. The phase boundary itself is divided into two parts: a second-order part for $g > g_c$ across which the average polymer density varies continuously, and a first-order part for $g < g_c$ across which the density has a jump discontinuity. The two pieces of the phase boundary are separated by the tricritical point $(g_c, \nu_c(g_c))$, known as the *theta point*. Tricritical behaviour differs from critical behaviour in the number of parameters that must be tuned. For critical behaviour, an experimentalist needs to tune a single variable to its critical value (given g, tune to $\nu_c(g)$). For tricritical behaviour, two variables must be tuned (tune to $(g_c, \nu_c(g_c))$). A mathematically rigorous theory of the mean-field tricritical polymer density transition has been lacking, and our purpose here is to provide such a theory. Our analysis could be extended to study the tricritical behaviour of n-component spins or higher-order multicritical points. Surprisingly, the mean-field theory of the density transition for the strictly self-avoiding walk has only very recently been developed [Deng et al. 2019; Slade 2020].

The upper critical dimension for the tricritical behaviour is predicted to be $d = 3$, and mean-field tricritical behaviour is predicted for the model on $\mathbb{Z}^d$ in dimensions $d > 3$. On the other hand, for the critical behaviour associated with the second-order part of the phase boundary, the upper critical dimension is instead $d = 4$.

Nonrigorous methods were used in the physics literature to study the density transition in dimensions 2 and 3, and in particular its tricritical behaviour, in the 1980s [Duplantier 1986; 1987; Duplantier and Saleur 1987a; 1987b]. In recent work with Lohmann, we applied a rigorous renormalisation group method to study the 3-dimensional tricritical point [Bauerschmidt et al. 2020], and proved that the tricritical two-point function has Gaussian $|x|^{-1}$ decay for the model on $\mathbb{Z}^3$. In [Golowich and Imbrie 1995], the transition across the second-order phase boundary was studied on a 4-dimensional hierarchical lattice, where a logarithmic correction to the mean-field behaviour of the density was proved. These references make use of an interpretation of the polymer model as the $n = 0$ version of an n-component spin model. We also implement this strategy, using an exact representation of the random walk model based on supersymmetry. After a transformation which can be regarded as a single block-spin renormalisation group step, this representation takes on a form which permits application of the Laplace method for integrals involving a large parameter.

In the mathematical literature, it has been more common to model the polymer collapse transition in terms of the *interacting self-avoiding walk* in which a walk with a self-repulsion receives an energetic reward for nearest-neighbour contacts. A review of the literature on this model can be found in [den Hollander 2009, Chapter 6]; more recent papers include [Bauerschmidt et al. 2017; Hammond and Helmuth 2019; Pétrélis and Torri 2018]. In our mean-field model set on the complete graph, there is no geometry, and the notion of collapse (a highly localised walk) is less meaningful. We therefore concentrate on the density transition and its tricritical behaviour.

1.2. *The model.*

1.2.1. *Definitions.* Let Λ be a finite set with N vertices; ultimately we are interested in the limit $N \to \infty$. Let $X = (X(t))_{t\in[0,\infty)}$ be the continuous-time simple random walk on the complete graph with vertex set Λ. This is the walk with generator Δ defined, for $f : \Lambda \to \mathbb{R}$, by

$$(\Delta f)_x = \frac{1}{N}\sum_{y\in\Lambda}(f_y - f_x) \qquad (x \in \Lambda). \tag{1-1}$$

Equivalently, when the walk is at $x \in \Lambda$, it steps to a uniformly chosen vertex in $\Lambda \setminus \{x\}$ after an exponentially distributed holding time with rate $1 - \frac{1}{N}$. The steps and holding times are all independent. We denote expectation for X with initial point $X(0) = x$ by E_x.

The *local time* of X at x up to time T is the random variable

$$L_{T,x} = \int_0^T \mathbb{1}_{X(t)=x}\,dt \qquad (x \in \Lambda), \tag{1-2}$$

which measures the amount of time spent by the walk at x up to time T. Let L_T denote the vector of all local times. Given a function $p : [0,\infty) \to [0,\infty)$ with $p(0) = 1$, we write $p_N(L_T) = \prod_{x\in\Lambda} p(L_{T,x})$. Let $x, y \in \Lambda$. Assuming the integrals exist, the *two-point function* is

$$G_{xy} = \int_0^\infty E_x(p_N(L_T)\mathbb{1}_{X(T)=y})\,dT, \tag{1-3}$$

and the *susceptibility* is

$$\chi = \sum_{y\in\Lambda} G_{xy} = \int_0^\infty E_x(p_N(L_T))\,dT. \tag{1-4}$$

The right-hand side is independent of $x \in \Lambda$.

We define the random variable L, the *length* of X, by its probability density function

$$f_L(T) = \frac{1}{\chi} E_x(p_N(L_T)) \qquad (T \geq 0), \tag{1-5}$$

which is also independent of $x \in \Lambda$. The expected value of the length is

$$\mathbb{E}L = \frac{1}{\chi} \int_0^\infty T\,E_x(p_N(L_T))\,dT. \tag{1-6}$$

The expected length can be written more compactly using a dot to represent differentiation with respect to ϵ at $\epsilon = 0$, when $p(s)$ is replaced by $p(s)e^{-\epsilon s}$. With this notation, since $T = \sum_{x\in\Lambda} L_{T,x}$,

$$\mathbb{E}L = -\frac{1}{\chi}\dot{\chi}. \tag{1-7}$$

Assuming the limit exists, the *density* of the walk is defined by

$$\rho = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}L. \tag{1-8}$$

1.2.2. *Example.* Although our results will be presented more generally, we are motivated by the example

$$p(t) = e^{-t^3-gt^2-\nu t} \qquad (t \geq 0), \tag{1-9}$$

where $g, \nu \in \mathbb{R}$ (we have set the coefficient of t^3 to equal 1; its specific value is unimportant). For p defined by (1-9), the two-point function becomes

$$G_{xy}(g,\nu) = \int_0^\infty E_x(e^{-\sum_{x\in\Lambda}(L_{T,x}^3+gL_{t,x}^2)}\mathbb{1}_{X(T)=y})e^{-\nu T}\,dT. \tag{1-10}$$

The above integral is finite for all $g, \nu \in \mathbb{R}$, since by Hölder's inequality

$$T = \sum_{x\in\Lambda} L_{T,x} \leq \Big(\sum_{x\in\Lambda} L_{T,x}^3\Big)^{1/3} |\Lambda|^{2/3}, \tag{1-11}$$

and also

$$\sum_{x\in\Lambda} L_{T,x}^2 \leq (\sup_x L_{T,x}) \sum_x L_{T,x} \leq T^2, \tag{1-12}$$

so

$$G_{xy}(g,\nu) \leq \int_0^\infty e^{-T^3|\Lambda|^{-2}+|g|T^2+|\nu|T}\,dt < \infty. \tag{1-13}$$

By definition,

$$\sum_{x\in\Lambda} L_{T,x}^2 = \int_0^T \int_0^T \mathbb{1}_{X(s)=X(t)}\, ds\, dt, \tag{1-14}$$

$$\sum_{x\in\Lambda} L_{T,x}^3 = \int_0^T \int_0^T \int_0^T \mathbb{1}_{X(s)=X(t)=X(u)}\, ds\, dt\, du. \tag{1-15}$$

Our interest lies in the case $g < 0$. In this case, walks X for which the local time has large ℓ^3-norm are penalised by the factor $e^{-\sum_{x\in\Lambda} L_{T,x}^3}$ (three-body repulsion), whereas those with large ℓ^2-norm are rewarded by the factor $e^{+\sum_{x\in\Lambda}|g|L_{T,x}^2}$ (two-body attraction). This is a model of a linear polymer in a solvent. The parameter ν is a chemical potential which controls the length of the polymer. The three-body repulsion models the excluded volume effect, and the two-body attraction models the effect of temperature or solvent quality. The competition between attraction and repulsion, together with the variable length mediated by the chemical potential, leads to a rich phase diagram.

1.2.3. *Effective potential.* The mean-field Ising model, known as the Curie–Weiss model, can be analysed in terms of the effective potential $V_{\text{Ising}}(\varphi) = \frac{\beta}{2}\varphi^2 - \log\cosh(\beta\varphi)$. In [Bauerschmidt et al. 2019, Section 1.4], this effective potential was derived as the result of a single block-spin renormalisation group step. Our approach is based on this idea.

For the mean-field polymer model with interaction p, we define the *effective potential* $V : [0, \infty) \to \mathbb{R}$ by

$$V(t) = t - \log(1 + v(t)), \qquad v(t) = \int_0^\infty p(s)e^{-s}\sqrt{\frac{t}{s}}\, I_1(2\sqrt{st})\, ds, \tag{1-16}$$

with I_1 the modified Bessel function of the first kind. We show in Proposition 1.1 that the integral $v(t)$ is finite if p is integrable. By definition, $V(0) = 0$.

The variable t corresponds to $\frac{1}{2}\varphi^2$ for the Ising effective potential. It is common in tricritical theory to encounter a triple-well potential; a double well for $V(t)$ in our setting corresponds to a triple well as a function of φ.

The effective potential occurs in integral representations of the two-point function, the susceptibility, and the expected length. In contrast to the analysis of the mean-field Ising model in [Bauerschmidt et al. 2019, Section 1.4], the integral representations involve the notions of fermions and supersymmetry as presented in [Bauerschmidt et al. 2019, Chapter 11]. Nevertheless, the integral representation reduces to a 1-dimensional Lebesgue integral. For example, as noted below (3-17), the two-point function at distinct points labelled 0, 1 has the integral representation

$$G_{01} = \int_0^\infty e^{-NV(t)}(NV'(t)(1 - V'(t)) + 2V''(t))(1 - V'(t))t\, dt. \tag{1-17}$$

Similarly G_{00}, χ, and $\mathbb{E}L$ are represented by integrals of the form $\int_0^\infty e^{-NV(t)}K(t)\, dt$ for suitable kernels K. The asymptotic behaviour of such integrals, as $N \to \infty$, can be computed using the Laplace method. This requires knowledge of the minimum structure of the effective potential V. The use of the minimum structure to predict the phase diagram is referred to as the Landau theory (see, e.g., [Arovas 2019, Section 7.6.4] where our variable t corresponds to m^2).

We assume throughout the paper that $p : [0, \infty) \to [0, \infty)$ is such that $e^{\epsilon s} p(s)$ is a Schwartz function for some $\epsilon > 0$; this assumption is a convenience that permits direct application of the integral representation given in Theorem 2.2. In particular, p is integrable. We also assume that $p(0) = 1$.

The following elementary proposition collects some basic facts about the effective potential. Weaker assumptions on p and a stronger analyticity conclusion for V are possible, but the proposition is sufficient for our needs as it is stated. The proposition shows that V is analytic on $(0, \infty)$ under our assumption that p is integrable. It also shows that the derivatives of the effective potential at $t = 0$ can be expressed in terms of the moments of $p(s)e^{-s}$ defined by

$$M_k = \int_0^\infty p(s)e^{-s} s^k ds \qquad (k = 0, 1, \ldots). \tag{1-18}$$

Such derivatives appear in Definition 1.2 and Theorems 1.3–1.4 below.

Proposition 1.1. (i) *If $\int_0^\infty p(s)\, ds < \infty$ then V is well-defined and analytic in $t \in (0, \infty)$. Moreover, if $p(s) \le O(e^{-\epsilon s})$ for some $\epsilon > 0$ then there exists $\delta > 0$ such that $V(t) \ge \delta t + O(1)$ as $t \to \infty$.*

(ii) *Derivatives of V at $t = 0$ (with the dot notation as indicated above (1-7)) are given by*

$$V'(0) = 1 - M_0, \quad V''(0) = M_0^2 - M_1, \quad V'''(0) = -\tfrac{1}{2}M_2 + 3M_1M_0 - 2M_0^3, \quad \dot{V}(0) = 0, \quad \dot{V}'(0) = M_1. \tag{1-19}$$

Proof. (i) The modified Bessel function I_1 is the entire function

$$I_1(z) = \sum_{k=0}^\infty \frac{1}{k!(k+1)!}\left(\frac{z}{2}\right)^{2k+1}. \tag{1-20}$$

Its asymptotic behaviour is $I_1(z) \sim \frac{z}{2}$ as $z \downarrow 0$ and $I_1(z) \sim 1/\sqrt{2\pi z}\, e^z$ as $z \to \infty$. Thus the integral $v(t)$ converges when p is integrable, and $V(t)$ is well-defined.

To prove that V is analytic on $(0, \infty)$, since $t \mapsto \sqrt{t}$ is analytic, it suffices to prove that the function $w(z) = v(z^2)$ is entire. By definition, for $z \in \mathbb{C}$,

$$w(z) = z \int_0^\infty p(s)e^{-s} s^{-1/2} I_1(2s^{1/2}z)\, ds. \tag{1-21}$$

The modified Bessel functions obey

$$|I_n(z)| \le I_n(|z|) \sim \frac{1}{\sqrt{2\pi|z|}} e^{|z|} \tag{1-22}$$

as $|z| \to \infty$ and the same asymptotics hold for their derivatives. This permits differentiation under the integral and guarantees existence of the derivative $w'(z)$.

For the lower bound on V, we apply the assumption $p(s) \le O(e^{-\epsilon s})$, the bounds $I_1(z) \le O(e^z)$ for $z \ge 1$ and $I_1(z) \le O(z)$ for $z \le 1$, and the inequality $2\sqrt{st} \le (1+\epsilon/2)s + (1-\delta)t$ with $1-\delta = (1+\epsilon/2)^{-1}$. Together, these lead to

$$v(t) \le C + C\int_1^\infty e^{-s-\epsilon s} e^{2\sqrt{st}}\, ds \le Ce^{(1-\delta)t} \int_0^\infty e^{-\epsilon s/2}\, ds \le O(e^{(1-\delta)t}), \tag{1-23}$$

and hence $V(t) = t - \log(1 + v(t)) \ge \delta t + O(1)$.

(ii) The Taylor expansion of $v(t)$ at $t=0$ is

$$v(t)=\int_0^\infty p(s)e^{-s}\left(\sum_{k=0}^\infty \frac{1}{k!(k+1)!}t^{k+1}s^k\right)ds=\sum_{k=0}^\infty \frac{1}{k!(k+1)!}M_k t^{k+1}. \tag{1-24}$$

In particular,

$$v(0)=0,\quad v'(0)=M_0,\quad v''(0)=M_1,\quad v^{(k)}(0)=\frac{M_{k-1}}{(k-1)!}\qquad (k\geq 1). \tag{1-25}$$

Computation gives

$$V'(t)=1-\frac{v'(t)}{1+v(t)},\qquad V''(t)=-\frac{v''(t)}{1+v(t)}+\left(\frac{v'(t)}{1+v(t)}\right)^2, \tag{1-26}$$

and the third derivative can be computed similarly. This leads to the statements for the derivatives of V with respect to t.

Finally, $\dot V(0)=0$ since $V(0)=0$ holds also when $p(s)$ is replaced by $p(s)e^{-\epsilon s}$, and $\dot V'(0)=M_1$ follows from $V'(0)=1-M_0$ and $\dot M_0=-M_1$. □

1.3. ***General results.*** In the following definition, we have in mind the situation where the effective potential V is defined by a function p which is parametrised by two real parameters (g,ν) as in (1-9). Different choices of parameters can correspond to different cases in the definition. For the specific example of (1-9), plots of the phase diagram and effective potential are given in Figures 2–3. However, our results and their proofs depend only on the qualitative features of the effective potential listed in the definition.

We say that the effective potential V *has a unique global minimum* $V(t_0)$ if $V(t)>V(t_0)$ for all $t\neq t_0$, and $\inf_{t\in[0,\infty):|t-t_0|\geq\epsilon}(V(t)-V(t_0))>0$ for all $\epsilon>0$. We say that V *has global minima* $V(t_0)=V(t_1)$ with $t_0\neq t_1$ if $V(t)>V(t_0)=V(t_1)$ for all $t\neq t_0,t_1$, and $\inf_{t\in[0,\infty):|t-t_0|\geq\epsilon,\,|t-t_1|\geq\epsilon}(V(t)-V(t_0))>0$ for all $\epsilon>0$.

Definition 1.2. We define two phases, two phase boundaries, and the tricritical point, in terms of the effective potential V as follows:

dilute phase: $V'(0)>0$, unique global minimum $V(0)=0$.

second-order curve: $V'(0)=0$, $V''(0)>0$, unique global minimum $V(0)=0$.

tricritical point: $V'(0)=V''(0)=0$, $V'''(0)>0$, unique global minimum $V(0)=0$.

first-order curve: $V'(0)>0$, global minima $V(0)=V(t_0)=0$ with $t_0>0$, $V''(t_0)>0$.

dense phase: unique global minimum $V(t_0)<0$ with $t_0>0$, $V''(t_0)>0$.

In principle, there are further possibilities such as n-th-order critical points. These do not occur in our example (1-9), so we do not consider them, but they could be handled in an analogous way.

The following two theorems give the asymptotic behaviour of the two-point function, the susceptibility, and the expected length, in the different regions of the phase diagram. The result for the susceptibility is a consequence of the result for the two-point function, together with the identity $\chi=G_{00}+(N-1)G_{01}$. As usual, the Gamma function is $\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt$ for $x>0$. The notation $f(N)\sim g(N)$ means $\lim_{N\to\infty}f(N)/g(N)=1$.

Theorem 1.3. *The two-point function has the following asymptotic behaviour*:

$$G_{00} \sim \begin{cases} 1 - V'(0) & (\textit{dilute phase and first-order curve}), \\ 1 & (\textit{second-order curve}), \\ 1 & (\textit{tricritical point}), \\ e^{N|V(t_0)|} \dfrac{1}{N^{1/2}} \dfrac{\sqrt{2\pi}}{V''(t_0)^{1/2}} (1 - V''(t_0)) & (\textit{dense phase}), \end{cases} \tag{1-27}$$

$$G_{01} \sim \begin{cases} \dfrac{(1-V'(0))^2}{V'(0)N} & (\textit{dilute phase}), \\ \dfrac{1}{\left(\frac{1}{2!}V''(0)N\right)^{1/2}} \Gamma(3/2) & (\textit{second-order curve}), \\ \dfrac{1}{\left(\frac{1}{3!}V'''(0)N\right)^{1/3}} \Gamma(4/3) & (\textit{tricritical point}), \\ e^{N|V(t_0)|} \dfrac{1}{N^{1/2}} \dfrac{\sqrt{2\pi}}{V''(t_0)^{1/2}} & (\textit{dense phase and first-order curve}). \end{cases} \tag{1-28}$$

The statement of the next theorem uses the notation $\dot{V}(t_0)$ and $\dot{V}'(0)$. The dot notation is as discussed above (1-7). Explicitly,

$$\dot{V}(t) = -\frac{\dot{v}(t)}{1+v(t)}, \qquad \dot{v}(t) = -\int_0^\infty p(s)e^{-s}\sqrt{st}\, I_1(2\sqrt{st})\, ds, \tag{1-29}$$

and, as usual, a prime denotes differentiation with respect to t.

Theorem 1.4. *The susceptibility χ and expected length $\mathbb{E}L$ have the following asymptotic behaviour*:

$$\chi \sim \begin{cases} \dfrac{1-V'(0)}{V'(0)} & (\textit{dilute phase}), \\ N^{1/2} \dfrac{\Gamma(3/2)}{\left(\frac{1}{2!}V''(0)\right)^{1/2}} & (\textit{second-order curve}), \\ N^{2/3} \dfrac{\Gamma(4/3)}{\left(\frac{1}{3!}V'''(0)\right)^{1/3}} & (\textit{tricritical point}), \\ e^{N|V(t_0)|} N^{1/2} \dfrac{\sqrt{2\pi}}{V''(t_0)^{1/2}} & (\textit{dense phase and first-order curve}), \end{cases} \tag{1-30}$$

$$\mathbb{E}L \sim \begin{cases} \dfrac{\dot{V}'(0)}{V'(0)(1-V'(0))} & (\textit{dilute phase}), \\ N^{1/2} \dfrac{1}{\Gamma(1/2)} \dfrac{\dot{V}'(0)}{\left(\frac{1}{2!}V''(0)\right)^{1/2}} & (\textit{second-order curve}), \\ N^{2/3} \dfrac{\Gamma(2/3)}{\Gamma(1/3)} \dfrac{1}{\left(\frac{1}{3!}V'''(0)\right)^{1/3}} & (\textit{tricritical point}), \\ N\dot{V}(t_0) & (\textit{dense phase and first-order curve}). \end{cases} \tag{1-31}$$

By Theorem 1.3, the two-point function remains bounded in the dilute phase, on the first- and second-order curves, and at the tricritical point. Also, G_{00} is asymptotically constant in the dilute phase, on the second-order curve, and at the tricritical point, whereas G_{01} decays at different rates in the different regions. In the dense phase, both G_{00} and G_{01} grow exponentially in N. Proposition 1.1(ii) shows that in all cases $1 - V'(0) > 0$, as is implied in particular in the dilute phase by the first asymptotic formula for G_{00}. The formula for G_{00} in the dense phase implies that $V''(t_0) < 1$; we do not have an independent general proof of that (though if t_0 is smooth in (g, ν) then it is true in the vicinity of the tricritical point where $t_0 = 0$ and $V''(0) = 0$).

Theorem 1.4 indicates that the susceptibility χ and expected length $\mathbb{E}L$ each have finite infinite-volume limits in the dilute phase. In the dense phase, χ grows exponentially with N. In the dense phase and on the first-order curve, $\mathbb{E}L$ is asymptotically linear in N. On the second-order curve, χ and $\mathbb{E}L$ are each of order $N^{1/2}$ (as in [Deng et al. 2019; Slade 2020] for the self-avoiding walk on the complete graph), whereas each is of order $N^{2/3}$ at the tricritical point. The density $\rho = \lim_{N\to\infty} N^{-1}\mathbb{E}L$ is zero except on the first-order curve and in the dense phase, where it is equal to $\dot{V}(t_0) > 0$.

The proofs of Theorems 1.3–1.4 are given in two steps. In Section 2, integral representations based on supersymmetry are derived; these involve the effective potential. In Section 3, the Laplace method is used to evaluate the asymptotic behaviour of the integrals.

1.4. ***Phase diagram for the example.*** For further interpretation of the phase diagram, we restrict attention in this section to the particular example

$$p(t) = e^{-t^3 - gt^2 - \nu t}, \tag{1-32}$$

for which we carry out numerical calculations to determine the structure of the effective potential. The numerical input we need is collected in Section 4.1, and we mention some of it here.

Two curves which provide bearings in the (g, ν) plane are determined by the equations $V'(0) = 0$ (i.e., $M_0 = 1$) and $V''(0) = 0$ (i.e., $M_1 = M_0^2$). The curves, which are plotted in Figure 2, intersect at the tricritical point (i.e., $M_0 = M_1 = 1$), which is

$$g_c = -3.2103..., \qquad \nu_c = 2.0772.... \tag{1-33}$$

At the tricritical point, numerical integration gives $M_2 = 1.4478...$ and $V'''(0) = 0.2762... > 0$. The first-order curve is the solid curve in Figure 2. The second-order curve is the portion of the dashed curve below the tricritical point. The dilute phase lies above the first- and second-order curves, and the dense phase comprises the other side of those curves. The *phase boundary* is the union of the first- and second-order curves together with the tricritical point; we regard this curve as a function $\nu_c(g)$ parametrised by g.

By Theorem 1.4, there is a transition as the phase boundary $g \mapsto (g, \nu_c(g))$ is traversed in the direction of decreasing g:

- On the second-order curve, χ and $\mathbb{E}L$ are of order $N^{1/2}$ and $\rho = 0$.
- At the tricritical point, χ and $\mathbb{E}L$ are of order $N^{2/3}$ and $\rho = 0$.
- On the first-order curve, χ is of order $N^{1/2}$, $\mathbb{E}L$ is of order N, and $\rho = \dot{V}(t_0) > 0$.

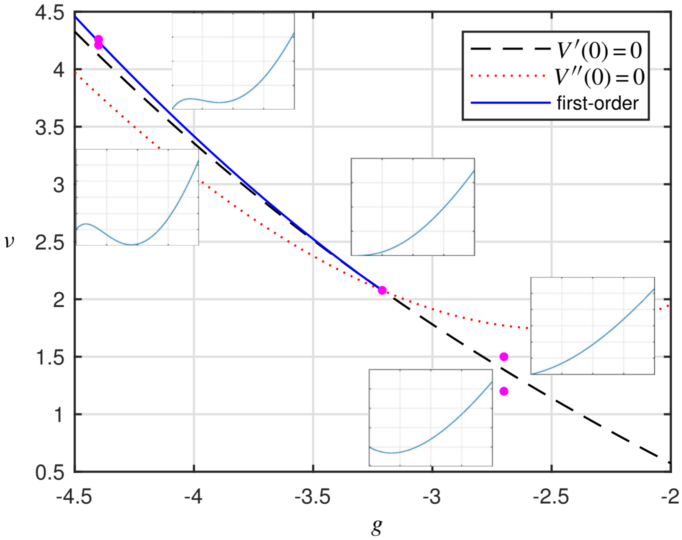


Figure 2. Phase diagram for $p(t) = e^{-t^3-gt^2-\nu t}$. The five marked points with their effective potentials are those shown in Figure 3.

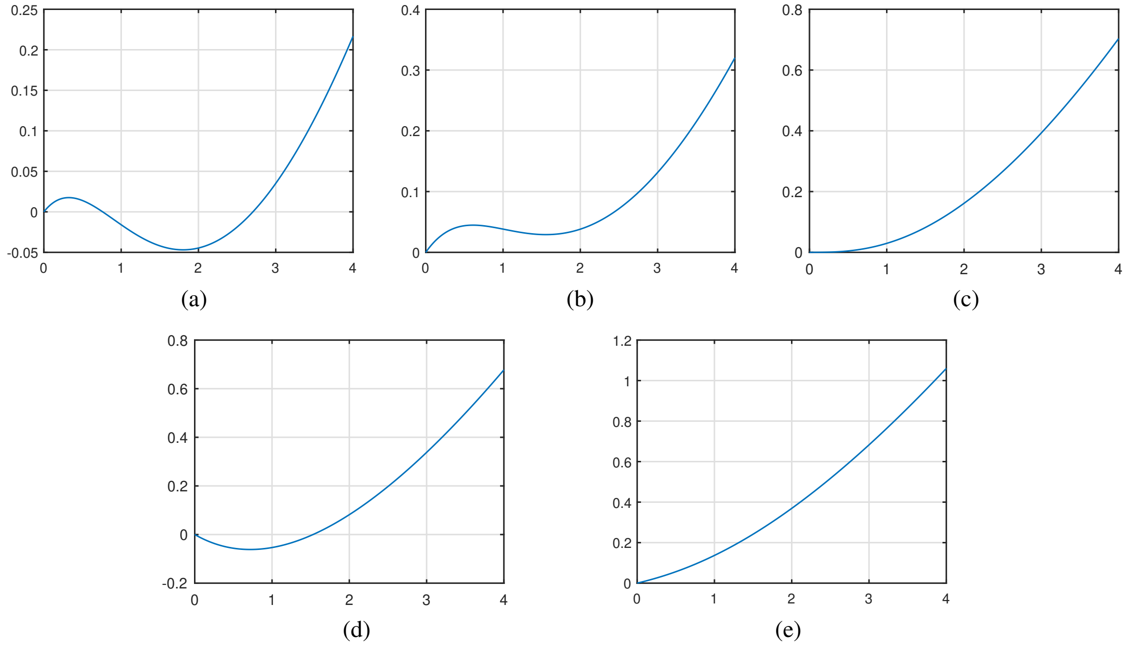
This is a density transition, from zero to positive density. Note that, by definition, $\dot{V} = dV/d\nu$ when p is given by (1-32).

First-order curve. The density ρ is discontinuous when crossing the first-order curve, since its value on the first-order curve is $\dot{V}(t_0) > 0$ whereas its value in the dilute phase is zero. However, the density is continuous at the first-order curve for the one-sided approach from the dense phase. This can be understood from the behaviour of the effective potential: as the first-order curve is approached from the dense phase, t_0 remains bounded away from zero and $\dot{V}(t_0)$ does not vanish (Figure 3 (a) and (b)). The density discontinuity on the first-order curve is in contrast to the continuous behaviour on the second-order curve. As the second-order curve (or tricritical point) is approached from the dense phase, t_0 decreases continuously to zero and $\dot{V}(t_0) \downarrow 0$ (Figure 3 (d) and (e)).

In the limit $N \to \infty$, the susceptibility has finite limit $(1 - V'(0))/V'(0)$ as the first-order curve is approached from the dilute phase, whereas it is divergent on the first-order curve. This is typical of a first-order transition.

Second-order curve and tricritical point. The detailed asymptotic behaviour of the divergence of the susceptibility and the vanishing of the density, at the second-order curve and tricritical point, are as described in the following theorem. The theorem also describes the phase boundary at the tricritical point. Its proof is given in Section 4.

Theorem 1.5 relies on numerical analysis of the effective potential for p given by (1-32), as discussed above. The precise conclusions from this numerical analysis are stated in Section 4.1. We emphasise that the effective potential is a function of a single real variable, and thus we believe that with effort this numerical input could be replaced by rigorous analysis (perhaps with computer assistance), but we do not pursue this.



(a) (b) (c)

(d) (e)

Figure 3. Effective potential V vs. t for several values of (g, ν): (a) dense phase near first-order curve $(-4.4, 4.21)$, (b) dilute phase near first-order curve $(-4.4, 4.26)$, (c) tricritical point, (d) dense phase near second-order curve $(-2.7, 1.2)$, (e) dilute phase near second-order curve $(-2.7, 1.5)$

In the theorem, we consider a line segment

$$(g(s), \nu(s)) = (g(0) + sm_1, \nu(0) + sm_2) \qquad (s \in [0, 1]) \tag{1-34}$$

that approaches a base point $(g(0), \nu(0))$ as $s \downarrow 0$. We write $\boldsymbol{m} = (m_1, m_2)$ for its direction. The base point may be either the tricritical point or a point on the second-order curve.

Theorem 1.5. *Let p be given by (1-32).*

(i) *The phase boundary $\nu_c(g)$ is differentiable with respect to g at the tricritical point, with slope $-M_2$. However, its left and right second derivatives differ at the tricritical point.*

(ii) *The vector $\boldsymbol{n} = (M_2, M_1)$ is normal along the second-order curve. Along a line segment (1-34) approaching a point on the second-order curve, or the tricritical point, from the dilute phase with direction satisfying $\boldsymbol{m} \cdot \boldsymbol{n} \neq 0$ (nontangential at the given point), the infinite-volume susceptibility $\chi = (1 - V'(0))/V'(0)$ diverges as*

$$\chi = \frac{1 - V'(0)}{V'(0)} \sim \frac{1}{|\boldsymbol{m} \cdot \boldsymbol{n}|s}. \tag{1-35}$$

(iii) *Along a line segment (1-34) approaching a point on the second-order curve from the dense phase with direction satisfying $\boldsymbol{m} \cdot \boldsymbol{n} \neq 0$ (nontangential at the given point), the density ρ vanishes as*

$$\rho \sim \frac{M_1}{1 - M_1} |\boldsymbol{m} \cdot \boldsymbol{n}| s \qquad (\textit{here } 1 - M_1 > 0). \tag{1-36}$$

There exists an arc of the second-order curve adjacent to the tricritical point, such that under tangential approach to a point on that arc, $\rho \sim Bs^2$ *with* $B > 0$.

There are positive constants B_0, B_1, B_2, B_3 *such that as the tricritical point is approached,*

$$\rho \sim \begin{cases} B_0(|\boldsymbol{m}\cdot\boldsymbol{n}|s)^{1/2} & (\textit{nontangentially from dense phase}), \\ B_1 s & (\textit{tangentially from second-order side}), \\ B_2 s & (\textit{tangentially from first-order side}), \\ B_3 s & (\textit{along first-order curve}). \end{cases} \tag{1-37}$$

(*For the first-order curve, the parametrisation is* $(g(s), \nu(s)) = (g_c - s, \nu_c(g-s))$.)

It is possible in general that the susceptibility could have different asymptotic behaviour for the approaches to the second-order curve and the tricritical point, but for the mean-field model there is no difference. However, for the density there is a difference.

2. Integral representation

In this section, we prove integral representations for the two-point function and expected length, in Propositions 2.7–2.8, via the supersymmetric version of the BFS–Dynkin isomorphism theorem [Bauerschmidt et al. 2019, Corollary 11.3.7]. These integral representations are in terms of the effective potential and provide the basis for the proofs of Theorems 1.3–1.4. We begin with brief background concerning Grassmann integration. Further background and history for the isomorphism theorem can be found in [Bauerschmidt et al. 2019, Chapter 11].

2.1. *Grassmann algebra and the integral representation.*

2.1.1. *Grassmann algebra.* We define a Grassmann algebra $\mathcal{N}_1$ with two generators $\psi, \bar\psi$ (the bar is only notational and is not a complex conjugate) to consist of linear combinations

$$K = a_0 + a_1\bar\psi + a_2\psi + a_3\bar\psi\psi, \tag{2-1}$$

where each a_i is a smooth function $a_i : \mathbb{R}^2 \to \mathbb{R}$ written $(u, v) \mapsto a_i(u, v)$, and where multiplication of the generators is anticommutative, i.e.,

$$\bar\psi\psi = -\psi\bar\psi, \qquad \psi\psi = 0, \qquad \bar\psi\bar\psi = 0. \tag{2-2}$$

To make the notation more symmetric, we combine $(u, v) \in \mathbb{R}^2$ into a complex variable ϕ by

$$\phi = u + iv, \qquad \bar\phi = u - iv. \tag{2-3}$$

We call $(\phi, \bar\phi)$ a *bosonic variable*, $(\psi, \bar\psi)$ a *fermionic variable*, and $\Phi = (\phi, \bar\phi, \psi, \bar\psi)$ a *supervariable*. Elements of the Grassmann algebra $\mathcal{N}_1$ are called *forms*. A form with $a_1 = a_2 = 0$ is called *even*. An important even form is

$$\Phi^2 = \phi\bar\phi + \psi\bar\psi. \tag{2-4}$$

The above discussion concerns a single boson pair and a single fermion pair. We also have need of the Grassmann algebra $\mathcal{N}_N$ with $2N$ anticommuting generators $(\psi_x, \bar{\psi}_x)_{x\in\Lambda}$, now with coefficients which are smooth functions from $\mathbb{R}^{2N}$ to $\mathbb{R}$. The *even subalgebra* consists of elements of $\mathcal{N}_N$ which only involve terms containing products of an even number of generators. We refer to $(\phi_x, \bar{\phi}_x)_{x\in\Lambda}$ and $(\psi_x, \bar{\psi}_x)_{x\in\Lambda}$ as the boson field and the fermion field, respectively. The combination $\Phi = (\phi_x, \bar{\phi}_x, \psi_x, \bar{\psi}_x)_{x\in\Lambda}$ is called a *superfield*, and we write

$$\Phi^2 = (\Phi_x^2)_{x\in\Lambda} = (\phi_x\bar{\phi}_x + \psi_x\bar{\psi}_x)_{x\in\Lambda}. \tag{2-5}$$

Two useful even forms in $\mathcal{N}_N$ are

$$(\Phi, \Phi) = \sum_{x\in\Lambda} \Phi_x^2 = \sum_{x\in\Lambda} (\phi_x\bar{\phi}_x + \psi_x\bar{\psi}_x), \tag{2-6}$$

$$(\Phi, -\Delta\Phi) = \sum_{x\in\Lambda} (\phi_x(-\Delta\bar{\phi})_x + \psi_x(-\Delta\bar{\psi})_x), \tag{2-7}$$

where $-\Delta$ is still defined by (1-1) when applied to the generators ψ and $\bar{\psi}$.

For $p \in \mathbb{N}$, consider a C^∞ function $F : \mathbb{R}^p \to \mathbb{R}$. Let $K = (K_j)_{j\le p}$ be a collection of even forms, and assume that the degree-zero part K_j^0 of each K_j (obtained by setting all fermionic variables to zero) is real. We define a form denoted $F(K)$ by Taylor series about the degree-zero part of K, i.e.,

$$F(K) = \sum_\alpha \frac{1}{\alpha!} F^{(\alpha)}(K^0)(K - K^0)^\alpha. \tag{2-8}$$

Here $\alpha = (\alpha_j)_{j\le p}$ is a multi-index, with $\alpha! = \prod_{j=1}^p \alpha_j!$ and $(K - K^0)^\alpha = \prod_{j=1}^p (K_j - K_j^0)^{\alpha_j}$. The order of the product is immaterial since each $K_j - K_j^0$ is even by assumption. Also, the summation terminates after finitely many terms since each $K_j - K_j^0$ is nilpotent.

For example, for $\Phi^2 \in \mathcal{N}_1$ given by (2-4), for smooth $F : \mathbb{R} \to \mathbb{R}$, the previous definition with $p = 1$ gives

$$F(\Phi^2) = F(\phi\bar{\phi}) + F'(\phi\bar{\phi})\psi\bar{\psi} = F(\phi\bar{\phi}) - F'(\phi\bar{\phi})\bar{\psi}\psi. \tag{2-9}$$

2.1.2. *Grassmann integration and the integral representation.* Given a form $K \in \mathcal{N}_N$, we write K_{2N} for its coefficient of $\bar{\psi}_1\psi_1 \cdots \bar{\psi}_N\psi_N$. This K_{2N} is a function of (u, v), i.e., a function on $\mathbb{R}^{2N}$. For example, for the form $K = F(\Phi^2)$ of (2-9), we have $N = 1$ and $K_2(u, v) = -F'(\phi\bar{\phi}) = -F'(u^2 + v^2)$. In general, the *superintegral* of K is defined by

$$\int_{\mathbb{R}^{2N}} D\Phi\, K = \frac{1}{\pi^N} \int_{\mathbb{R}^{2N}} K_{2N}(u, v)\, du\, dv, \tag{2-10}$$

assuming that K_{2N} decays sufficiently rapidly that the Lebesgue integral on the right-hand side exists. The notation $D\Phi$ signifies that K is a form for the superfield $\Phi = (\phi_x, \bar{\phi}_x, \psi_x, \bar{\psi}_x)_{x\in\Lambda}$. This will be useful to distinguish superfields when more than one are in play. The factor π^{-N} in the definition simplifies the conclusions of the next example and theorem.

Example 2.1. If $F:\mathbb{R}\to\mathbb{R}$ decays sufficiently rapidly then

$$\int_{\mathbb{R}^2} D\Phi\, F(\Phi^2) = F(0). \tag{2-11}$$

In fact, after conversion to polar coordinates, using $\frac{1}{\pi}\,du\,dv = dr^2\,\frac{1}{2\pi}d\theta$, the definition gives

$$\int_{\mathbb{R}^2} D\Phi\, F(\Phi^2) = -\int_{\mathbb{R}^2} F'(r^2)\frac{1}{\pi}\,du\,dv = -\int_0^\infty F'(t)dt = F(0), \tag{2-12}$$

as claimed.

The supersymmetric version of the BFS–Dynkin isomorphism theorem (see, for example, [Bauerschmidt et al. 2019, Corollary 11.3.7]) relates random walks and superfields via an exact equality, as follows.

Theorem 2.2. *Let $F:\mathbb{R}^N\to\mathbb{R}$ be such that $e^{\epsilon\sum_{z\in\Lambda}t_z}F(t)$ is a Schwartz function for some $\epsilon>0$. Then*

$$\int_0^\infty E_x\big(F(L_T)\mathbb{1}_{X(T)=y}\big)\,dT = \int_{\mathbb{R}^{2N}} D\Phi\, e^{-(\Phi,-\Delta\Phi)}F(\Phi^2)\,\bar{\phi}_x\phi_y, \tag{2-13}$$

where Δ is the generator (defined in (1-1)) of the random walk $X=(X(t))_{t\geq 0}$ with expectation E_x, and L_T is the local time.

By definition, the two-point function (1-3) and expected length (1-6) are given by expressions like the left-hand side of (2-13), which therefore can be rewritten as the right-hand side.

2.1.3. *Block-spin renormalisation.* The next lemma gives a way to rewrite the exponential factor on the right-hand side of (2-13) as an integral over a single *constant* block-spin superfield $\mathrm{Z}=(\zeta,\bar{\zeta},\xi,\bar{\xi})$. The application of this lemma can be regarded as a single block-spin renormalisation group step, as in [Bauerschmidt et al. 2019, Section 1.4]. For the statement of the lemma, we use the notation

$$(\mathrm{Z}-\Phi,\mathrm{Z}-\Phi) = \sum_{x\in\Lambda}(\mathrm{Z}-\Phi_x)^2 = \sum_{x\in\Lambda}\big((\zeta-\phi_x)(\bar{\zeta}-\bar{\phi}_x)+(\xi-\psi_x)(\bar{\xi}-\bar{\psi}_x)\big). \tag{2-14}$$

Lemma 2.3. *For a superfield $\Phi=(\phi,\bar{\phi},\psi,\bar{\psi})=(\phi_x,\bar{\phi}_x,\psi_x,\bar{\psi}_x)_{x\in\Lambda}$,*

$$e^{-(\Phi,-\Delta\Phi)} = \int_{\mathbb{R}^2} D\mathrm{Z}\, e^{-(\mathrm{Z}-\Phi,\mathrm{Z}-\Phi)}. \tag{2-15}$$

Proof. Let $A\phi = N^{-1}\sum_{x\in\Lambda}\phi_x$. Since cross terms vanish,

$$\begin{aligned}\sum_{x\in\Lambda}(\zeta-\phi_x)(\bar{\zeta}-\bar{\phi}_x) &= \sum_{x\in\Lambda}((\zeta-A\phi)+(A\phi-\phi_x))((\bar{\zeta}-A\bar{\phi})+(A\bar{\phi}-\bar{\phi}_x))\\ &= N(\zeta-A\phi)(\bar{\zeta}-A\bar{\phi})+\sum_{x\in\Lambda}(A\phi-\phi_x)(A\bar{\phi}-\bar{\phi}_x).\end{aligned} \tag{2-16}$$

By definition of Δ, and since $\sum_{x\in\Lambda}(\Delta f)_x=0$, the last term on the right-hand side is

$$\sum_{x\in\Lambda}(A\phi-\phi_x)(A\bar{\phi}-\bar{\phi}_x) = \sum_{x\in\Lambda}(A\phi-\phi_x)(\Delta\bar{\phi})_x = -\sum_{x\in\Lambda}\phi_x(\Delta\bar{\phi})_x. \tag{2-17}$$

The fermionic part is completely analogous. Therefore, with $A\Phi = (A\phi, A\bar{\phi}, A\psi, A\bar{\psi})$,

$$(\mathrm{Z}-\Phi, \mathrm{Z}-\Phi) = N(\mathrm{Z}-A\Phi)^2 + (\Phi, -\Delta\Phi), \tag{2-18}$$

and hence

$$\int_{\mathbb{R}^2} D\mathrm{Z}\, e^{-(\mathrm{Z}-\Phi,\mathrm{Z}-\Phi)} = e^{-(\Phi,-\Delta\Phi)} \int_{\mathbb{R}^2} D\mathrm{Z}\, e^{-N(\mathrm{Z}-A\Phi)^2}. \tag{2-19}$$

In the integral on the right-hand side, we make the change of variables $\zeta \mapsto \zeta + A\phi$, $\xi \mapsto \xi + A\psi$, and similarly for $\bar{\zeta}, \bar{\xi}$. The bosonic change of variables is the usual one for Lebesgue integration, and the fermionic change of variables maintains the same $\bar{\xi}\xi$ term in $e^{-N(\mathrm{Z}-A\Phi)^2}$. Thus the integral is unchanged and hence is equal to $\int_{\mathbb{R}^2} D\mathrm{Z}\, e^{-N\mathrm{Z}^2}$, which is 1 by (2-11). This completes the proof. □

2.2. *Effective potential.*

Definition 2.4. Given a (smooth) function $p : [0, \infty) \to [0, \infty)$ such that the following integral exists, the *effective potential* $V : [0, \infty) \to \mathbb{R}$ is defined by

$$e^{-V(\mathrm{Z}^2)} = \int_{\mathbb{R}^2} D\Phi\, e^{-(\mathrm{Z}-\Phi)^2} p(\Phi^2). \tag{2-20}$$

That the right-hand side truly is a function of the form Z^2 is proved in Lemma 2.10, which we defer to Section 2.4. The consistency of this definition of V with the formula given in (1-16) is established in Proposition 2.5.

The proof of the next proposition appeals to Lemma 2.10. Recall that I_1 is a modified Bessel function of the first kind.

Proposition 2.5. *Fix* $\mathrm{Z} = (\zeta, \bar{\zeta}, \xi, \bar{\xi})$. *For any bounded smooth function* $p : [0, \infty) \to [0, \infty)$ *such that the integrals exist*,

$$\int_{\mathbb{R}^2} D\Phi\, e^{-(\mathrm{Z}-\Phi)^2} p(\Phi^2) = e^{-\mathrm{Z}^2}\left(p(0) + \int_0^\infty p(s)e^{-s}\sqrt{\frac{\mathrm{Z}^2}{s}}\, I_1(2\sqrt{\mathrm{Z}^2 s})\,ds\right), \tag{2-21}$$

and hence

$$V(t) = t - \log(p(0) + v(t)), \qquad v(t) = \int_0^\infty p(s)e^{-s}\sqrt{\frac{t}{s}}\, I_1(2\sqrt{ts})\,ds. \tag{2-22}$$

Proof. We denote the left-hand side of (2-21) by $F = F(\zeta, \bar{\zeta}, \xi, \bar{\xi})$. By Lemma 2.10, F is a function of Z^2 so it suffices to prove that

$$F(\zeta, \bar{\zeta}, 0, 0) = e^{-|\zeta|^2}\left(p(0) + \int_0^\infty p(s)e^{-s}|\zeta|\frac{1}{\sqrt{s}} I_1(2|\zeta|\sqrt{s})\,ds\right). \tag{2-23}$$

Let $\tilde{p}(s) = p(s)e^{-s}$. With $\xi = \bar{\xi} = 0$, the integrand of the left-hand side of (2-21) becomes

$$e^{-|\zeta|^2} e^{\zeta\bar{\phi}+\bar{\zeta}\phi}\, \tilde{p}(\Phi^2) = e^{-|\zeta|^2} e^{\zeta\bar{\phi}+\zeta\phi}\big(\tilde{p}(|\phi|^2) + \tilde{p}'(|\phi|^2)\psi\bar{\psi}\big). \tag{2-24}$$

Therefore, by the definition (2-10) of the integral, and with $I_0(z)=\frac{1}{2\pi}\int_0^{2\pi}e^{z\cos\theta}d\theta$ the modified Bessel function of the first kind,

$$\begin{aligned}
F(\zeta,\bar\zeta,0,0) &= -e^{-|\zeta|^2}\int_{\mathbb{R}^2} e^{\zeta r e^{-i\theta}+\bar\zeta r e^{i\theta}}\,\tilde p'(r^2)\frac{dr^2\,d\theta}{2\pi}\\
&= -e^{-|\zeta|^2}\int_0^\infty ds\,\tilde p'(s)\int_0^{2\pi} e^{\zeta\sqrt{s}e^{-i\theta}+\bar\zeta\sqrt{s}e^{i\theta}}\frac{d\theta}{2\pi}\\
&= -e^{-|\zeta|^2}\int_0^\infty ds\,\tilde p'(s)\int_0^{2\pi} e^{2|\zeta|\sqrt{s}\cos\theta}\frac{d\theta}{2\pi}\\
&= -e^{-|\zeta|^2}\int_0^\infty \tilde p'(s)I_0(2|\zeta|\sqrt{s})\,ds\\
&= e^{-|\zeta|^2}\Big(\tilde p(0)+\int_0^\infty \tilde p(s)I_0'(2|\zeta|\sqrt{s})|\zeta|\frac{1}{\sqrt{s}}ds\Big),
\end{aligned} \tag{2-25}$$

where we used integration by parts for the last equality, together with our assumption that p is bounded at infinity (this can certainly be weakened). Since $I_0'=I_1$, the proof is complete. □

Next, for later use, we state and prove a lemma that shows how the effective potential arises in various integrals. As usual, we write $V'=dV/dt$ and $\dot V=dV/d\epsilon|_{\epsilon=0}$ (with the ϵ-dependence as in (1-7); see (1-29)). We also write $Q'=1-V'$. We define forms $k_*=k_*(\mathrm{Z}^2,\zeta,\bar\zeta)$ by

$$\begin{gathered}
k_0=\zeta\,Q'(\mathrm{Z}^2),\qquad \bar k_0=\bar\zeta\,Q'(\mathrm{Z}^2),\qquad k_{00}=Q'(\mathrm{Z}^2)+Q'(\mathrm{Z}^2)^2|\zeta|^2-V''(\mathrm{Z}^2)|\zeta|^2,\\
k_{0+}=\zeta(Q'(\mathrm{Z}^2)\dot V(\mathrm{Z}^2)+\dot V'(\mathrm{Z}^2)),\qquad \bar k_{0+}=\bar\zeta(Q'(\mathrm{Z}^2)\dot V(\mathrm{Z}^2)+\dot V'(\mathrm{Z}^2)),\\
k_+=\dot V(\mathrm{Z}^2),\qquad k_{00+}=k_{00}\dot V(\mathrm{Z}^2)+\big(1+2Q'(\mathrm{Z}^2)|\zeta|^2\big)\dot V'(\mathrm{Z}^2)+\dot V''(\mathrm{Z}^2)|\zeta|^2.
\end{gathered} \tag{2-26}$$

Lemma 2.6. *The following integral formulas hold*:

$$\begin{gathered}
\int_{\mathbb{R}^2} D\Phi\,\phi p(\Phi^2)e^{-(\mathrm{Z}-\Phi)^2}=k_0e^{-V(\mathrm{Z}^2)},\qquad \int_{\mathbb{R}^2} D\Phi\,\bar\phi p(\Phi^2)e^{-(\mathrm{Z}-\Phi)^2}=\bar k_0e^{-V(\mathrm{Z}^2)},\\
\int_{\mathbb{R}^2} D\Phi\,\bar\phi\phi p(\Phi^2)e^{-(\mathrm{Z}-\Phi)^2}=k_{00}e^{-V(\mathrm{Z}^2)},\qquad \int_{\mathbb{R}^2} D\Phi\,\Phi^2 p(\Phi^2)e^{-(\mathrm{Z}-\Phi)^2}=k_+e^{-V(\mathrm{Z}^2)},\\
\int_{\mathbb{R}^2} D\Phi\,\phi\Phi^2 p(\Phi^2)e^{-(\mathrm{Z}-\Phi)^2}=k_{0+}e^{-V(\mathrm{Z}^2)},\qquad \int_{\mathbb{R}^2} D\Phi\,\bar\phi\Phi^2 p(\Phi^2)e^{-(\mathrm{Z}-\Phi)^2}=\bar k_{0+}e^{-V(\mathrm{Z}^2)},\\
\int_{\mathbb{R}^2} D\Phi\,\bar\phi\phi\Phi^2 p(\Phi^2)e^{-(\mathrm{Z}-\Phi)^2}=k_{00+}e^{-V(\mathrm{Z}^2)}.
\end{gathered} \tag{2-27}$$

Proof. Given $h:\Lambda\to\mathbb{C}$, let $(\mathrm{Z}+h)^2=(\zeta+h)(\bar\zeta+\bar h)+\xi\bar\xi$. There is no fermionic partner for h in $(\mathrm{Z}+h)^2$. Completion of the square and the definition of V give

$$\int_{\mathbb{R}^2} D\Phi\,e^{-(\mathrm{Z}-\Phi)^2}p(\Phi^2)e^{h\bar\phi+\bar h\phi}=e^{h\bar h}e^{h\bar\zeta+\bar h\zeta}\int_{\mathbb{R}^2} D\Phi\,e^{-(\mathrm{Z}-\Phi+h)^2}p(\Phi^2)=e^{h\bar h}e^{h\bar\zeta+\bar h\zeta}e^{-V((\mathrm{Z}+h)^2)}. \tag{2-28}$$

Therefore, using $d\bar{h}/dh = 0$, we obtain

$$\begin{aligned}\int_{\mathbb{R}^2} D\Phi\, \bar{\phi} p(\Phi^2) e^{-(\mathrm{Z}-\Phi)^2} &= \frac{\partial}{\partial h}\bigg|_{h=0} \int_{\mathbb{R}^2} D\Phi\, e^{-(\mathrm{Z}-\Phi)^2} p(\Phi^2) e^{h\bar{\phi}+\bar{h}\phi} \\ &= \frac{\partial}{\partial h}\bigg|_{h=0} e^{h\bar{h}} e^{h\bar{\zeta}+\bar{h}\zeta} e^{-V((\mathrm{Z}+h)^2)} = \bar{\zeta}(1-V'(\mathrm{Z}^2))e^{-V(\mathrm{Z}^2)}. \end{aligned} \tag{2-29}$$

The first and third equalities in (2-27) follow similarly. For the fourth, with $V^{(\epsilon)}$ the effective potential for $p(s)e^{-\epsilon s}$, we use

$$\begin{aligned}\int_{\mathbb{R}^2} D\Phi\, \Phi^2 p(\Phi^2) e^{-(\mathrm{Z}-\Phi)^2} &= -\frac{\partial}{\partial \epsilon}\bigg|_{\epsilon=0} \int_{\mathbb{R}^2} D\Phi\, p(\Phi^2) e^{-\epsilon\Phi^2} e^{-(\mathrm{Z}-\Phi)^2} \\ &= -\frac{\partial}{\partial \epsilon}\bigg|_{\epsilon=0} e^{-V^{(\epsilon)}(\mathrm{Z}^2)} = \dot{V}(\mathrm{Z}^2)e^{-V(\mathrm{Z}^2)}. \end{aligned} \tag{2-30}$$

The remaining three identities follow, e.g., from

$$\int_{\mathbb{R}^2} D\Phi\, \bar{\phi}\Phi^2 p(\Phi^2) e^{-(\mathrm{Z}-\Phi)^2} = -\frac{\partial}{\partial \epsilon}\bigg|_{\epsilon=0} \int_{\mathbb{R}^2} D\Phi\, \bar{\phi} p(\Phi^2) e^{-\epsilon\Phi^2} e^{-(\mathrm{Z}-\Phi)^2}, \tag{2-31}$$

together with differentiation of the right-hand sides of the first three identities with respect to ϵ. □

2.3. ***Two-point function and expected length.*** We now have what is needed to prove integral representations for the two-point function and expected length, in the next two propositions.

Proposition 2.7. *The two-point function is given by*

$$G_{01} = \int_{\mathbb{R}^2} D\mathrm{Z}\, e^{-NV(\mathrm{Z}^2)}(1-V'(\mathrm{Z}^2))^2|\zeta|^2, \tag{2-32}$$

$$G_{00} = (1-V'(0)) + G_{01} - \int_{\mathbb{R}^2} D\mathrm{Z}\, e^{-NV(\mathrm{Z}^2)} V''(\mathrm{Z}^2)|\zeta|^2. \tag{2-33}$$

Proof. By the definition of the two-point function in (1-3), followed by the supersymmetric BFS–Dynkin isomorphism (2-13) and the block-spin transformation of Lemma 2.3,

$$\begin{aligned} G_{xy} &= \int_0^\infty E_x(p_N(L_T)\mathbb{1}_{X(T)=y})\, dT \\ &= \int_{\mathbb{R}^{2N}} D\Phi\, e^{-(\Phi,-\Delta\Phi)}\bar{\phi}_x\phi_y p_N(\Phi^2) \\ &= \int_{\mathbb{R}^2} D\mathrm{Z} \int_{\mathbb{R}^{2N}} D\Phi\, \bar{\phi}_x\phi_y p_N(\Phi^2) e^{-(\mathrm{Z}-\Phi)^2}. \end{aligned} \tag{2-34}$$

The integral over $\mathbb{R}^{2N}$ on the right-hand side of (2-34) factorises into a product of N integrals over $\mathbb{R}^2$ (each an integral with respect to Φ_x at a single point x). With the definition of the effective potential in

Definition 2.4, and with the first line of (2-27), this leads to

$$\begin{aligned} G_{01} &= \int_{\mathbb{R}^2} DZ\, e^{-(N-2)V(Z^2)} \Big(\int_{\mathbb{R}^2} D\Phi_0\, \bar\phi_0\, p(\Phi_0^2) e^{-(Z-\Phi_0)^2} \Big) \Big(\int_{\mathbb{R}^2} D\Phi_1\, \phi_1\, p(\Phi_1^2) e^{-(Z-\Phi_1)^2} \Big) \\ &= \int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} (1 - V'(Z^2))^2 |\zeta|^2. \end{aligned} \tag{2-35}$$

Similarly, by the third equality of (2-27),

$$\begin{aligned} G_{00} &= \int_{\mathbb{R}^2} DZ\, e^{-(N-1)V(Z^2)} \int_{\mathbb{R}^2} D\Phi_0\, \bar\phi_0 \phi_0\, p(\Phi_0^2) e^{-(Z-\Phi_0)^2} \\ &= \int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} \big((1 - V'(Z^2)) + (1 - V'(Z^2))^2 |\zeta|^2 - V''(Z^2)|\zeta|^2 \big) \\ &= (1 - V'(0)) + G_{01} - \int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} V''(Z^2) |\zeta|^2, \end{aligned} \tag{2-36}$$

where in the last line we used (2-11) for the first term and (2-35) for the second. This completes the proof. □

For the expected length, the general procedure is the same. With k_* defined by (2-26), let

$$K_{0xy} = \begin{cases} \bar k_0 k_0 k_+ & (x = 1,\ y = 2), \\ k_{00} k_+ & (x = 0,\ y = 1), \\ \bar k_0 k_{0+} & (x = y = 1), \\ k_{00+} & (x = y = 0). \end{cases} \tag{2-37}$$

Proposition 2.8. *The expected length is given by*

$$\begin{aligned} \mathbb{E} L = \frac{1}{\chi} \Big((N-1)(N-2) \int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} K_{012} \\ + (N-1) \int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} (K_{001} + 2K_{011}) + \int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} K_{000} \Big). \end{aligned} \tag{2-38}$$

Proof. By (1-6) and $T = \sum_{x \in \Lambda} L_{T,x}$,

$$\mathbb{E} L = \frac{1}{\chi} \sum_{x,y \in \Lambda} \int_0^\infty E_0(L_{T,y}\, p_N(L_T) \mathbb{1}_{X(T)=x}) dT. \tag{2-39}$$

By the supersymmetric BFS–Dynkin isomorphism (2-13) followed by the block-spin transformation of Lemma 2.3,

$$\begin{aligned} \mathbb{E} L &= \frac{1}{\chi} \sum_{x,y \in \Lambda} \int_{\mathbb{R}^{2N}} D\Phi\, \bar\phi_0 \phi_x \Phi_y^2\, p_N(\Phi^2) e^{-(\Phi, -\Delta \Phi)} \\ &= \frac{1}{\chi} \sum_{x,y \in \Lambda} \int_{\mathbb{R}^2} DZ \int_{\mathbb{R}^{2N}} D\Phi\, \bar\phi_0 \phi_x \Phi_y^2\, p_N(\Phi^2) e^{-(Z-\Phi)^2}. \end{aligned} \tag{2-40}$$

By symmetry, it suffices to show that

$$\int_{\mathbb{R}^{2N}} D\Phi\, \bar\phi_0\phi_x \Phi_y^2 p(\Phi^2) e^{-(\mathrm{Z}-\Phi)^2} = e^{-NV(\mathrm{Z}^2)} K_{0xy}. \tag{2-41}$$

For the case of $0, x, y$ distinct, since the $N-3$ integrals for the factors with $z \neq 0, x, y$ are the same,

$$\begin{aligned}\int_{\mathbb{R}^{2N}} & D\Phi\, \bar\phi_0\phi_x \Phi_y^2 p(\Phi^2) e^{-(\mathrm{Z}-\Phi)^2} \\ &= e^{-(N-3)V(\mathrm{Z}^2)} \biggl(\int_{\mathbb{R}^2} D\Phi_0\, \bar\phi_0 p(\Phi_0^2) e^{-(\mathrm{Z}-\Phi_0)^2}\biggr) \\ &\quad \times \biggl(\int_{\mathbb{R}^2} D\Phi_x\, \phi_x p(\Phi_x^2) e^{-(\mathrm{Z}-\Phi_x)^2}\biggr) \biggl(\int_{\mathbb{R}^2} D\Phi_y\, \Phi_y^2 p(\Phi_y^2) e^{-(\mathrm{Z}-\Phi_y)^2}\biggr),\end{aligned} \tag{2-42}$$

and (2-41) follows from the definitions of k_0, $\bar k_0$, k_+. The other cases are similar. □

2.4. *Supersymmetry.* In this section, we prove Lemma 2.10, which was used in the proof of Proposition 2.5. It is possible to give a more direct proof of Proposition 2.5 without using the notion of supersymmetry. However, the proof using Lemma 2.10 is particularly elegant.

The *supersymmetry generator* is the antiderivation defined by

$$Q = \psi \frac{\partial}{\partial \phi} + \bar\psi \frac{\partial}{\partial \bar\phi} - \phi \frac{\partial}{\partial \psi} + \bar\phi \frac{\partial}{\partial \bar\psi}. \tag{2-43}$$

We say that $F = F(\phi, \bar\phi, \psi, \bar\psi)$ is *supersymmetric* if $QF = 0$. The next lemma is [Brydges and Imbrie 2003, Lemma A.4].

Lemma 2.9. *If F is even and supersymmetric then $F = f(\Phi^2)$ for some function f.*

Proof. We write $F = G + H\psi\bar\psi$, and use subscripts to denote partial derivatives. It suffices to show that there is a function f such that $G(\phi, \bar\phi) = f(|\phi|^2)$ and $H(\phi, \bar\phi) = f'(|\phi|^2)$. Since

$$0 = QF = G_\phi \psi + G_{\bar\phi}\bar\psi - \phi H \bar\psi - \bar\phi H \psi, \tag{2-44}$$

we see that $G_\phi = \bar\phi H$ and $G_{\bar\phi} = \phi H$. Therefore,

$$\begin{aligned}\frac{d}{d\theta} G(\phi e^{i\theta}, \bar\phi e^{-i\theta}) &= G_\phi(\phi e^{i\theta}, \bar\phi e^{-i\theta}) \phi i e^{i\theta} + G_{\bar\phi}(\phi e^{i\theta}, \bar\phi e^{-i\theta}) \bar\phi(-i) e^{-i\theta} \\ &= \bar\phi e^{-i\theta} H(\phi e^{i\theta}, \bar\phi e^{-i\theta}) \phi i e^{i\theta} + \phi e^{i\theta} H(\phi e^{i\theta}, \bar\phi e^{-i\theta}) \bar\phi(-i) e^{-i\theta} = 0.\end{aligned} \tag{2-45}$$

This implies that there is a function f as required. □

Lemma 2.10. *The integral $\int_{\mathbb{R}^2} D\Phi\, e^{-(\mathrm{Z}-\Phi)^2} p(\Phi^2)$ is an even supersymmetric form, and hence is a function of Z^2.*

Proof. Let $F = F(\zeta, \bar\zeta, \xi, \bar\xi) = \int_{\mathbb{R}^2} D\Phi\, e^{-(\mathrm{Z}-\Phi)^2} p(\Phi^2)$. Since $p(\Phi^2)$ is even, only even contributions in ψ, $\bar\psi$ from

$$e^{-(\mathrm{Z}-\Phi)^2} = e^{-|\zeta-\phi|^2}(1 - (\xi - \psi)(\bar\xi - \bar\psi)) \tag{2-46}$$

can contribute to the integral. Thus, within the integral, the above right-hand side can be replaced by $e^{-|\zeta-\phi|^2}(1-\xi\bar{\xi}-\psi\bar{\psi})$, and we see that F is even in $\xi, \bar{\xi}$.

To see that F is supersymmetric, let Q_{Z} act on $\mathrm{Z}=(\zeta,\bar{\zeta},\xi,\bar{\xi})$ and Q_Φ on $\Phi=(\phi,\bar{\phi},\psi,\bar{\psi})$. By definition,

$$e^{-(\mathrm{Z}-\Phi)^2}=e^{-\Phi^2}e^{-\mathrm{Z}^2}e^K \quad \text{with} \quad K=\zeta\bar{\phi}+\phi\bar{\zeta}+\xi\bar{\psi}+\psi\bar{\xi}. \tag{2-47}$$

Let $\tilde{p}(\Phi^2)=p(\Phi^2)e^{-\Phi^2}$. Since Q_{Z} is an anti-derivation, since $Q_{\mathrm{Z}}e^{-\mathrm{Z}^2}=0$ and $Q_\Phi\tilde{p}(\Phi^2)=0$ (by [Bauerschmidt et al. 2019, Example 11.4.4]), and since $Q_{\mathrm{Z}}e^K=-Q_\Phi e^K$,

$$\begin{aligned} Q_{\mathrm{Z}}F &= e^{-\mathrm{Z}^2}Q_{\mathrm{Z}}\biggl(\int_{\mathbb{R}^2} D\Phi\, e^K\tilde{p}(\Phi^2)\biggr) = e^{-\mathrm{Z}^2}\int_{\mathbb{R}^2} D\Phi\,(Q_{\mathrm{Z}}e^K)\tilde{p}(\Phi^2) \\ &= e^{-\mathrm{Z}^2}\int_{\mathbb{R}^2} D\Phi\,(-Q_\Phi e^K)\tilde{p}(\Phi^2) = -e^{-\mathrm{Z}^2}\int_{\mathbb{R}^2} D\Phi\, Q_\Phi(e^K\tilde{p}(\Phi^2)). \end{aligned} \tag{2-48}$$

The last integrand is in the image of Q_Φ, so the integral is zero (see [Bauerschmidt et al. 2019, Section 11.4.1]), and hence F is supersymmetric. By Lemma 2.9, F is therefore a function of Z^2. □

3. Proof of general results: Theorems 1.3–1.4

The proofs of Theorems 1.3–1.4 amount to application of the Laplace method to the integrals of Propositions 2.7–2.8. The application of the Laplace method depends on whether: (i) the global minimum of the effective potential is attained at zero and only at zero, or (ii) it is attained at a point $t_0>0$ with $V(t_0)<0$ or $V(t_0)=0$. Case (i) concerns the dilute phase, the second-order curve, and the tricritical point, while case (ii) concerns the dense phase and first-order curve.

3.1. *Laplace method.*

3.1.1. *Laplace method: minimum at endpoint.* For the dilute phase, the second-order curve, and the tricritical point, we use the following theorem, which can be found, e.g., in [Olver 1997, p.81]. The theorem can be extended to an asymptotic expansion to all orders, (see [Olver 1997, p.86] or [Olver 1970, p.233]), but we do not need the extension. In a corollary to the theorem, we adapt its statement to integrals of the form appearing in Propositions 2.7–2.8.

Theorem 3.1. *Suppose that $V, q:[a,b)\to\mathbb{R}$ ($b=\infty$ is allowed) are such that*

(i) *V has a unique global minimum $V(a)$ (as defined at the beginning of Section 1.3),*

(ii) *V' and q are continuous in a neighbourhood of a, except possibly at a,*

(iii) *as $t\to a^+$, $V(t)\sim V(a)+v_0(t-a)^\mu$, $V'(t)\sim\mu v_0(t-a)^{\mu-1}$, $q(t)\sim q_0(t-a)^{\lambda-1}$, with $v_0,\mu,\lambda>0$ and $q_0\neq 0$,*

(iv) *$e^{-NV(t)}q(t)$ is integrable for large N.*

Then

$$\int_a^b e^{-NV(t)}q(t)dt \sim e^{-NV(a)}\frac{q_0}{\mu(v_0N)^{\lambda/\mu}}\Gamma\biggl(\frac{\lambda}{\mu}\biggr). \tag{3-1}$$

Corollary 3.2. *Suppose that the hypotheses on V of Theorem 3.1 hold with $a = 0$ and $b = \infty$. Suppose that $F : [0, \infty)^2 \to \mathbb{R}$ is such that $q(t) = V'(t)F(t,t)$ and $q(t) = \partial_1 F(t,t)$ obey hypotheses* (ii) *and* (iv) *of Theorem 3.1, and that, as $t \downarrow 0$,*

$$F(t,t) \sim q_0 t^{\lambda_0}, \qquad \partial_1 F(t,t) \sim \lambda_1 r_0 t^{\lambda_1 - 1}. \tag{3-2}$$

If $\lambda_1 > \lambda_0 \geq 0$ then

$$\int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} F(Z^2, |\zeta|^2) \sim e^{-NV(0)} \frac{q_0}{(v_0 N)^{\lambda_0/\mu}} \Gamma\left(\frac{\mu + \lambda_0}{\mu}\right). \tag{3-3}$$

If $\lambda_1 = \lambda_0 > 0$ then

$$\int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} F(Z^2, |\zeta|^2) \sim e^{-NV(0)} \frac{q_0 - r_0}{(v_0 N)^{\lambda_0/\mu}} \Gamma\left(\frac{\mu + \lambda_0}{\mu}\right), \tag{3-4}$$

where the right-hand side is interpreted as $e^{-NV(0)} o(N^{-\lambda_0/\mu})$ if $q_0 = r_0$. For any $\lambda_1 > 0$ and $\lambda_0 \geq 0$, the integral is at most $e^{-NV(0)} O(N^{-\lambda_0/\mu} + N^{-\lambda_1/\mu})$.

Proof. By definition of the integral, and since $\bar{\xi}\xi = \frac{1}{\pi} dx dy = \frac{1}{2\pi} dr^2 d\theta$ (as in (2-11)),

$$\begin{aligned}
&\int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} F(Z^2, |\zeta|^2) \\
&\qquad = \int_{\mathbb{R}^2} DZ\, e^{-NV(|\zeta|^2)} (1 - NV'(|\zeta|^2)\xi\bar{\xi}) \big(F(|\zeta|^2, |\zeta|^2) + \partial_1 F(|\zeta|^2, |\zeta|^2)\xi\bar{\xi}\big) \\
&\qquad = \int_{\mathbb{R}^2} DZ\, e^{-NV(|\zeta|^2)} \big(NV'(|\zeta|^2) F(|\zeta|^2, |\zeta|^2) - \partial_1 F(|\zeta|^2, |\zeta|^2)\big)\bar{\xi}\xi \\
&\qquad = \int_0^\infty e^{-NV(t)} (NV'(t)F(t,t) - \partial_1 F(t,t))\, dt.
\end{aligned} \tag{3-5}$$

Now we apply Theorem 3.1. Since $V'(t)F(t,t) \sim \mu v_0 t^{\mu-1} q_0 t^{\lambda_0}$, the power of N arising for this term is $NN^{-(\mu+\lambda_0)/\mu} = N^{-\lambda_0/\mu}$. If $\lambda_1 > \lambda_0 \geq 0$ then this dominates the power $N^{-\lambda_1/\mu}$ from the $\partial_1 F(t,t)$ term and yields (3-3). If $\lambda_1 = \lambda_0 > 0$ then both terms contribute the same power of N, and (3-4) follows from $\Gamma((\mu+\lambda_1)/\mu) = (\lambda_1/\mu)\Gamma(\lambda_1/\mu)$. Finally, the general upper bound follows immediately from the above considerations. □

3.1.2. *Laplace method: minimum at interior point.* The following theorem from [Olver 1997, p.127] more than covers our needs for the case where V attains its unique global minimum in an open interval. Its analyticity assumption could be weakened, but the analyticity does hold in our setting.

Theorem 3.3. *Let $a \in [-\infty, \infty)$ and $b \in (-\infty, \infty]$. Suppose that $V, q : (a,b) \to \mathbb{R}$ are analytic, and that V has a unique global minimum at $t_0 \in (a,b)$ (as defined at the beginning of Section 1.3) with $V'(t_0) = 0$ and $V''(t_0) > 0$. Then, assuming that the integral is finite for some N,*

$$\int_a^b e^{-NV(t)} q(t) dt \sim 2e^{-NV(t_0)} \sum_{s=0}^{\infty} \Gamma(s + 1/2) \frac{b_s}{N^{s+1/2}}, \tag{3-6}$$

with (*all functions evaluated at* t_0)

$$b_0 = \frac{q}{(2V'')^{1/2}}, \tag{3-7}$$

$$b_1 = \left(2q'' - \frac{2V'''q'}{V''} + \left[\frac{5V'''^2}{6V''^2} - \frac{V''''}{2V''}\right]q\right)\frac{1}{(2V'')^{3/2}}, \tag{3-8}$$

and with b_s *as given in* [Olver 1997] *for* $s \geq 2$.

Corollary 3.4. *Suppose that* $V : (0,\infty) \to \mathbb{R}$ *is analytic and has a unique global minimum at* $t_0 \in (0,\infty)$ (*as defined at the beginning of Section 1.3*) *with* $V'(t_0) = 0$ *and* $V''(t_0) > 0$. *Given* $F : (0,\infty)^2 \to \mathbb{R}$, *suppose that the functions* $C(t) = V'(t)F(t,t)$ *and* $D(t) = \partial_1 F(t,t)$ *are analytic on* (a,b) *and that* $\int_0^\infty e^{-NV(t)}q(t)\,dt$ *is finite for* $q = C$ *and* $q = D$. *Then*

$$\int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} F(Z^2, |\zeta|^2) \sim 2e^{-NV(t_0)} \sum_{s=0}^{\infty} \frac{1}{N^{s+1/2}} (c_{s+1}\Gamma(s+3/2) - d_s\Gamma(s+1/2)), \tag{3-9}$$

where the coefficients c_s *and* d_s *are the coefficients* b_s *computed when the function* q *in Theorem 3.3 is replaced by* C *and* D, *respectively. Assuming that* $\partial_2 F(t_0,t_0) \neq 0$, *we have in particular*

$$\int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} F(Z^2, |\zeta|^2) \sim e^{-NV(t_0)} \frac{1}{N^{1/2}} \frac{\sqrt{2\pi}}{V''(t_0)^{1/2}} \partial_2 F(t_0,t_0). \tag{3-10}$$

Proof. The full expansion follows from Theorem 3.3 and (3-5). Let $A = \sqrt{2}/V''(t_0)^{1/2}\partial_2 F(t_0,t_0)$. For (3-10), since $\Gamma(1/2) = \sqrt{\pi}$, it suffices to show that $c_1 - 2d_0 = A$. Let $F' = d/dt\, F(t,t) = \partial_1 F(t,t) + \partial_2 F(t,t)$. Then

$$C' = V''F + V'F', \qquad C'' = V'''F + 2V''F' + V'F''. \tag{3-11}$$

Since $V'(t_0) = 0$ we find from (3-7)–(3-8) that

$$\begin{aligned} c_1 - 2d_0 &= \frac{1}{(2V'')^{3/2}}\left(2C'' - \frac{2V'''C'}{V''}\right) - 2\frac{\partial_1 F}{(2V'')^{1/2}} \\ &= \frac{1}{(2V'')^{3/2}}\left(2(V'''F + 2V''F') - \frac{2V'''V''F}{V''}\right) - 2\frac{\partial_1 F}{(2V'')^{1/2}}, \end{aligned} \tag{3-12}$$

and after simplification the right-hand side is equal to A. □

On the first-order curve, V has global minima $V(0) = V(t_0)$ with $V'(0) > 0$ and $V''(t_0) > 0$ (by smoothness of V, also $V'(t_0) = 0$). The following corollary covers the cases we need.

Corollary 3.5. *Suppose that* $V : (0,\infty) \to \mathbb{R}$ *is analytic and has global minima* $V(0) = V(t_0) = 0$ *for* $t_0 \in (0,\infty)$ (*as defined at the beginning of Section 1.3*) *with* $V'(0) > 0$, $V'(t_0) = 0$, *and* $V''(t_0) > 0$. *With the notation of Corollary 3.2, assume that* $\lambda_1 \geq \lambda_0 \geq 1$, *and with the notation of Corollary 3.4, assume that* $\partial_2 F(t_0,t_0) \neq 0$. *Then, for* F *obeying the assumptions of Corollaries 3.2 and 3.4,*

$$\int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} F(Z^2, |\zeta|^2) \sim \frac{1}{N^{1/2}} \frac{\sqrt{2\pi}}{V''(t_0)^{1/2}} \partial_2 F(t_0,t_0). \tag{3-13}$$

If instead $\lambda_1 > \lambda_0 = 0$, *then the right-hand side of* (3-13) *is at most* $O(1)$.

Proof. By (3-5),

$$\int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} F(Z^2, |\zeta|^2) = \int_0^\infty e^{-NV(t)}(NV'(t)F(t,t) - \partial_1 F(t,t))\, dt. \tag{3-14}$$

We divide the integral on the right-hand side into integrals over $\left(0, \frac{1}{2}t_0\right)$ and $\left(\frac{1}{2}t_0, \infty\right)$. Exactly as in the proof of Corollary 3.2 with $\mu = 1$ (only changing the integration interval), if $\lambda_1 \geq \lambda_0 \geq 1$ then the former integral is at most $O(N^{-1})$. Exactly as in the proof of Corollary 3.4, the latter integral is asymptotic to the right-hand side of (3-13), which dominates $O(N^{-1})$. If instead $\lambda_0 = 0$ then by Corollary 3.2 there can be a contribution from $t = 0$ which is $O(1)$. □

3.2. *Two-point function and susceptibility.* We now prove Theorem 1.3 and the part of Theorem 1.4 that concerns the susceptibility. For convenience, we restate Theorem 1.3 as the following proposition.

Proposition 3.6. *The two-point function has the asymptotic behaviour*:

$$G_{01} \sim \begin{cases} \dfrac{(1-V'(0))^2}{V'(0)N}\Gamma(2/1) & (\textit{dilute phase}), \\[2ex] \dfrac{1}{\left(\frac{1}{2!}V''(0)N\right)^{1/2}}\Gamma(3/2) & (\textit{second-order curve}), \\[2ex] \dfrac{1}{\left(\frac{1}{3!}V'''(0)N\right)^{1/3}}\Gamma(4/3) & (\textit{tricritical point}), \\[2ex] \dfrac{e^{-NV(t_0)}}{N^{1/2}}\dfrac{\sqrt{2\pi}}{V''(t_0)^{1/2}} & (\textit{dense phase and first-order curve}), \end{cases} \tag{3-15}$$

$$G_{00} \sim \begin{cases} 1 - V'(0) & (\textit{dilute phase and first-order curve}), \\ 1 & (\textit{second-order curve}), \\ 1 & (\textit{tricritical point}), \\ \dfrac{e^{-NV(t_0)}}{N^{1/2}}\dfrac{\sqrt{2\pi}}{V''(t_0)^{1/2}}(1 - V''(t_0)) & (\textit{dense phase}). \end{cases} \tag{3-16}$$

Proof. By Proposition 2.7,

$$G_{01} = \int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)}(1 - V'(Z^2))^2 |\zeta|^2. \tag{3-17}$$

(Note that (1-17) then follows via (3-5).) The integrability of (3-17) follows from the lower bound on V of Proposition 1.1(i). For the first three cases of (3-15), we apply Corollary 3.2 with

$$\mu = 1, \qquad v_0 = V'(0) \qquad \text{(dilute phase)}, \tag{3-18}$$

$$\mu = 2, \qquad v_0 = \tfrac{1}{2!}V''(0) \qquad \text{(second-order curve)}, \tag{3-19}$$

$$\mu = 3, \qquad v_0 = \tfrac{1}{3!}V'''(0) \qquad \text{(tricritical point)}. \tag{3-20}$$

The integrand of (3-17) involves

$$F_{01}(Z^2,|\zeta|^2)=(1-V'(Z^2))^2|\zeta|^2, \tag{3-21}$$

for which

$$\begin{aligned} F_{01}(t,t)&=(1-V'(t))^2t,\\ \partial_1F_{01}(t,t)&=2(1-V'(t))(-V''(t))t. \end{aligned} \tag{3-22}$$

From this, we see that

$$\lambda_0=1,\qquad \lambda_1\geq 2,\qquad q_0=(1-V'(0))^2\qquad \text{(dilute phase and second-order curve)}, \tag{3-23}$$

$$\lambda_0=1,\qquad \lambda_1=3,\qquad q_0=1\qquad \text{(tricritical point)}. \tag{3-24}$$

In all three cases $\lambda_1>\lambda_0$, so Corollary 3.2 gives, as desired,

$$G_{01}\sim\frac{q_0}{(v_0N)^{\lambda_0/\mu}}\Gamma\left(\frac{\mu+\lambda_0}{\mu}\right) \tag{3-25}$$

(recall that $V'(0)=0$ on the second-order curve).

For the first three cases of (3-16), by Proposition 2.7,

$$G_{00}=(1-V'(0))+G_{01}-\int_{\mathbb{R}^2}DZ\,e^{-NV(Z^2)}V''(Z^2)|\zeta|^2. \tag{3-26}$$

The integral has

$$F_{00}(Z^2,|\zeta|^2)=V''(Z^2)|\zeta|^2, \tag{3-27}$$

so

$$F_{00}(t,t)=V''(t)t,\qquad \partial_1F_{00}(t,t)=V'''(t)t. \tag{3-28}$$

The dilute, second-order, and tricritical cases have, respectively,

$$\lambda_0\geq 1,\quad \lambda_1\geq 2;\qquad \lambda_0=1,\quad \lambda_1\geq 2;\qquad \lambda_0=\lambda_1=2.$$

In all cases, the integral decays as a power of N, and since we have proved above that G_{01} also decays, we conclude that $G_{00}\to 1-V'(0)$ (with $V'(0)=0$ on the second-order curve and at the tricritical point by definition).

For the dense phase, we have

$$\partial_2F_{01}(t,t)=(1-V'(t))^2\quad\text{and}\quad \partial_2F_{00}(t,t)=V''(t), \tag{3-29}$$

and the result follows immediately from Corollary 3.4.

By definition, on the first-order curve V has global minima $V(0)=V(t_0)=0$, with $t_0\neq 0$. The hypotheses of Corollary 3.5 hold for G_{01}, with $\mu=1$ by the assumption that $V'(0)>0$ on the first-order curve, and with $\lambda_0=1$ and $\lambda_1=2$ by (3-22). The desired asymptotic formula for G_{01} then follows from (3-13). Similarly, for the integral in (3-26), we have $\mu=1$, $\lambda_0=1$, $\lambda_1\geq 2$, so by Corollary 3.5 the integral is asymptotic to a multiple of $N^{-1/2}$. It is therefore the constant term $1-V'(0)$ in (3-26) that dominates for G_{00}. □

Proof of Theorem 1.4: susceptibility. It follows from Proposition 2.7 and $\chi = G_{00} + (N-1)G_{01}$ that the susceptibility obeys

$$\chi \sim \begin{cases} \dfrac{1-V'(0)}{V'(0)}\Gamma(2/1) & \text{(dilute phase)},\\[2ex] N^{1/2}\dfrac{1}{\big(\frac{1}{2!}V''(0)\big)^{1/2}}\Gamma(3/2) & \text{(second-order curve)},\\[2ex] N^{2/3}\dfrac{1}{\big(\frac{1}{3!}V'''(0)\big)^{1/3}}\Gamma(4/3) & \text{(tricritical point)},\\[2ex] e^{-NV(t_0)}N^{1/2}\dfrac{\sqrt{2\pi}}{V''(t_0)^{1/2}} & \text{(dense phase and first-order curve)}, \end{cases} \tag{3-30}$$

as stated in Theorem 1.4. □

We remark that there is a mismatch for G_{01} and for the susceptibility as the dense phase approaches the second-order curve. For the susceptibility, the limiting value from the dense phase (as $t_0 \downarrow 0$) is

$$\chi \to N^{1/2}\sqrt{2\pi}(V''(0))^{-1/2}, \tag{3-31}$$

which is twice as big as the value

$$N^{1/2}\Gamma(3/2)\big(\tfrac{1}{2}V''(0)\big)^{-1/2} = N^{1/2}\tfrac{1}{2}\sqrt{\pi}\big(\tfrac{1}{2}V''(0)\big)^{-1/2} \tag{3-32}$$

on the second-order curve. The reason for this is clear from the proof: in the dense phase the susceptibility receives a contribution from both sides of the minimum of V at t_0, whereas on the second-order curve it only receives a contribution from the right-hand side of the minimum at 0 (cf. the two insets at the second-order curve in Figure 2).

3.3. *Expected length.* Given the asymptotic behaviour for χ in (3-30), the asymptotic formulas for the expected length stated in Theorem 1.4 will follow once we prove that

$$\chi\mathbb{E}L \sim \begin{cases} \dfrac{\dot V'(0)}{(V'(0))^2} & \text{(dilute phase)},\\[2ex] N\dfrac{\dot V'(0)}{V''(0)} & \text{(second-order curve)},\\[2ex] N^{4/3}\dfrac{\Gamma(2/3)}{3}\dfrac{1}{\big(\frac{1}{3!}V'''(0)\big)^{2/3}} & \text{(tricritical point)},\\[2ex] N^{3/2}e^{-NV(t_0)}\dfrac{\sqrt{2\pi}}{V''(t_0)^{1/2}}\dot V(t_0) & \text{(dense phase including first-order curve)}. \end{cases} \tag{3-33}$$

The proof of (3-33) is based on the following lemma. Recall that $Q' = 1 - V'$.

Lemma 3.7. *In the dilute phase, on the second-order curve, at the tricritical point, and on the first-*

order curve (*the latter for the minimum of V at $t=0$*), *the forms K_{0xy} defined in* (2-37) *have parameters $q_0, r_0, \lambda_0, \lambda_1$ as in* Corollary 3.2 *given by*:

$$\text{For } K_{012}\colon \quad q_0 = Q'(0)^2\dot V'(0), \quad r_0 = \tfrac{1}{2}Q'(0)^2\dot V'(0), \quad \lambda_0=\lambda_1=2. \tag{3-34}$$

$$\text{For } K_{001}\colon \quad q_0 = Q'(0)\dot V'(0), \quad r_0 = Q'(0)\dot V'(0), \quad \lambda_0=\lambda_1=1. \tag{3-35}$$

$$\text{For } K_{011}\colon \quad q_0 = Q'(0)\dot V'(0), \quad \lambda_0=1,\ \lambda_1=2. \tag{3-36}$$

$$\text{For } K_{000}\colon \quad q_0 = \dot V'(0), \quad \lambda_0=0,\ \lambda_1=1. \tag{3-37}$$

Proof. The desired results can be read off from the following.

By (2-37), $K_{012}(\mathrm{Z}^2, |\zeta|^2) = Q'(\mathrm{Z}^2)^2|\zeta|^2\dot V(\mathrm{Z}^2)$, so

$$K_{012}(t,t) \sim Q'(0)^2\dot V'(0)t^2, \qquad \partial_1 K_{012}(t,t) \sim Q'(0)^2\dot V'(0)t. \tag{3-38}$$

Similarly, $K_{001}(\mathrm{Z}^2, |\zeta|^2) = \big(Q'(\mathrm{Z}^2) + Q'(\mathrm{Z}^2)^2|\zeta|^2 - V''(\mathrm{Z}^2)|\zeta|^2\big)\dot V(\mathrm{Z}^2)$ obeys

$$K_{001}(t,t) \sim Q'(0)\dot V'(0)t, \qquad \partial_1 K_{001}(t,t) \sim Q'(0)\dot V'(0). \tag{3-39}$$

Next, $K_{011}(\mathrm{Z}^2, |\zeta|^2) = |\zeta|^2 Q'(\mathrm{Z}^2)\big(Q'(\mathrm{Z}^2)\dot V(\mathrm{Z}^2) + \dot V'(\mathrm{Z}^2)\big)$ obeys

$$K_{011}(t,t) \sim Q'(0)\dot V'(0)t, \partial_1 K_{011}(t,t) \sim \big(Q''(0)\dot V'(0) + Q'(0)^2\dot V'(0) + Q'(0)\dot V''(0)\big)t. \tag{3-40}$$

For the last case,

$$K_{000}(\mathrm{Z}^2, |\zeta|^2) \\ = \big(Q'(\mathrm{Z}^2) + Q'(\mathrm{Z}^2)^2|\zeta|^2 - V''(\mathrm{Z}^2)|\zeta|^2\big)\dot V(\mathrm{Z}^2) + (1+2Q'(\mathrm{Z}^2)|\zeta|^2)\dot V'(\mathrm{Z}^2) + \dot V''(\mathrm{Z}^2)|\zeta|^2 \tag{3-41}$$

obeys

$$K_{000}(t,t) \sim \dot V'(0), \qquad \partial_1 K_{000}(t,t) \sim Q'(0)\dot V'(0) + \dot V''(0). \tag{3-42}$$

This completes the proof. □

Proof of Theorem 1.4: *expected length.* It suffices to prove (3-33). By Proposition 2.8,

$$\chi\mathbb{E}L = (N-1)(N-2)\int_{\mathbb{R}^2} D\mathrm{Z}\, e^{-NV(\mathrm{Z}^2)}K_{012} \\ +(N-1)\int_{\mathbb{R}^2} D\mathrm{Z}\, e^{-NV(\mathrm{Z}^2)}(K_{001}+2K_{011}) + \int_{\mathbb{R}^2} D\mathrm{Z}\, e^{-NV(\mathrm{Z}^2)}K_{000}. \tag{3-43}$$

It will turn out that all three terms on the right-hand side contribute in the dilute phase, but in all other cases only the first term contributes to the leading behaviour. The integrability of (3-43) follows from the lower bound on V of Proposition 1.1(i).

Consider first the dilute phase. We apply Lemma 3.7 and Corollary 3.2 with $\mu = 1$, $v_0 = V'(0)$ and immediately obtain

$$\int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} K_{012} \sim \frac{(1 - V'(0))^2 \dot{V}'(0)}{(V'(0)N)^2}, \tag{3-44}$$

$$\int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} K_{001} = o(N^{-1}), \tag{3-45}$$

$$\int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} K_{011} \sim \frac{(1 - V'(0)) \dot{V}'(0)}{V'(0)N}, \tag{3-46}$$

$$\int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} K_{000} \sim \dot{V}'(0). \tag{3-47}$$

Therefore, in the dilute phase, as stated in (3-33),

$$\chi \mathbb{E} L \sim \dot{V}'(0)\left(\frac{(1 - V'(0))^2}{(V'(0))^2} + 2\frac{1 - V'(0)}{V'(0)} + 1\right) = \frac{\dot{V}'(0)}{(V'(0))^2}. \tag{3-48}$$

Consider next the second-order curve ($\mu = 2$, $v_0 = \frac{1}{2!}V''(0)$) and the tricritical point ($\mu = 3$, $v_0 = \frac{1}{3!}V'''(0)$). For these cases, Lemma 3.7 and Corollary 3.2 give

$$\int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} K_{012}(Z^2, |\zeta|^2) \sim \frac{(1 - V'(0))^2 \dot{V}'(0)}{\mu (v_0 N)^{2/\mu}} \Gamma(2/\mu). \tag{3-49}$$

By definition, on the second-order curve $V'(0) = 0$, and at the tricritical point $V'(0) = 0$ and $\dot{V}'(0) = M_1 = 1$ (recall Proposition 1.1(ii)). By Lemma 3.7 and Corollary 3.2, the integrals involving K_{001} and K_{011} are at most $O(N^{-1/\mu})$, and the one involving K_{000} is at most $O(1)$. Since the latter are multiplied by N and 1 respectively, these terms contribute order $N^{1-1/\mu}$ and N^0, and this is less than the K_{012} term which is multiplied by N^2 and hence is order $N^{2-2/\mu}$. This proves the second-order and tricritical cases of (3-33).

Next, we consider the dense phase. Let $t_0 > 0$ be the location of the global minimum of V. We have $V(t_0) < 0$, $V'(t_0) = 0$, and $V''(t_0) > 0$. By Corollary 3.4 (note that V and the various K_* satisfy the analyticity hypotheses by definition),

$$\int_{\mathbb{R}^2} DZ\, e^{-NV(Z^2)} K_{0xy} \sim e^{-NV(t_0)} \frac{A_{0xy}}{N^{1/2}}, \qquad A_{0xy} = \frac{\sqrt{2\pi}}{V''(t_0)^{1/2}} \partial_2 K_{0xy}(t_0, t_0). \tag{3-50}$$

There are order N^2 terms with $0, x, y$ distinct, order N terms where only two are distinct, and a single term where $0 = x = y$. Since each term has the same $N^{-1/2} e^{-NV(t_0)}$ behaviour, only the case with $0, x, y$ distinct can contribute to $\chi \mathbb{E} L$. Since $K_{012} = (1 - V'(Z^2))^2 \dot{V}(Z^2) |\zeta|^2$ obeys

$$\partial_2 K_{012}(t_0, t_0) = (1 - V'(t_0))^2 \dot{V}(t_0) = \dot{V}(t_0), \tag{3-51}$$

Corollary 3.4 gives

$$\chi \mathbb{E} L \sim N^2 e^{-NV(t_0)} \frac{1}{N^{1/2}} \frac{\sqrt{2\pi}}{V''(t_0)^{1/2}} \dot{V}(t_0), \tag{3-52}$$

as stated in (3-33).

Finally, we consider the first-order curve, with global minima $V(0)=V(t_0)=0$. We apply Corollary 3.5 to each of the integrals in (3-43), using $\mu = 1$ and the values of λ_i stated in Lemma 3.7. After taking into account the N-dependent factors in the terms of (3-43), we conclude from Corollary 3.5 that $\chi \mathbb{E} L$ has the same asymptotic behaviour on the first-order curve as it does in the dense phase, now with $V(t_0) = 0$, namely,

$$\chi \mathbb{E} L \sim N^{3/2} \frac{\sqrt{2\pi}}{V''(t_0)^{1/2}} \dot{V}(t_0). \tag{3-53}$$

This completes the proof. □

4. Phase diagram for the example: proof of Theorem 1.5

In this section we prove Theorem 1.5, which concerns the differentiability of the phase boundary $\nu_c(g)$ at the tricritical point g_c, and the asymptotic behaviour of the susceptibility and density, for the specific example $p(t) = e^{-t^3-gt^2-\nu t}$. According to (1-16), the effective potential is the function $V : [0, \infty) \to \mathbb{R}$ defined by

$$V(t) = t - \log(1 + v(t)), \qquad v(t) = \int_0^\infty e^{-t^3-gt^2-(\nu+1)t} \sqrt{\frac{t}{s}} I_1(2\sqrt{ts})\,ds. \tag{4-1}$$

We emphasise that V is a function of a single real variable, so in principle complete information could be extracted with sufficient effort.

4.1. ***Numerical input.*** As discussed before its statement, our proof of Theorem 1.5 relies on a numerical analysis of the effective potential (4-1), whose conclusions are summarised in Figure 2 which for convenience we repeat here in abridged form as Figure 4.

The dashed and dotted curves in Figure 4 are respectively the curves defined implicitly by $V'(0) = 1 - M_0 = 0$ and $V''(0) = M_0^2 - M_1 = 0$. Above the dashed curve $V'(0) > 0$ and below the dashed curve $V'(0) < 0$. Above the dotted curve $V''(0) < 0$ and below the curve $V''(0) > 0$. Below the curve $V'(0) = 0$ there is a unique solution to $V'(t_0) = 0$. The two curves $V'(0) = 0$ and $V''(0) = 0$ intersect at the tricritical point, which is

$$g_c = -3.2103..., \quad \nu_c = 2.0772.... \tag{4-2}$$

The solid curve (for $g \le g_c$) is the first-order curve and the dashed curve (for $g \ge g_c$) is the second-order curve. These two curves satisfy the conditions of Definition 1.2. Together the first- and second-order curves define the phase boundary $\nu_c(g)$. Points below the phase boundary are in the dense phase in the sense of Definition 1.2, and points above the phase boundary are in the dilute phase in the sense of Definition 1.2. The first few moments at the tricritical point are

$$M_0^c = M_1^c = 1, \quad M_2^c = 1.4478..., \quad M_3^c = 2.4062..., \quad M_4^c = 4.3315.... \tag{4-3}$$

To distinguish moments at the tricritical point (g_c, ν_c) from moments computed at other points (g, ν), we write the former as M_i^c and the latter simply as M_i.

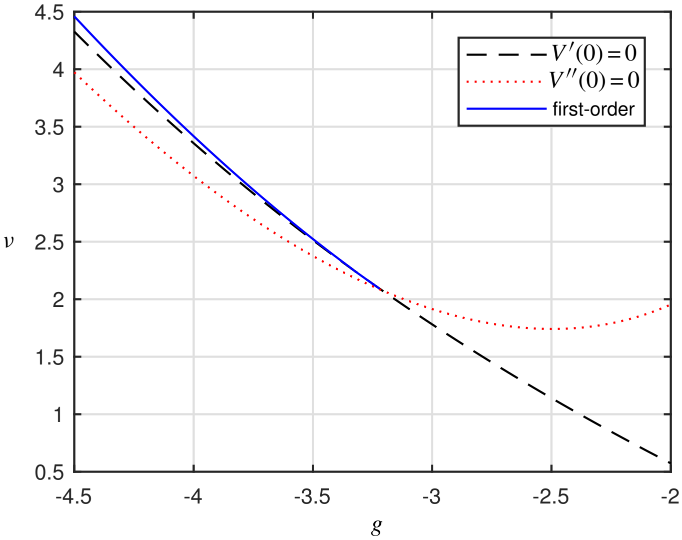


Figure 4. Phase diagram for $p(t)=e^{-t^3-gt^2-\nu t}$.

4.2. ***First derivative of phase boundary.*** We prove Theorem 1.5(i) in two lemmas. Together, the lemmas show that the phase boundary is differentiable but not twice differentiable at the tricritical point, with derivative $-M_2^c$. In this section, we prove the differentiability. The proof of the inequality of the left and right second derivatives at the tricritical point is deferred to Lemma 4.6. Here and throughout Section 4, we use the following elementary facts about derivatives of moments, for $i \geq 0$:

$$M_{i,g}=-M_{i+2},\quad M_{i,\nu}=-M_{i+1},\quad M_{i,gg}=M_{i+4},\quad M_{i,g\nu}=M_{i+3},\quad M_{i,\nu\nu}=M_{i+2}. \tag{4-4}$$

Lemma 4.1. (i) *The tangent to the curve $M_0=1$ (the entire dashed curve in Figure 4, including the second-order curve and the tricritical point) has slope $\nu_g=-M_2/M_1$.*

(ii) *The tangent to the first-order curve at the tricritical point also has slope $\nu_{c,g}=-M_2^c/M_1^c=-M_2^c$.*

Proof. (i) We write the curve $M_0=1$ as $\nu=\nu(g)$. Implicit differentiation of $M_0(g,\nu(g))=1$ with respect to g, together with (4-4), gives $0=M_{0,g}+M_{0,\nu}\nu_g=-M_2-M_1\nu_g$, so $\nu_g=-M_2/M_1$.

(ii) On the first-order curve $\nu=\nu_c(g)$ (for $g\leq g_c$), by Definition 1.2 there are two solutions to $V(t_0)=0$: $t_0=0$ and a positive $t_0>0$, which by definition characterises the first-order curve. At the positive root, $V'(t_0)=0$. At the tricritical point, $t_0=0$ and $V(0)=V'(0)=V''(0)=0$. By continuity, as $g\uparrow g_c$ we have $t_0\to 0$. Total derivatives with respect to g are denoted $\mathring{f}=df/dg$.

We differentiate $V(t_0(g,\nu_c(g)),g,\nu_c(g))=0$ with respect to g and obtain

$$V'\mathring{t}_0+V_g+V_\nu\nu_{c,g}=0. \tag{4-5}$$

For every $g\leq g_c$, $V'(t_0(g,\nu_c(g)),g,\nu_c(g))=0$, so the first term is constant in g and is equal to zero. Also, V_ν is nonzero on the first-order curve, so $\nu_{c,g}=-V_g/V_\nu$. We parametrise the first order curve as $(g(s),\nu(s))=(g_c-s,\nu_c(g_c-s))$ for $s\geq 0$. By definition, the slope of the tangent to the first-order curve is $\lim_{s\downarrow 0} \nu_{c,g}$. Also, by definition, by (1-24), and by (4-4), as $s \downarrow 0$ we have

$$\frac{V_g}{V_\nu} = \frac{v_g}{v_\nu} \sim \frac{M_{0,g}t_0}{M_{0,\nu}t_0} = \frac{M_2}{M_1} \sim M_2^c, \tag{4-6}$$

where we also used $M_1^c = 1$ and the fact that $t_0(g(s), \nu(s)) \to 0$ as $s \downarrow 0$. Therefore, at the tricritical point, $\nu_{c,g} = -M_2^c$. □

4.3. ***Susceptibility.*** We now prove Theorem 1.5(ii). Given a point $(g(0), \nu(0))$ on the second-order curve, or the tricritical point, we fix a vector $\boldsymbol{m} = (m_1, m_2)$ with base at $(g(0), \nu(0))$, which is nontangential to the second-order curve and pointing into the dilute phase. By Lemma 4.1, $\boldsymbol{n} = (M_2, M_1)$ is normal to the curve and pointing into the dilute phase, so $\boldsymbol{m} \cdot \boldsymbol{n} > 0$ (moments are evaluated at $(g(0), \nu(0))$). We define a line segment in the dilute phase that starts at our fixed point by

$$(g(s), \nu(s)) = (g(0) + sm_1, \nu(0) + sm_2) \qquad (s \in [0, 1]), \tag{4-7}$$

and set $\chi(s) = \chi(g(s), \nu(s))$. We set $M_0(s) = M_0(g(s), \nu(s))$ and define other functions similarly.

By (1-30), the infinite-volume susceptibility in the dilute phase is given by

$$\chi = \frac{1 - V'(0)}{V'(0)} = \frac{M_0(s)}{1 - M_0(s)}. \tag{4-8}$$

By (4-4) and the chain rule, $M_0(s) \sim 1 - (M_2(0)m_1 + M_1(0)m_2)s = 1 - (\boldsymbol{m} \cdot \boldsymbol{n})s$ as $s \downarrow 0$. This gives

$$\chi \sim \frac{1}{(\boldsymbol{m} \cdot \boldsymbol{n})s}, \tag{4-9}$$

which proves that χ diverges as stated in (1-35).

4.4. ***Density.*** The effective potential V is smooth in (g, ν), has a uniquely attained global minimum at $t = 0$ on the second-order curve (with $V(0) = 0$), has a uniquely attained global minimum at $t_0 > 0$ in the dense phase (with $V(t_0) < 0$), and attains its global minimum at both 0 and t_0 on the first-order curve (with $V(0) = V(t_0) = 0$). By smoothness of V, $t_0 \to 0$ as the second-order curve or tricritical point is approached. In the dense phase, the density is given by $\rho = \dot{V}(t_0)$, so as $t_0 \to 0$, $\dot{V}(t_0) \sim \dot{V}'(0)t_0 = M_1 t_0$, and at the tricritical point $M_1^c = 1$.

To prove Theorem 1.5(iii), it therefore suffices to prove the following Propositions 4.2, 4.3 and 4.5 for the asymptotic behaviour of t_0. The values of the constants A, B_i in the propositions are specified in their proofs. The fact that the phase boundary is not twice differentiable at the tricritical point is proved in Lemma 4.6, to complete the proof of Theorem 1.5(i).

Let $(g(0), \nu(0))$ be a given point on the second-order curve, or the tricritical point. We again use the normal $\boldsymbol{n} = (M_2, M_1)$ which points into the dilute phase. As in (4-7) we fix a vector $\boldsymbol{m} = (m_1, m_2)$, but now with $\boldsymbol{m} \cdot \boldsymbol{n} < 0$ so that $\boldsymbol{m}$ points into the dense phase, and we define a line segment that starts at our given point by

$$(g(s), \nu(s)) = (g(0) + sm_1, \nu(0) + sm_2) \qquad (s \in [0, 1]). \tag{4-10}$$

We consider the asymptotic behaviour of the density ρ along this segment, as $s \downarrow 0$. Let $M_i(s) = M_i(g(s), \nu(s))$ for $i = 0, 1, 2, \ldots$. An ingredient in the proofs is the asymptotic formula, as $s \downarrow 0$,

$$M_i(s) = M_i - (m_1 M_{i+2} + m_2 M_{i+1})s + \tfrac{1}{2}(m_1^2 M_{i+4} + 2m_1 m_2 M_{i+3} + m_2^2 M_{i+2})s^2 + O(s^3), \tag{4-11}$$

with all moments on the right-hand side evaluated at $s = 0$. This follows from Taylor's theorem and (4-4).

4.4.1. *Approach to second-order curve from dense phase.*

Proposition 4.2. *As the second-order curve is approached,*

$$t_0 \sim \begin{cases} \dfrac{|\boldsymbol{m}\cdot\boldsymbol{n}|}{1-M_1} s & (\textit{nontangentially from dense phase}), \\ As^2 & (\textit{tangentially from dense phase}). \end{cases} \tag{4-12}$$

The constant A is strictly positive at least along some arc of the second-order curve adjacent to the tricritical point.

Proof. We parametrise the approach to a point $(g(0), \nu(0))$ on the second order curve as $(g(s), \nu(s))$ with $s \downarrow 0$, as in (4-10), and we write $V_s(t) = V(g(s), \nu(s); t)$. On the second order curve, $V_0'(0) = 0$, $V_0''(0) > 0$ and 0 is a global minimum of V_0. On the other hand, for $s > 0$ there is a unique solution $t_0 = t_0(s) > 0$ to $V_s'(t_0(s)) = 0$. The condition $V'(t_0) = 0$ is equivalent to $1 + v(t_0) = v'(t_0)$, so

$$1 + v(0) + v'(0)t_0 = v'(0) + v''(0)t_0 + O(t_0^2). \tag{4-13}$$

By (1-24), this gives

$$1 + M_0 t_0 = M_0 + M_1 t_0 + O(t_0^2). \tag{4-14}$$

Therefore, since $t_0 = o_s(1)$, $1 - M_1(0) > 0$, and $M_0(s) - 1 = o_s(1)$,

$$t_0 = \frac{M_0 - 1}{M_0 - M_1}(1 + O(t_0)) \sim \frac{M_0 - 1}{1 - M_1(0)}. \tag{4-15}$$

Second-order curve nontangentially from dense phase. By (4-15) and (4-11),

$$t_0 \sim \frac{|\boldsymbol{m}\cdot\boldsymbol{n}|}{1 - M_1(0)} s. \tag{4-16}$$

This proves the result for the nontangential approach, for which $\boldsymbol{m}\cdot\boldsymbol{n} < 0$.

Second-order curve tangentially from dense phase. For the tangential approach, we choose $\boldsymbol{m} = (M_1, -M_2)$ so that $\boldsymbol{m}\cdot\boldsymbol{n} = 0$. By (4-11) we have $M_0(s) \sim 1 + as^2$ with $a = \frac{1}{2}(M_1^2 M_4 - 2M_1 M_2 M_3 + M_2^3)$, so now

$$t_0 \sim \frac{a}{1 - M_1(0)} s^2 \tag{4-17}$$

(all moments are evaluated at $s = 0$ here). At the tricritical point, $a^c = M_4^c - 2M_2^c M_3^c + (M_2^c)^3 > 0$ by (4-3) (cf. Lemma 4.6). By continuity, a remains positive at least along some arc of the second-order curve adjacent to the tricritical point. The constant A is $A = (1 - M_1)^{-1} a$. □

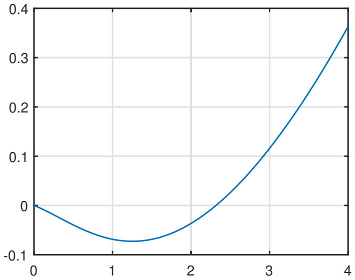


Figure 5. Effective potential V vs. t on tangent line (first-order side), at $(g,\nu)=(-3.700, 2.786)$. The point t_0 is the location of the minimum.

4.4.2. *Approach to tricritical point from dense phase.* We parametrise the approach to the tricritical point (g_c,ν_c) as $(g(s),\nu(s))$ with $s\downarrow 0$, as in (4-10), and we write $V_s(t)=V(g(s),\nu(s);t)$. For both tangential and nontangential approaches to the tricritical point from the dense phase, the approach is from below the curve $V'(0)=1-M_0=0$ and there is one solution $t_0=t_0(s)>0$ to $V_s'(t_0)=0$. An example is depicted in Figure 5. For this approach, $V_s'(0)=1-M_0<0$, and $V_s''(0)=M_0^2-M_1$ can have either sign or equal zero (the dotted curve in Figure 4 is the curve $V''(0)=0$).

Proposition 4.3. *As the tricritical point is approached,*

$$t_0\sim\begin{cases} B_0(|\boldsymbol{m}\cdot\boldsymbol{n}|s)^{1/2} & (\textit{nontangentially from dense phase}),\\ B_1 s & (\textit{tangentially from second-order side}),\\ B_2 s & (\textit{tangentially from first-order side}).\end{cases}\tag{4-18}$$

The constants B_0, B_1, B_2 are all strictly positive.

The proof of Proposition 4.3 uses the following elementary lemma, which gives an estimate for how fast $t_0\to 0$.

Lemma 4.4. *For $s\in[0,1]$, let $f_s:[0,\infty)\to\mathbb{R}$ be smooth functions, with f_s and its derivatives uniformly continuous* (*and hence uniformly bounded*) *in small s,t. Suppose that $f_0(0)=f_0'(0)=0$, $f_0''(0)>0$, and that $f_s(0)<0$ for $s>0$. Then for small $s>0$ there is a unique root $t_0=t_0(s)$ of $f_s(t_0(s))=0$ and as $s\downarrow 0$,*

$$t_0(s)=O(|f_s(0)|^{1/2}+|f_s'(0)|).\tag{4-19}$$

Proof. By the assumptions that $f_0''(0)>0$ and that $f_s''(t)$ is uniformly continuous in small s,t, there are $\delta,c>0$ such that $f_s''(t)\geq 2c$ for $s,t\leq\delta$, and hence $f_s(t)\geq f_s(0)+f_s'(0)t+ct^2=-|f_s(0)|+f_s'(0)t+ct^2$ for $s,t\leq\delta$. Since the positive root to $-a+bt+ct^2=0$ with $a,c>0$ is

$$t=\frac{1}{2c}(-b+\sqrt{b^2+4ac})\leq\frac{1}{2c}(2\max\{-b,0\}+2\sqrt{ac})=\frac{\max\{-b,0\}}{c}+\sqrt{\frac{a}{c}},\tag{4-20}$$

we see that $f_s(t)>0$ if $t>\sqrt{|f_s(0)|/c}+\max\{-f_s'(0),0\}/c$. In particular, for s small enough that $\sqrt{|f_s(0)|/c}+|f_s'(0)|/c\leq\delta$, it follows by continuity and $f_s(0)<0$ that f_s has a root $t_0=t_0(s)\in[0,\delta]$ to $f_s(t_0)=0$ satisfying $t_0=O(\sqrt{|f_s(0)|}+|f_s'(0)|)$. Since f_s is convex on $[0,\delta]$, this root is unique. □

Proof of Proposition 4.3. We apply Lemma 4.4 to $f_s(t) = V_s'(t)$. The hypotheses are satisfied: $f_s(0) = 1 - M_0(s) < 0$, $f_0'(0) = 0$, and by (1-19) and (4-3),

$$\alpha = f_0''(0) = V_0'''(0) = -\tfrac{1}{2}M_2^c + 1 > 0. \tag{4-21}$$

Taylor expansion of $V'(t_0) = 0$ gives

$$V'(0) + V''(0)t_0 + \tfrac{1}{2}V'''(0)t_0^2 + O(t_0^3) = 0. \tag{4-22}$$

It follows from Lemma 4.4 that

$$t_0 = \frac{-V''(0) \pm \sqrt{V''(0)^2 - 2V'''(0)(V'(0) + O(|V'(0)|^{3/2} + |V''(0)|^3)}}{V'''(0)} \tag{4-23}$$

Tricritical point nontangentially from dense phase. By (4-11), in this case we have

$$V_s'(0) = 1 - M_0(s) \sim -|\boldsymbol{m}\cdot\boldsymbol{n}|s, \qquad V_s''(0) \sim 1 - M_1(s) = O(s). \tag{4-24}$$

Therefore the error term in (4-23) is $O(s^{3/2})$ and hence

$$t_0 \sim \alpha^{-1}(2\alpha|\boldsymbol{m}\cdot\boldsymbol{n}|s)^{1/2}. \tag{4-25}$$

The constant B_0 is $B_0 = (2\alpha^{-1})^{1/2} = 2(2 - M_2^c)^{-1/2}$.

Tricritical point tangentially from second-order side. The slope of the tangent line at the tricritical point is $-M_2^c$, so the tangential approach is parametrised by (4-10) with $\boldsymbol{m} = (1, -M_2^c)$. On this tangent line, $M_0 > 1$ and $M_0^2 - M_1 > 0$ (see Figure 4). As above (4-17), $M_0(s) \sim 1 + a^c s^2$ with $a^c = \frac{1}{2}((M_1^c)^2 M_4^c - 2M_1^c M_2^c M_3^c + (M_2^c)^3) > 0$, and by (4-11), $M_1(s) \sim 1 - bs$ with $b = M_3^c - (M_2^c)^2 > 0$ by (4-3). Therefore,

$$V_s'(0) = 1 - M_0(s) \sim -a^c s^2, \qquad V_s''(0) = M_0(s)^2 - M_1(s) \sim bs. \tag{4-26}$$

As in the previous case, we apply Lemma 4.4 and (4-23) but this time with an error term which is $O(s^3)$. This leads to

$$t_0 \sim \alpha^{-1}(-bs + \sqrt{b^2s^2 + 2\alpha a s^2}) = \alpha^{-1}(-b + \sqrt{b^2 + 2\alpha a})s. \tag{4-27}$$

The constant B_1 is $B_1 = \alpha^{-1}(-b + \sqrt{b^2 + 2\alpha a}) > 0$.

Tricritical point tangentially from first-order side. Now we parametrise the tangential approach as in (4-10) with $\boldsymbol{m} = (-1, M_2^c)$. Now $M_0 > 1$ and $M_0^2 - M_1 < 0$ (see Figure 4). The formula (4-23) again applies with error term $O(s^3)$. Again $M_0(s) \sim 1 + as^2$, but now

$$M_1(s) \sim 1 + bs \tag{4-28}$$

due to the replacement of $\boldsymbol{m}$ by $-\boldsymbol{m}$. Therefore,

$$t_0 \sim \alpha^{-1}(bs + \sqrt{b^2s^2 + 2\alpha a s^2}) = \alpha^{-1}(b + \sqrt{b^2 + 2\alpha a})s. \tag{4-29}$$

The constant B_2 is $B_2 = \alpha^{-1}(b + \sqrt{b^2 + 2\alpha a}) > 0$. □

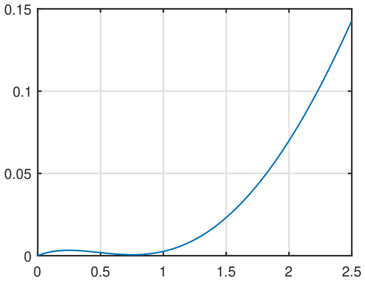


Figure 6. Effective potential V vs. t on first-order curve, at $(g, \nu) = (-3.700, 2.864)$. The point t_0 is the unique positive solution to $V(t_0) = 0$.

4.4.3. *Approach to tricritical point along first-order curve.* We now complete the proof of Theorem 1.5, by proving the last case of Theorem 1.5(iii) in Proposition 4.5 and the remaining part of Theorem 1.5(i) (namely the failure of the phase boundary to be twice differentiable) in Lemma 4.6. The focus is on the location t_0 of the nonzero minimum of the effective potential on the first-order curve. Figure 6 shows a typical V.

Proposition 4.5. *As the tricritical point is approached along the first-order curve parametrised by* $(g(s), \nu(s)) = (g_c - s, \nu_c(g_c - s))$,

$$t_0 \sim B_3 s \qquad (\textit{along first-order curve}). \tag{4-30}$$

The constant B_3 is strictly positive.

Proof. On the first-order curve, $V_s'(0) > 0$ and $V_s''(0) < 0$ (see Figure 4), and there is a unique $t_0 > 0$ such that $V_s(t_0) = 0$. At this minimum, $V_s'(t_0) = 0$ and $V_s''(t_0) > 0$. Under Taylor expansion, together with the fact that $V_s(0) = 0$, the equations $V_s(t_0) = 0$ and $V_s'(t_0) = 0$ become

$$0 = t_0\big(V_s'(0) + \tfrac{1}{2!}V_s''(0)t_0 + \tfrac{1}{3!}V_s'''(0)t_0^2 + O(t_0^3)\big) \tag{4-31}$$

$$0 = V_s'(0) + V_s''(0)t_0 + \tfrac{1}{2!}V_s''(0)t_0^2 + O(t_0^3). \tag{4-32}$$

By the chain rule, $\frac{d}{ds}\nu_c(g(s)) = -\nu_{c,g}(g(s))$, and by Lemma 4.1, $-\nu_{c,g}(g(s)) \to M_2^c$ as $s \downarrow 0$. Therefore, as in (4-28),

$$M_1(s) \sim 1 + bs, \tag{4-33}$$

with $b = M_3^c - (M_2^c)^2 > 0$. Also, since

$$\frac{d}{ds}M_0(s) = -M_{0,g}(s) - M_{0,\nu}(s)\nu_c'(g_c - s) \to M_2^c - M_1^c(-M_2^c/M_1^c) = 0 \qquad (s \to 0), \tag{4-34}$$

it follows that

$$V_s''(0) = M_0(s)^2 - M_1(s) \sim -bs. \tag{4-35}$$

We substitute this into (4-31)–(4-32) and use

$$V'''(0) = \alpha = 1 - \tfrac{1}{2}M_2^c > 0, \tag{4-36}$$

to find that

$$V'(0) - \frac{bs}{2}t_0 + \frac{\alpha}{6}t_0^2 = O(t_0(t_0+s)^2), \tag{4-37}$$

$$V'(0) - bst_0 + \frac{\alpha}{2}t_0^2 = O(t_0(t_0+s)^2), \tag{4-38}$$

and hence

$$V'(0) - \frac{bs}{4}t_0 = O(t_0(t_0+s)^2). \tag{4-39}$$

We substitute (4-39) into either of (4-31)–(4-32), and after cancellation of a factor t_0, we obtain

$$-\frac{bs}{4} + \frac{\alpha}{6}t_0 = O(t_0+s)^2. \tag{4-40}$$

Since $t_0 = o_s(1)$, this gives

$$t_0 \sim \frac{3b}{2\alpha}s = B_3 s, \tag{4-41}$$

so $B_3 = 3b/(2\alpha) = 3(M_3^c - (M_2^c)^2)/(2 - M_2^c) > 0$ and the proof is complete. □

Note that the conclusion (4-30) agrees with the naive argument (ignoring the error term) that the quadratic in (4-31) will have a unique root precisely when the discriminant vanishes, i.e., when

$$V_s'(0) = \frac{3}{8}\frac{V_s''(0)^2}{V_s'''(0)}, \tag{4-42}$$

and in this case the root is $t_0 = \frac{3}{2}|V_s''(0)|/V_s'''(0) \sim \frac{3}{2}bs/\alpha$.

Finally, we prove the remaining part of Theorem 1.5(iii), concerning the behaviour of the density as the tricritical point is approached along the first-order curve.

Lemma 4.6. (i) *At the tricritical point, the second derivative of the second-order curve is given by*

$$\nu_{gg} = M_4^c - 2M_3^c M_2^c + (M_2^c)^3 > 0. \tag{4-43}$$

(ii) *At the tricritical point, the second derivative of the first-order curve is instead*

$$\nu_{c,gg} = M_4^c - 2M_3^c M_2^c + (M_2^c)^3 + 3b^2/(4\alpha), \tag{4-44}$$

with $b = M_3^c - (M_2^c)^2 > 0$ *and with* $\alpha = 1 - \frac{1}{2}M_2^c > 0$.

Proof. (i) The second-order curve $\nu = \nu(g)$ is given by $V'(0) = 0$. Since $V'(0) = 1 - M_0$, a first differentiation with respect to g gives $0 = M_{0,g} + M_{0,\nu}\nu_g$, and a second differentiation gives

$$0 = M_{0,gg} + 2M_{0,\nu g}\nu_g + M_{0,\nu\nu}\nu_g^2 + M_{0,\nu}\nu_{gg}. \tag{4-45}$$

We use $\nu_g = -M_2/M_1$ from Lemma 4.1(i), (4-4), and $M_1^c = 1$ to see that, at the tricritical point,

$$\nu_{c,gg} = M_4^c - 2M_3^c M_2^c + (M_2^c)^3. \tag{4-46}$$

By (4-3), $\nu_{c,gg} > 0$.

(ii) By (4-5), on the first-order curve we have $V_g + V_\nu \nu_{c,g} = 0$. Total derivatives with respect to g are denoted $\mathring{f} = df/dg$. Differentiation with respect to g gives

$$\mathring{V}_g + \mathring{V}_\nu \nu_{c,g} + V_\nu \nu_{c,gg} = 0, \tag{4-47}$$

so, since $V_\nu \sim M_1^c t_0 = t_0$ by (1-24),

$$\nu_{c,gg} = -\frac{1}{V_\nu}(\mathring{V}_g + \mathring{V}_\nu \nu_{c,g}) \sim -\frac{1}{t_0}(\mathring{V}_g + \mathring{V}_\nu \nu_{c,g}). \tag{4-48}$$

We need to compute the coefficient of t_0 in $\mathring{V}_g + \mathring{V}_\nu \nu_{c,g}$. Since

$$\mathring{V}_g = V_g' \mathring{t}_0 + V_{gg} + V_{g\nu} \nu_{c,g}, \tag{4-49}$$

$$\mathring{V}_\nu = V_\nu' \mathring{t}_0 + V_{\nu g} + V_{\nu\nu} \nu_{c,g}, \tag{4-50}$$

we obtain

$$\nu_{c,gg} \sim -\frac{1}{t_0}\big((V_g' + V_\nu' \nu_{c,g})\mathring{t}_0 + V_{gg} + V_{\nu g}\nu_{c,g} + (V_{g\nu} + V_{\nu\nu}\nu_{c,g})\nu_{c,g}\big). \tag{4-51}$$

From (1-24),

$$V_{gg} = -\frac{v_{gg}}{1+v} + \frac{v_g^2}{(1+v)^2} \sim -v_{gg} + O(t_0^2) \sim -M_{0,gg} t_0 \sim -M_4^c t_0. \tag{4-52}$$

Similarly,

$$V_{g\nu} \sim -M_3^c t_0 \quad \text{and} \quad V_{\nu\nu} \sim -M_2^c t_0. \tag{4-53}$$

Therefore, since $\nu_{c,g} \sim -M_2^c$ by Lemma 4.1(ii), and since $-\mathring{t}_0 \to B_3 = 3b/(2\alpha)$ as $g \uparrow g_c$ by Proposition 4.5,

$$\nu_{c,gg} \sim (M_4^c - 2M_2^c M_3^c + (M_2^c)^3) + B_3 \frac{1}{t_0}(V_g' + V_\nu' \nu_{c,g}). \tag{4-54}$$

Let $f(g) = V_g' + V_\nu' \nu_{c,g}$. Since $f(g_c) = M_2^c + M_1^c(-M_2^c) = 0$, we need the next order term. Direct computation gives

$$V_g' + V_\nu' \nu_{c,g} = \frac{1}{(1+v)^2}\left(-v_g'(1+v) + v' v_g - (-v_\nu'(1+v) + v' v_\nu)\frac{v_g}{v_\nu}\right). \tag{4-55}$$

By (1-24) and (4-4), together with $M_0^c = M_1^c = 1$

$$v = M_0 t + \tfrac{1}{2} M_1 t^2 + \cdots \sim t + \tfrac{1}{2} t^2 + \cdots , \tag{4-56}$$

$$v' = M_0 + M_1 t + \cdots \sim 1 + t + \cdots , \tag{4-57}$$

$$v_g = M_{0,g} t + \tfrac{1}{2} M_{1,g} t^2 + \cdots \sim -M_2^c t - \tfrac{1}{2} M_3^c t^2 + \cdots , \tag{4-58}$$

$$v_\nu = M_{0,\nu} t + \tfrac{1}{2} M_{1,\nu} t^2 + \cdots \sim -t - \tfrac{1}{2} M_2^c t^2 + \cdots , \tag{4-59}$$

$$v_g' \sim -M_2^c - M_3^c t, \tag{4-60}$$

$$v_\nu' \sim -1 - M_2^c t. \tag{4-61}$$

Therefore,

$$\begin{aligned} V_g' + V_\nu' \nu_{c,g} &\sim (M_2^c + M_3^c t)(1+t) - M_2 t - ((1+M_2^c t)(1+t) - t)\frac{M_2^c + \frac{1}{2} M_3^c t}{1+\frac{1}{2} M_2^c t} \\ &\sim M_2^c + M_3^c t - (1+M_2 t)\big(M_2^c + \tfrac{1}{2} M_3^c t\big)\big(1 - \tfrac{1}{2} M_2^c t\big) \\ &\sim \tfrac{1}{2}(M_3^c - (M_2^c)^2)t = \frac{b}{2}t, \end{aligned} \tag{4-62}$$

and the proof is complete. □

Acknowledgements

The work of Slade was supported in part by NSERC of Canada. We thank David Brydges for discussions concerning Section 2 and for showing us the proof of Proposition 2.5 via Lemma 2.10. We are grateful to Anthony Peirce and Kay MacDonald for assistance with Matlab programming, and to an anonymous referee for helpful advice.

References

[Arovas 2019] D. Arovas, "Thermodynamics and statistical mechanics", Lecture notes, 2019, https://courses.physics.ucsd.edu/2017/Spring/physics210a/LECTURES/BOOK_STATMECH.pdf.

[Bauerschmidt et al. 2017] R. Bauerschmidt, G. Slade, and B. C. Wallace, "Four-dimensional weakly self-avoiding walk with contact self-attraction", *J. Stat. Phys.* **167**:2 (2017), 317–350. MR

[Bauerschmidt et al. 2019] R. Bauerschmidt, D. C. Brydges, and G. Slade, *Introduction to a renormalisation group method*, Lecture Notes in Mathematics **2242**, Springer, 2019. MR

[Bauerschmidt et al. 2020] R. Bauerschmidt, M. Lohmann, and G. Slade, "Three-dimensional tricritical spins and polymers", *J. Math. Phys.* **61**:3 (2020), [art. id. 033302]. MR

[Bousquet-Mélou et al. 2005] M. Bousquet-Mélou, A. J. Guttmann, and I. Jensen, "Self-avoiding walks crossing a square", *J. Phys. A* **38**:42 (2005), 9159–9181. MR

[Brydges and Imbrie 2003] D. C. Brydges and J. Z. Imbrie, "Green's function for a hierarchical self-avoiding walk in four dimensions", *Comm. Math. Phys.* **239**:3 (2003), 549–584. MR

[Deng et al. 2019] Y. Deng, T. M. Garoni, J. Grimm, A. Nasrawi, and Z. Zhou, "The length of self-avoiding walks on the complete graph", *J. Stat. Mech. Theory Exp.* 10 (2019), [art. id. 103206]. MR

[Duminil-Copin et al. 2014] H. Duminil-Copin, G. Kozma, and A. Yadin, "Supercritical self-avoiding walks are space-filling", *Ann. Inst. Henri Poincaré Probab. Stat.* **50**:2 (2014), 315–326. MR

[Duplantier 1986] B. Duplantier, "Tricritical Polymer Chains in or below Three Dimensions", *Europhysics Letters* (*EPL*) **1**:10 (may 1986), 491–498.

[Duplantier 1987] B. Duplantier, "Geometry of polymer chains near the theta-point and dimensional regularization", *J. Chem. Phys.* **86**:7 (1987), 4233–4244. MR

[Duplantier and Saleur 1987a] B. Duplantier and H. Saleur, "Exact critical properties of two-dimensional dense self-avoiding walks", *Nuclear Phys. B* **290**:3 (1987), 291–326. MR

[Duplantier and Saleur 1987b] B. Duplantier and H. Saleur, "Exact Tricritical Exponents for Polymers at the θ Point in Two-dimensions", *Phys. Rev. Lett.* **59** (1987), 539–542.

[de Gennes 1979] P. G. de Gennes, *Scaling Concepts in Polymer Physics*, Cornell University Press, Ithaca, N.Y., 1979.

[Golowich and Imbrie 1995] S. E. Golowich and J. Z. Imbrie, "The broken supersymmetry phase of a self-avoiding random walk", *Comm. Math. Phys.* **168**:2 (1995), 265–319. MR

[Hammond and Helmuth 2019] A. Hammond and T. Helmuth, "Self-attracting self-avoiding walk", *Probab. Theory Related Fields* **175**:3-4 (2019), 677–719. MR

[den Hollander 2009] F. den Hollander, *Random polymers*, vol. 1974, Lecture Notes in Mathematics **37**, Springer, 2009. MR

[Madras 1995] N. Madras, "Critical behaviour of self-avoiding walks that cross a square", *J. Phys. A* **28**:6 (1995), 1535–1547. MR

[Olver 1970] F. W. J. Olver, "Why steepest descents?", *SIAM Rev.* **12** (1970), 228–247. MR

[Olver 1997] F. W. J. Olver, *Asymptotics and special functions*, AKP Classics, A K Peters, Ltd., Wellesley, MA, 1997. MR

[Pétrélis and Torri 2018] N. Pétrélis and N. Torri, "Collapse transition of the interacting prudent walk", *Ann. Inst. Henri Poincaré D* **5**:3 (2018), 387–435. MR

[Slade 2020] G. Slade, "Self-avoiding walk on the complete graph", online publication March 2020. To appear *J. Math. Soc. Japan*.

Received 1 Nov 2019. Accepted 24 Apr 2020.

ROLAND BAUERSCHMIDT: rb812@cam.ac.uk
Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Wilberforce Road, Cambridge, United Kingdom

GORDON SLADE: slade@math.ubc.ca
Department of Mathematics, University of British Columbia, Vancouver BC, Canada

PROBABILITY and
MATHEMATICAL
PHYSICS

msp

Vol. 1, No. 1, 2020
https://doi.org/10.2140/pmp.2020.1.205

OBSERVABLES OF COLOURED STOCHASTIC VERTEX MODELS AND THEIR POLYMER LIMITS

ALEXEI BORODIN AND MICHAEL WHEELER

In the context of the coloured stochastic vertex model in a quadrant, we identify a family of observables whose averages are given by explicit contour integrals. The observables are certain linear combinations of q-moments of the coloured height functions of the model. In a polymer limit, this yields integral representations for moments of partition functions of strict-weak, semidiscrete Brownian, and continuum Brownian polymers with varying beginning and ending points of the polymers.

1. Introduction

At least since the work of Kardar [1987], moments of polymer partition functions and related quantities have been an indispensable tool in analyzing models from the so-called Kardar–Parisi–Zhang (KPZ) universality class in (1+1) dimensions [Kardar et al. 1986], such as various (integrable) models of directed polymers in a random environment, exclusion and zero-range processes, and random growth models. An advanced (although nonrigorous) replica analysis of moments allowed Calabrese, Le Doussal and Rosso [2010] and Dotsenko [2010] to derive the long-time asymptotics of the so-called narrow wedge solution to the KPZ equation. An alternative approach to the asymptotics of this solution based on the pioneering work of Tracy and Widom [2008a; 2008b; 2009] was developed rigorously by Amir, Corwin and Quastel [2011] and in the physics literature by Sasamoto and Spohn [2010]. (These two approaches were mostly reconciled by Borodin, Corwin and Sasamoto [2014].) The moment method became rigorous with the appearance of explicit integral representations for q-moments of suitable q-deformed models together with the asymptotic analysis of their generating functions in [Borodin and Corwin 2014]. (A simpler approach to the asymptotics through moments was later suggested in [Borodin 2018]

MSC2010: 60K35.
Keywords: observables, stochastic vertex models, polymer models.

and [Borodin and Olshanski 2017].) Many papers with analysis of q-moments of various integrable probabilistic systems have been written since then. The stochastic six vertex model, first introduced by Gwa and Spohn [1992], was also found to be accessible via such a route in work of Borodin, Corwin and Gorin [2016b]; see also works by Borodin, Corwin, Petrov and Sasamoto [2015a; 2015b], Corwin and Petrov [2016] and Borodin and Petrov [2017; 2018a] for various approaches to the q-moments of this model.

Kuniba, Mangazeev, Maruyama and Okado [2016] introduced Yang–Baxter integrable *coloured* stochastic vertex models; see also works by Bosnjak and Mangazeev [2016] and Aggarwal, Borodin and Bufetov [2019]. Colours correspond to the simple roots in the underlying quantum affine algebra $U_q(\widehat{\mathfrak{sl}_{n+1}})$, with $n = 1$ corresponding to the *colourless* or rank-1 case considered previously.

Our recent paper [Borodin and Wheeler 2018] offered an extensive algebraic analysis of these coloured models and uncovered certain distributional correspondences between coloured and (much more studied) colourless ones. Such correspondences were further extended by Borodin and Bufetov [2019] and Borodin, Gorin and Wheeler [2019], and they gave access to various unknown marginals of the coloured models. However, so far no explicit formulas for observables of the coloured models have been found, apart from those that arise through matching with colourless models. The primary goal of this paper is to remedy this fact.

We obtain explicit integral representations for certain linear combinations of q-moments of coloured height functions for the coloured stochastic vertex model in a quadrant. Further, following a path worked out in [Borodin et al. 2019], we degenerate the (fully fused) coloured vertex model to directed polymers, thus obtaining formulas for joint moments of polymer partition functions with different starting points in the same noise field. The limiting objects include the KPZ equation (equivalently, the continuum Brownian polymer), the O'Connell–Yor semi-discrete Brownian polymer [2001], the strict-weak or Gamma polymer of Corwin, Seppäläinen and Shen [2015] and O'Connell and Ortmann [2015], and the Beta polymer of Barraquand and Corwin [2017].

Let us describe our results in more detail.

The coloured stochastic vertex model in a quadrant can be viewed as a Markovian recipe of constructing random coloured up-right paths in $\mathbb{Z}_{\geq 1} \times \mathbb{Z}_{\geq 1}$ with the colours labelled by natural numbers. Let us also initially assume that no horizontal edge of the lattice can be occupied by more than a single path; this restriction will eventually be removed. The model depends on a quantization parameter $q \in \mathbb{C}$, a spin parameter $s \in \mathbb{C}$, and row *rapidities* denoted by $x_1, x_2, \ldots \in \mathbb{C}$.

Along the boundary of the quadrant, we demand that no paths enter the quadrant from the bottom. On the other hand, a single coloured path enters the quadrant from the left in each row. We assume that the colours of the paths entering on the left are weakly increasing in the upward direction, and denote by $\lambda_1 \geq 0$ the number of paths of colour 1, by $\lambda_2 \geq 0$ the number of paths of colour 2, etc. Let us also denote by

$$\ell_k = \lambda_1 + \cdots + \lambda_k, \quad k \geq 1, \tag{1-1}$$

the partial sums of this sequence and also set $\ell_0 = 0$.

Once the paths are specified along the boundary, they progress in the up-right direction within the quadrant using certain interaction probabilities, also known as *vertex weights*. For each vertex of the

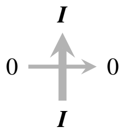
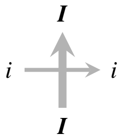
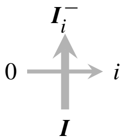
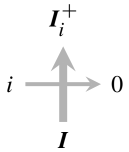
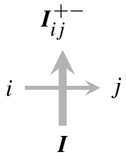
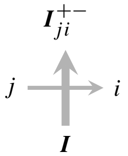
 $\frac{1-sxq^{\boldsymbol{I}_{\geq 1}}}{1-sx}$	 $\frac{s(sq^{I_i}-x)q^{\boldsymbol{I}_{\geq i+1}}}{1-sx}$	 $\frac{sx(q^{I_i}-1)q^{\boldsymbol{I}_{\geq i+1}}}{1-sx}$
 $\frac{1-s^2q^{\boldsymbol{I}_{\geq 1}}}{1-sx}$	 $\boldsymbol{I}_{ij}^{+-}$ i j $\boldsymbol{I}$ $\frac{sx(q^{I_j}-1)q^{\boldsymbol{I}_{\geq j+1}}}{1-sx}$	 $\boldsymbol{I}_{ji}^{+-}$ j i $\boldsymbol{I}$ $\frac{s^2(q^{I_i}-1)q^{\boldsymbol{I}_{\geq i+1}}}{1-sx}$

Table 1. The vertex weights.

lattice, once we know the colours of the entering paths along the bottom and left adjacent edges, we decide on the colours of the exiting paths along the top and right edges according to those probabilities.[1] They are given by Table 1, where it is assumed that x is the rapidity of the row to which the vertex belongs, $1 \leq i < j$, $\boldsymbol{I} = (I_1, I_2, \dots)$ denotes a vector whose coordinates I_k are equal to the number of paths of colour k that enter the vertex from the bottom, and $\boldsymbol{I_i}^{\pm} = \boldsymbol{I} \pm \boldsymbol{e}_i$ with $\boldsymbol{e_i}$ being the standard basis vector with 1 as its i-th coordinate and all other coordinates equal to 0. We also use the notation $\boldsymbol{I}_{ab}^{+-} = \boldsymbol{I} + \boldsymbol{e}_a - \boldsymbol{e}_b$ and $\boldsymbol{I}_{\geq a} = I_a + I_{a+1} + \cdots$. The weight of any vertex that does not fall into one of the six categories in Table 1 is set to zero. See Section 2 for the origin of these weights.

The *stochasticity* of the weights encodes the fact that these weights add up to 1 when summed over all possible states of the outgoing edges with the states of incoming edges being fixed. When the parameters of the model are such that all the weights are nonnegative, we obtain bona fide transition probabilities. However, if we agree to deal with complex-valued discrete distributions, we will not actually need the positivity assumption for our main algebraic result.

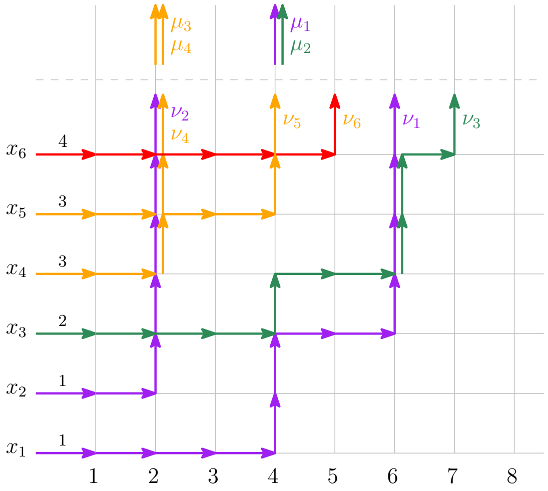
Let us now fix $n \geq 1$ and focus on the state of the model between row n and row $n+1$. That is, let us record the locations where the paths of colours $1, 2, \dots$ exit the n-th row upwards as an n-dimensional coloured vector (equivalently, a *coloured composition*) ν with coordinates in $\mathbb{Z}_{\geq 1}$. By colour conservation that our weights observe, the counts of different colours in such a vector are provided by a (finite) sequence $\lambda = (\lambda_1, \lambda_2, \dots)$ corresponding to the colours that enter via the left edges of the first n rows; we call λ the *labelling composition*.[2] We will write the coordinates of ν as

$$\nu = (\nu_1 \geq \cdots \geq \nu_{\ell_1} \mid \nu_{\ell_1+1} \geq \cdots \geq \nu_{\ell_2} \mid \nu_{\ell_2+1} \geq \cdots \geq \nu_{\ell_3} \mid \cdots), \tag{1-2}$$

where the groups separated by vertical bars list the coordinates of the paths between rows n and $n+1$ of colour 1, colour 2, etc. See Figure 1 for an example.

[1] Such decisions at different vertices are independent.

[2] For convenience of notation, we assume that the left-incoming colours in rows n and $n+1$ are different, so that we do not need to split the last coordinate in this sequence.



$$\kappa = (1, 1, 2, 0)$$

$$\mu = (4\,|\,4\,|\,2, 2) = 1^{\mathbf{0}} 2^{(0,0,2,0)} 3^{\mathbf{0}} 4^{(1,1,0,0)} 5^{\mathbf{0}} 6^{\mathbf{0}} \cdots$$

$$\lambda = (2, 1, 2, 1)$$

$$\nu = (6, 2\,|\,7\,|\,4, 2\,|\,5)$$

$$H_1^{\mu}(4) = H_2^{\mu}(4) = 1$$

$$H_3^{\mu}(4) = H_4^{\mu}(4) = 0$$

$$H_1^{\nu}(4) = H_2^{\nu}(4) = H_3^{\nu}(4) = H_4^{\nu}(4) = 1$$

$$H_1^{\nu/\mu} = H_2^{\nu/\mu}(4) = 0$$

$$H_3^{\nu/\mu}(4) = H_4^{\nu/\mu}(4) = 1$$

Figure 1. A possible configuration of the coloured vertex model.

Our observables on random ν's are also indexed by coloured compositions. Let us fix a composition $\kappa = (\kappa_1, \kappa_2, \dots)$ satisfying $\kappa \le \lambda$ coordinate-wise (otherwise the observable vanishes identically), and a coloured composition μ with coordinates in $\mathbb{Z}_{\ge 1}$ whose colour counts are given by κ. It is also convenient to parametrize μ differently by writing $\mu = 1^{\boldsymbol{m}^{(1)}} 2^{\boldsymbol{m}^{(2)}} \cdots$, where the vectors $\boldsymbol{m}^{(j)} = (m_1^{(j)}, m_2^{(j)}, \dots)$ are such that their coordinates $m_i^{(j)}$ count the number of parts of μ of colour i equal to j; see Figure 1.

Yet another way to describe coloured compositions is through their *coloured height functions* defined as follows; see Figure 1:

$$H_i^{\rho}(x) = \#\{j : \mathrm{colour}(\rho_j) = i,\ \rho_j \ge x\};$$
$$H_{>i}^{\rho}(x) = \sum_{k>i} H_k^{\rho}(x), \qquad H_{\ge i}^{\rho}(x) = \sum_{k \ge i} H_k^{\rho}(x); \qquad H_*^{\nu/\mu} \equiv H_*^{\nu} - H_*^{\mu}.$$

Here colour(ρ_j) refers to the number of the block in the splitting of the form (1-2) for the coloured composition ρ that ρ_j belongs to.

For κ-coloured μ as above, let us now define an observable $\mathcal{O}_\mu$, whose values on λ-coloured ν's are given by

$$\mathcal{O}_\mu(\nu) = \prod_{i,j \ge 1} q^{m_i^{(j)} H_{>i}^{\nu/\mu}(j+1)} \binom{H_i^{\nu/\mu}(j+1)}{m_i^{(j)}}_q$$
$$= \prod_{i,j \ge 1} \frac{(q^{H_{>i}^{\nu/\mu}(j+1)} - q^{H_{\ge i}^{\nu/\mu}(j+1)})(q^{H_{>i}^{\nu/\mu}(j+1)} - q^{H_{\ge i}^{\nu/\mu}(j+1)-1}) \cdots (q^{H_{>i}^{\nu/\mu}(j+1)} - q^{H_{\ge i}^{\nu/\mu}(j+1)-m_i^{(j)}+1})}{(q;q)_{m_i^{(j)}}}.$$

In the *rainbow case* $\lambda = (1, 1, \dots, 1)$ of all colours being different, $\mathcal{O}_\mu$ simplifies to

$$\mathcal{O}_\mu^{\text{rainbow}}(\nu) = \prod_{i,j\,:\,m_i^{(j)}=1} \mathbf{1}_{H_i^{\nu/\mu}(j+1)=1}\, q^{H_{>i}^{\nu/\mu}(j+1)} = \prod_{i,j\,:\,m_i^{(j)}=1} \frac{q^{H_{>i+1}^{\nu/\mu}(j+1)} - q^{H_{>i}^{\nu/\mu}(j+1)}}{q-1},$$

and in the *colour-blind case* $\lambda = (n)$ it is given by

$$\mathcal{O}_{\mu}^{\text{colour-blind}}(\nu) = \frac{(1-q^{H^{\nu}(\mu_1+1)})(1-q^{H^{\nu}(\mu_2+1)-1})\cdots(1-q^{H^{\nu}(\mu_m+1)-m+1})}{\prod_{j\geq 1}(q;q)_{\text{mult}_j(\mu)}},$$

where $H^{\nu} = H^{\nu}_{\geq 1}$ is the colour-blind height function, $\text{mult}_j(\mu) = \#\{i : \mu_i = j\}$, and $m = \ell(\mu)$ is the number of parts in μ.

In order to write down an integral representation for the average of $\mathcal{O}_{\mu}$, we need to introduce certain rational functions $f_{\mu}(\kappa; z_1, \dots, z_m)$ with the number of variables $m = \ell(\mu)$ equal to the number of parts of μ. In the rainbow case of pairwise distinct colours, coloured compositions are the same as uncoloured ones, and for anti-dominant compositions $\delta = (\delta_1 \leq \delta_2 \leq \dots \leq \delta_m)$ these functions are completely factorized:

$$f_{\delta}(z_1, \dots, z_m) = \frac{\prod_{j\geq 0}(s^2;q)_{\text{mult}_j(\delta)}}{\prod_{i=1}^{m}(1-sz_i)} \prod_{i=1}^{m}\left(\frac{z_i - s}{1-sz_i}\right)^{\delta_i}.$$

Note that we dropped κ from the notation for f_{δ} as it plays no role in the rainbow situation. For non-anti-dominant μ's, one way to define f_{μ} is by the following recursion that allows us to move step by step from anti-dominant compositions toward the dominant ones: If $\mu_i < \mu_{i+1}$ for some $1 \leq i \leq m-1$, then

$$T_i \cdot f_{\mu}(z_1, \dots, z_m) = f_{(\mu_1,\dots,\mu_{i+1},\mu_i,\dots,\mu_m)}(z_1, \dots, z_m),$$

where

$$T_i \equiv q - \frac{z_i - qz_{i+1}}{z_i - z_{i+1}}(1-\mathfrak{s}_i), \quad 1 \leq i \leq m-1,$$

with elementary transpositions $\mathfrak{s}_i$ acting by

$$\mathfrak{s}_i \cdot h(z_1, \dots, z_m) := h(z_1, \dots, z_{i+1}, z_i, \dots, z_m),$$

are the Demazure–Lusztig operators of the polynomial representation of the Hecke algebra of type A_{m-1}. The rainbow functions f_{μ} were thoroughly studied in [Borodin and Wheeler 2018] under the name of *spin nonsymmetric Hall-Littlewood functions*; they also play a central role in the present work.

For generic, not necessarily rainbow κ and κ-coloured μ, f_{μ} is defined as a suitable sum of rainbow functions. Concretely, let $\theta : \{1, \dots, m\} \to \{1, 2, \dots\}$ be the unique monotone map such that $|\theta^{-1}(j)| = \kappa_j$ for all $j \geq 1$. Then for any composition $\varkappa$ with m parts, we can define a κ-coloured composition $\theta_*(\varkappa)$ by colouring the coordinate $\varkappa_i$ by colour j if and only if $\theta(i) = j$, for all $i = 1, \dots, m$. With this, we set

$$f_{\mu}(\kappa, z_1, \dots, z_m) = \sum_{\varkappa : \theta_*(\varkappa) = \mu} f_{\varkappa}(z_1, \dots, z_m).$$

Finally, denote by $c_1 < \cdots < c_{\alpha}$ the colours of parts of μ, and denote by $\mathfrak{m}_1, \dots, \mathfrak{m}_{\alpha} \geq 1$ the number of parts of μ of colours $c_1, \dots, c_{\alpha}$, respectively ($\mathfrak{m}_j$'s are simply renumbered nonzero coordinates of κ). Set $\mathfrak{m}[a,b] = \mathfrak{m}_a + \mathfrak{m}_{a+1} + \cdots + \mathfrak{m}_b$ and recall that

$$\ell_k = \lambda_1 + \cdots + \lambda_k \quad \text{for } k \geq 1, \ \ell_0 = 0.$$

We can now state the main result of this paper.

Theorem 1.1. *With the above notation, we have*

$$\mathbb{E}\,\mathcal{O}_\mu = \frac{q^{\sum_{u\geq 1}\sum_{i>j} m_i^{(u)} m_j^{(u)}}}{\prod_{j\geq 1}(s^2;q)_{|\boldsymbol{m}^{(j)}|}} \frac{(-s)^{\mu_1+\mu_2+\dots}}{(2\pi\sqrt{-1})^m} \oint\cdots\oint_{\text{around}\{x_j^{-1}\}} \prod_{1\leq i<j\leq m} \frac{y_j-y_i}{y_j-qy_i}$$
$$\times \prod_{k=1}^{\alpha}\Biggl(\sum_{j=0}^{\mathfrak{m}_k} \frac{(-1)^j q^{\binom{\mathfrak{m}_k-j}{2}}}{(q;q)_j(q;q)_{\mathfrak{m}_k-j}} \prod_{p>\mathfrak{m}[1,k-1]}^{j+\mathfrak{m}[1,k-1]} \prod_{a>\ell_{c_{k-1}}}^{n} \frac{1-qx_ay_p}{1-x_ay_p} \prod_{r>j+\mathfrak{m}[1,k-1]}^{\mathfrak{m}[1,k]} \prod_{b>\ell_{c_k}}^{n} \frac{1-qx_by_r}{1-x_by_r}\Biggr)$$
$$\times f_\mu(\kappa; y_1^{-1},\dots,y_m^{-1}) \prod_{i=1}^{m} \frac{(y_i-s)dy_i}{y_i^2}, \tag{1-3}$$

*where (positively oriented) integration contours are chosen to encircle all points $\{x_j^{-1}\}_{j=1}^n$ and no other singularities of the integrand, or as q-**nested** closed simple curves with y_i-contour containing $q^{-1}\cdot(y_j$-contour) for all $i<j$, and all of the contours encircling $\{x_j^{-1}\}_{j=1}^n$. The contours can also be chosen to either encircle or not encircle the point* 0.

Our proof of Theorem 1.1 is deeply rooted in the formalism of spin nonsymmetric Hall-Littlewood functions f_μ developed in [Borodin and Wheeler 2018]. The key new ingredient is the idea to treat skew-Cauchy identities for these functions as averages of certain observables over the measure given by terms of the nonskew Cauchy identity. The latter can then be identified, under a certain specialization, with the sum over states of the stochastic vertex model along a horizontal line. Accessing observables via deformed Cauchy identities was previously used in [Borodin and Petrov 2017; 2018a] in the colour-blind case, but the mechanism of deformation was different, and the path to integral representations was more complex. It should be noted, however, that [Borodin and Petrov 2018a] was able to reach an integral representation for fully inhomogeneous vertex models (which was later exploited analytically in [Borodin and Petrov 2018b]). So far we have not been able to reach the same level of inhomogeneity in the coloured case, although some inhomogeneity can be added, and it is actually necessary for the limit to polymers.

Let us briefly mention some algebraic corollaries of Theorem 1.1.

First, in the colour-blind case $\lambda=(n)$, (1-3) readily leads to a formula for q-moments of the height function with a completely factorized integrand that could be viewed as a source of all the major asymptotic advances in the area. This colour-blind reduction is discussed in Section 6B.

Next, as the dependence on the values of the coordinates of μ is concentrated in the f_μ-factor and f_μ's are eigenfunctions of a transfer-matrix of our vertex model, the expectations $\mathcal{O}_\mu$ satisfy certain difference equations. Those can be seen as evidence that $\mathcal{O}_\mu(\nu)$ is actually a *duality functional* for our model; see Section 6C for details. Duality has served as a major tool for analyzing (q-)moments since [Kardar 1987]. It would be very interesting to see if our prospective duality functionals can be related to those obtained by Kuan [2018].

In the rainbow case $\lambda=(1,1,\dots,1)$, together with simplification of observables $\mathcal{O}_\mu$ to $\mathcal{O}_\mu^{\text{rainbow}}$ mentioned above, the integrand also simplifies due to the lack of j-summations. This case carries additional colour-position symmetry both in the left-hand side of (1-3) and in the right-hand side; see Section 6D. When applied to $\mathcal{O}_\mu$, this yields another set of potential duality functionals. Let us also note that for anti-dominant μ, when f_μ completely factorizes, the result can alternatively be obtained from the colour-blind case by applying a (highly nontrivial) shift-invariance property of [Borodin et al. 2019].

Finally, Theorem 1.1 allows for *stochastic fusion*: A cluster of neighbouring rows with same left-entering colours and rapidities forming a geometric progression can be collapsed to a single "fat" row whose edges are allowed to carry multiple paths. One can further analytically continue in the parameters $q^{\text{number of rows in a cluster}}$, obtaining a corresponding result in the fully fused model.[3] Details of this procedure can be found in [Borodin and Wheeler 2018; Borodin et al. 2019], and the resulting version of Theorem 1.1 is Corollary 6.9 in Section 6E.

The source of analytic corollaries of Theorem 1.1 is the fact that the coloured stochastic vertex model and its fully fused version degenerate, in various limits, to a variety of other probabilistic systems; see [Borodin and Wheeler 2018, Chapter 12] and [Borodin et al. 2019] for some of those degenerations, and the chart in the introduction to [Borodin et al. 2019] for a "big picture". In this text we only consider, in Section 7, the limit into directed random polymers that was worked out in [Borodin et al. 2019]. We obtain versions of Theorem 1.1 for random Beta-polymers (first considered in [Barraquand and Corwin 2017]), strict-weak or Gamma-polymers (first considered in [Corwin et al. 2015; O'Connell and Ortmann 2015]), O'Connell–Yor semidiscrete Brownian polymers [O'Connell and Yor 2001], and fully continuous (also known as *continuum*) Brownian polymers (equivalently, the stochastic heat equation with multiplicative noise or the KPZ equation).

Two simplifications happen in these limits. First, the presence of colours in our vertex models translates into varying starting points of the polymers, while the general definitions of the polymer models remain the same as in the colourless situation. Second, the observables simplify to pure moments of partition functions (no linear combinations necessary) in three of the four polymer models we consider. Let us illustrate what happens on the example of the continuum Brownian polymer; see Section 7E.

Let $\mathcal{Z}^{(y)}(t,x)$ be the unique solution of the following stochastic partial differential equation with the initial condition

$$\mathcal{Z}^{(y)}_t = \tfrac{1}{2}\mathcal{Z}^{(y)}_{xx} + \eta(t,x)\mathcal{Z}^{(y)}, \qquad t>0,\ x\in\mathbb{R}; \qquad \mathcal{Z}^{(y)}(0,x)=\delta(x-y),$$

where $\eta=\eta(t,x)$ is the two-dimensional white noise. See, e.g., [Quastel 2012] and references therein for an extensive literature on this equation and its close relation to continuum Brownian path integrals and the Kardar–Parisi–Zhang equation.

We are interested in evaluating the average of the product of several $\mathcal{Z}^{(y)}(t,x)$ with varying x's and y's. We will use the notation $\varkappa_j$ for different values of x, and we will group those according to which of them correspond to the same value of y. In order to do that, we will talk about a coloured real-valued vector $\varkappa=(\varkappa_1,\dots,\varkappa_m)$, with coordinates of the same colour corresponding to the same y's. Further, we will simply use the term "colour" for those y's and denote them as $s_1<\dots<s_\alpha$; they are also real-valued. Denote

$$m_i^{(x)} = \#\{j : \varkappa_j = x \text{ and has colour } s_i\}, \qquad \mathfrak{m}_i=\sum_x m_i^{(x)}, \qquad 1\le i\le\alpha,\ x\in\mathbb{R}.$$

[3] Our original setup can be viewed as a partially fused model because vertical edges are allowed to carry multiple paths; it could have been obtained from the *fundamental* model with no more than one path on any edge by clustering columns.

Then the limiting version of the fused version of Theorem 1.1 reads (see Proposition 7.9)

$$\mathbb{E}\Bigg[\prod_{(i,x)\,:\,m_i^{(x)}>0} \frac{(\mathcal{Z}^{(s_i)}(t,x))^{m_i^{(x)}}}{m_i^{(x)}!}\Bigg] = \frac{1}{(2\pi\sqrt{-1})^m}\int\cdots\int \prod_{1\le i<j\le m}\frac{w_j-w_i}{w_j-w_i-1} \times \prod_{k=1}^{\alpha}\frac{\exp\Big(-s_k\cdot\sum_{r>\mathfrak{m}[1,k-1]}^{\mathfrak{m}[1,k]} w_r\Big)}{\mathfrak{m}_k!}\cdot \mathfrak{e}_{\varkappa}(w_1,\dots,w_m)\prod_{i=1}^{m} e^{tw_i^2/2}dw_i, \tag{1-4}$$

where the integration is over upwardly oriented lines $w_i = a_i + \sqrt{-1}\cdot\mathbb{R}$ with $\Re a_j > \Re a_i + 1$ for $j > i$. The functions $\mathfrak{e}_\varkappa$ in the integrand are the limiting versions of the functions f_μ in (1-3), and they are defined as follows. In the rainbow case, in the dominant sector $\kappa_1 \ge \kappa_2 \ge \ldots \ge \kappa_m$ one has

$$\mathfrak{e}_\varkappa(w_1,\dots,w_m) = \exp(\varkappa_1 w_1 + \cdots + \varkappa_m w_m),$$

and for $\varkappa_i > \varkappa_{i+1}$ for some $1 \le i \le m-1$ one uses the exchange relations

$$\mathfrak{T}_i\cdot\mathfrak{e}_\varkappa = \mathfrak{e}_{(\varkappa_1,\dots,\varkappa_{i+1},\varkappa_i,\dots,\varkappa_m)}, \qquad \mathfrak{T}_i \equiv 1 - \frac{w_i - w_{i+1}+1}{w_i - w_{i+1}}(1-\mathfrak{s}_i), \qquad 1 \le i \le m-1,$$

to extend the definition to all rainbow vectors $\varkappa$. For a more general colouring, we use the (unique) colour-identifying monotone map θ from $\{1,\dots,m\}$ to the set of colours $\{s_i\}$, and define $\mathfrak{e}_\varkappa = \sum_{\text{rainbow}\,\varkappa' : \theta_*(\varkappa')=\varkappa}\mathfrak{e}_{\varkappa'}$.

The moment formula (1-4) has a certain shift-invariance, see Remark 7.10 below, that is partially explained by the KPZ-level degeneration of the results of [Borodin et al. 2019]. This shift-invariance is, furthermore, a corollary of the conjecture in the introduction to [Borodin et al. 2019]. It does not imply that conjecture though, because the moments are well-known to not determine the distributions of $\mathcal{Z}$'s uniquely.[4]

2. Preliminaries

The goal of this section is to summarize previously proved results that we will need later on. The notation and exposition largely follow [Borodin and Wheeler 2018].

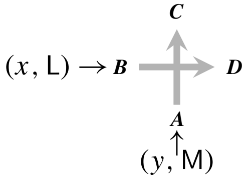
2A. ***The weights.*** The vertex models that we consider assign weights to finite collections of finite paths drawn on a square grid. Each vertex for which there exists a path that enters and exits it produces a weight that depends on the configuration of all the paths that go through this vertex. The total weight for a collection of paths is the product of weights of the vertices that the paths traverse. We tacitly assume the normalization in which the weight of an empty vertex is always equal to 1.

Our paths are going to be *coloured*, i.e., each path carries a colour that will typically be a (positive) natural number, with colour 0 reserved for the absence of a path. Our vertex weights will actually depend on the ordering of the (nonzero) paths' colours, rather than on their exact values. The paths will always

[4] That conjecture was very recently proved, by two different methods, in [Dauvergne 2020] and [Galashin 2020].

travel upward in the vertical direction, and in the horizontal direction a path can travel rightward or leftward, depending on the region of the grid it is in; this choice will always be explicitly specified.

A basic family of vertex weights that we will use is denoted as



$$= W_{\mathsf{L},\mathsf{M}}\left(\frac{x}{y}; q; \boldsymbol{B} \underset{\boldsymbol{A}}{\overset{\boldsymbol{C}}{+}} \boldsymbol{D}\right) \equiv W_{\mathsf{L},\mathsf{M}}(x/y; q; \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}), \tag{2-1}$$

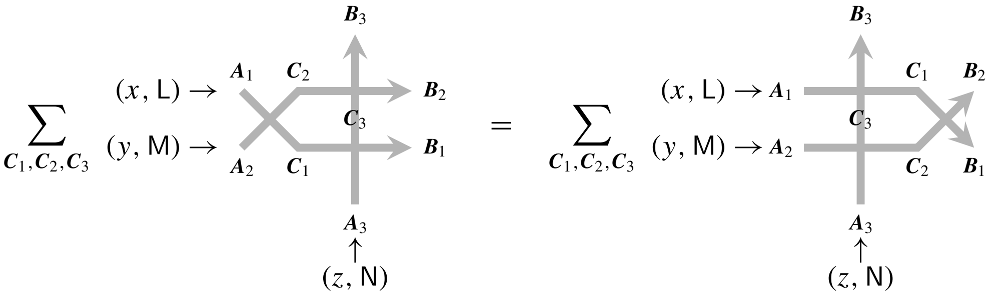
where x, y are inhomogeneity parameters (or *rapidities*) associated to the row and column, horizontal edges carry no more than L paths, vertical edges carry no more than M paths, all paths are of colours $1, \dots, n$ for some $n \geq 1$, and the specific sets of colours on the edges are encoded by compositions (or, equivalently, vectors with nonnegative entries) $\boldsymbol{A} = (A_1, \dots, A_n)$, $\boldsymbol{B} = (B_1, \dots, B_n)$, $\boldsymbol{C} = (C_1, \dots, C_n)$, $\boldsymbol{D} = (D_1, \dots, D_n)$ subject to the constraints

$$|\boldsymbol{A}|, |\boldsymbol{C}| \leq \mathsf{M}, \qquad |\boldsymbol{B}|, |\boldsymbol{D}| \leq \mathsf{L}, \tag{2-2}$$

where we use the notation $|\cdot|$ to denote the sum of all parts of a composition.

These remarkable weights come from a family of stochastic R-matrices constructed in [Kuniba et al. 2016] via symmetric tensor representations of the quantized affine algebra $U_q(\widehat{\mathfrak{sl}_{n+1}})$; see also [Kuan 2018; Bosnjak and Mangazeev 2016; Aggarwal et al. 2019; Borodin and Wheeler 2018]. Let us list some of their properties.

The *Yang–Baxter equation* for these weights can be written graphically as



(2-3)

See also [Borodin and Wheeler 2018, (C.1.2)] for the corresponding formula.

These weights enjoy a *transpositional symmetry*

$$W_{\mathsf{L},\mathsf{M}}\left(x; q; \boldsymbol{B} \underset{\boldsymbol{A}}{\overset{\boldsymbol{C}}{+}} \boldsymbol{D}\right) = W_{\mathsf{M},\mathsf{L}}\left(\frac{q^{\mathsf{M}-\mathsf{L}}}{x}; \frac{1}{q}; \boldsymbol{A} \underset{\boldsymbol{B}}{\overset{\boldsymbol{D}}{+}} \boldsymbol{C}\right). \tag{2-4}$$

They are *stochastic* in the following sense:

$$\sum_{\boldsymbol{C},\boldsymbol{D}} W_{\mathsf{L},\mathsf{M}}\left(x; q; \boldsymbol{B} \underset{\boldsymbol{A}}{\overset{\boldsymbol{C}}{+}} \boldsymbol{D}\right) = 1, \tag{2-5}$$

where the sum is over all compositions $\boldsymbol{C} = (C_1, \dots, C_n)$ and $\boldsymbol{D} = (D_1, \dots, D_n)$ with $|\boldsymbol{C}| \leq \mathsf{M}$, $|\boldsymbol{D}| \leq \mathsf{L}$.

They can also be given by the following explicit formula originating from [Bosnjak and Mangazeev 2016]; see also [Borodin and Wheeler 2018, Theorem C.1.1]:

$$W_{\mathsf{L},\mathsf{M}}\left(x;q;\;\begin{array}{ccc} & \boldsymbol{C} & \\ & \uparrow & \\ \boldsymbol{B} & \rightarrow & \boldsymbol{D} \\ & | & \\ & \boldsymbol{A} & \end{array}\right) = (\mathbf{1}_{\boldsymbol{A}+\boldsymbol{B}=\boldsymbol{C}+\boldsymbol{D}})\, x^{|\boldsymbol{D}|-|\boldsymbol{B}|} q^{|\boldsymbol{A}|\mathsf{L}-|\boldsymbol{D}|\mathsf{M}} \\ \times \sum_{\boldsymbol{P}} \Phi(\boldsymbol{C}-\boldsymbol{P}, \boldsymbol{C}+\boldsymbol{D}-\boldsymbol{P}; q^{\mathsf{L}-\mathsf{M}}x, q^{-\mathsf{M}}x)\,\Phi(\boldsymbol{P},\boldsymbol{B}; q^{-\mathsf{L}}/x, q^{-\mathsf{L}}), \qquad (2\text{-}6)$$

where the sum is over compositions $\boldsymbol{P} = (P_1, \dots, P_n)$ such that $0 \le P_i \le \min(B_i, C_i)$ for all $1 \le i \le n$; and for any two compositions $\lambda, \mu \in \mathbb{N}^n$ such that $\lambda_i \le \mu_i$ for all $1 \le i \le n$, we used the notation

$$\Phi(\lambda,\mu;x,y) := \frac{(x;q)_{|\lambda|}(y/x;q)_{|\mu-\lambda|}}{(y;q)_{|\mu|}}(y/x)^{|\lambda|}\, q^{\sum_{i<j}(\mu_i-\lambda_i)\lambda_j} \prod_{i=1}^{n} \binom{\mu_i}{\lambda_i}_q .$$

The substitution of $\mathsf{L} = \mathsf{M} = 1$ into $W_{\mathsf{L},\mathsf{M}}$ returns the (stochastic version of) the fundamental R-matrix for $U_q(\widehat{\mathfrak{sl}_{n+1}})$:

$$W_{1,1}\left(\frac{x}{y};q;\;\begin{array}{ccc} & \boldsymbol{C} & \\ & \uparrow & \\ \boldsymbol{B} & \rightarrow & \boldsymbol{D} \\ & | & \\ & \boldsymbol{A} & \end{array}\right) = \begin{cases} R_{y/x}(\boldsymbol{A}^*, \boldsymbol{B}^*; \boldsymbol{C}^*, \boldsymbol{D}^*), & |\boldsymbol{A}|, |\boldsymbol{B}|, |\boldsymbol{C}|, |\boldsymbol{D}| \le 1, \\ 0, & \text{otherwise}, \end{cases} \qquad (2\text{-}7)$$

where we have defined

$$\boldsymbol{I}^* = \begin{cases} 0, & \boldsymbol{I} = \boldsymbol{0}, \\ i, & \boldsymbol{I} = \boldsymbol{e}_i \quad (i\text{-th standard basis vector}), \end{cases}$$

for any composition $\boldsymbol{I} = (I_1, \dots, I_n)$ such that $|\boldsymbol{I}| \le 1$, and where $R_{y/x}$ denotes the fundamental R-matrix depicted as (with $z = y/x$)

$$R_z(i,j;k,\ell) = \begin{array}{ccc} & k & \\ & \uparrow & \\ j & \longrightarrow & \ell \\ & | & \\ & i & \end{array}, \qquad i,j,k,\ell \in \{0,1,\dots,n\}, \qquad (2\text{-}8)$$

and with matrix elements summarized in Table 2, in which we assume that $0 \le i < j \le n$. Observe that these weights are manifestly stochastic.

The general weights $W_{\mathsf{L},\mathsf{M}}$ can be reconstructed from the fundamental ones in Table 2 via the procedure of *stochastic fusion*. To state how it works we need a bit of notation.

Let $\mathsf{N} \ge 1$, and consider a vector of nonnegative integers $(i_1, \dots, i_{\mathsf{N}}) \in \{0, 1, \dots, n\}^M$. From this we define another vector,

$$\mathcal{C}(i_1, \dots, i_{\mathsf{N}}) := (I_1, \dots, I_n), \qquad I_a = \#\{k : i_k = a\}, \qquad 1 \le a \le n, \qquad (2\text{-}9)$$

which keeps track of the multiplicity of each colour $1 \le a \le n$ within $(i_1, \dots, i_{\mathsf{N}})$. Set

$$\operatorname{inv}(i_1, \dots, i_{\mathsf{N}}) = \#\{1 \le a < b \le \mathsf{N} : i_a > i_b\}, \qquad \widetilde{\operatorname{inv}}(i_1, \dots, i_{\mathsf{N}}) = \#\{1 \le a < b \le \mathsf{N} : i_a < i_b\},$$

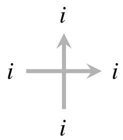
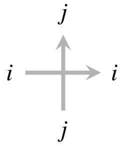
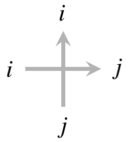
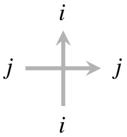
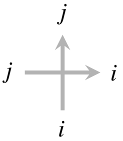
 1	 $\dfrac{q(1-z)}{1-qz}$	 $\dfrac{1-q}{1-qz}$
	 $\dfrac{1-z}{1-qz}$	 $\dfrac{(1-q)z}{1-qz}$

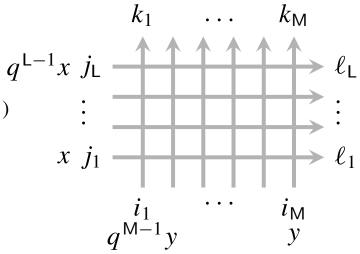
Table 2. The matrix elements of the fundamental R-matrix.

and, denoting $I_0 := \mathsf{N} - \sum_{a=1}^{n} I_a$,

$$Z_q(\mathsf{N};\boldsymbol{I}) = \sum_{\mathcal{C}(i_1,\ldots,i_\mathsf{N})=\boldsymbol{I}} q^{\mathrm{inv}(i_1,\ldots,i_\mathsf{N})} = \sum_{\mathcal{C}(j_1,\ldots,j_\mathsf{N})=\boldsymbol{I}} q^{\widetilde{\mathrm{inv}}(j_1,\ldots,j_\mathsf{N})} = \frac{(q;q)_\mathsf{N}}{(q;q)_{I_0}(q;q)_{I_1}\ldots(q;q)_{I_n}}.$$

Then (see [Borodin et al. 2019, Appendix]),

$$W_{\mathsf{L},\mathsf{M}}\left(\frac{x}{y};q;\ \boldsymbol{B}\underset{\boldsymbol{A}}{\overset{\boldsymbol{C}}{+}}\boldsymbol{D}\right) = \frac{1}{Z_q(\mathsf{M};\boldsymbol{A})Z_q(\mathsf{L};\boldsymbol{B})} \sum_{\substack{\mathcal{C}(j_1,\ldots,j_\mathsf{L})=\boldsymbol{B}\\ \mathcal{C}(\ell_1,\ldots,\ell_\mathsf{L})=\boldsymbol{D}}} q^{\widetilde{\mathrm{inv}}(j_1,\ldots,j_\mathsf{L})} \sum_{\substack{\mathcal{C}(i_1,\ldots,i_\mathsf{M})=\boldsymbol{A}\\ \mathcal{C}(k_1,\ldots,k_\mathsf{M})=\boldsymbol{C}}} q^{\mathrm{inv}(i_1,\ldots,i_\mathsf{M})} \tag{2-10}$$



where the figure on the right denotes the corresponding partition function with R-weights given by (2-8) and Table 2 (thus, summation over all possible states of interior edges is assumed), and the colours on the boundary edges, as well as row and column rapidities, are explicitly indicated.[5]

A key feature of the fused R-vertices that allows stacking them together is their *q-exchangeability*: In equation (2-10), the sum over $(k_1,\ldots,k_\mathsf{M})$ can be omitted at the expense of adding the factor of $q^{-\mathrm{inv}(k_1,\ldots,k_\mathsf{M})}Z_q(\mathsf{M};\boldsymbol{C})$ to the summands, and, independently, the sum over $(\ell_1,\ldots,\ell_\mathsf{L})$ can be removed at the expense of adding the factor of $q^{-\widetilde{\mathrm{inv}}(\ell_1,\ldots,\ell_\mathsf{L})}Z_q(\mathsf{L};\boldsymbol{D})$; see [Borodin and Wheeler 2018, Proposition B.2.2] for a proof.

[5] As in (2-8), the spectral parameter of an R-vertex is assumed to be equal to the ratio of the column rapidity and the row rapidity.

In what follows we will also need partially fused weights defined as (see (2-4))

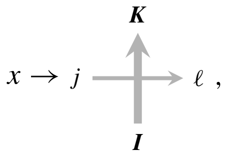
$$L_x^{\mathrm{stoch}}(\boldsymbol{A},b;\boldsymbol{C},d) := W_{1,\mathsf{M}}\left(\frac{x}{s};q;\begin{array}{c}\boldsymbol{C}\\ e_b \overset{\uparrow}{\longrightarrow} e_d\\ \boldsymbol{A}\end{array}\right)\Bigg|_{q^{\mathsf{M}}\to s^{-2}}, \tag{2-11}$$

$$M_x^{\mathrm{stoch}}(\boldsymbol{A},b;\boldsymbol{C},d) := W_{1,\mathsf{M}}\left(\frac{s}{x};\frac{1}{q};\begin{array}{c}\boldsymbol{C}\\ e_b \overset{\uparrow}{\longrightarrow} e_d\\ \boldsymbol{A}\end{array}\right)\Bigg|_{q^{\mathsf{M}}\to s^{-2}}, \tag{2-12}$$

where the substitution of a generic complex parameter s^2 for $q^{-\mathsf{M}}$ is based on the fact that the right-hand sides are rational in q^{M}. The weights L_x^{stoch} are tabulated in Table 1.

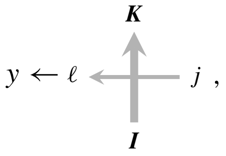
In addition, we will use gauge transformed (nonstochastic) versions

$$(-s)^{-\mathbf{1}_{\ell\geq 1}} L_x^{\mathrm{stoch}}(\boldsymbol{I},j;\boldsymbol{K},\ell) =: L_x(\boldsymbol{I},j;\boldsymbol{K},\ell) =$$



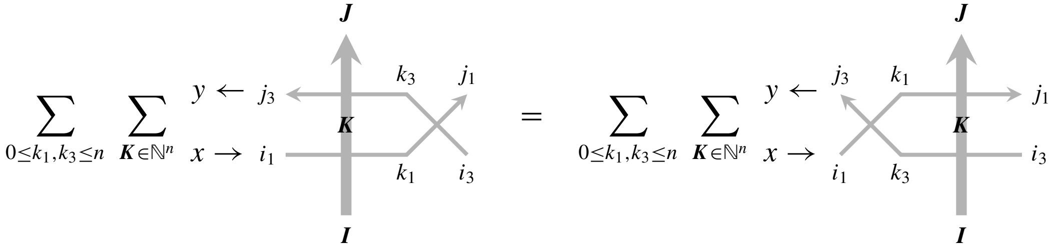
(2-13)

$$(-s)^{\mathbf{1}_{j\geq 1}} M_x^{\mathrm{stoch}}(\boldsymbol{I},j;\boldsymbol{K},\ell) =: M_x(\boldsymbol{I},j;\boldsymbol{K},\ell) =$$



(2-14)

that, as a consequence of (2-3), satisfy the following version of the Yang–Baxter equation; see [Borodin and Wheeler 2018, (2.3.5)]:


$\sum_{0\le k_1,k_3\le n} \sum_{\boldsymbol{K}\in\mathbb{N}^n}$
J
y ← j_3
k_3
j_1
K
x → i_1
k_1
i_3
I
=
$\sum_{0\le k_1,k_3\le n} \sum_{\boldsymbol{K}\in\mathbb{N}^n}$
J
y ←
j_3
k_1
j_1
K
x →
i_1
k_3
i_3
I

(2-15)

with the spectral parameter of the R-vertex equal to $(qxy)^{-1}$.

The explicit values of the weights (2-13) are summarized in Table 3 where we assume that $1 \le i < j \le n$, and the notation $\boldsymbol{I}_{[k,n]}$ stands for $\sum_{a=k}^{n} I_a$. For $n = 1$, these weights correspond to the image of the universal R-matrix for the quantum affine group $U_q(\widehat{\mathfrak{sl}_2})$ in the tensor product of its vector representation (horizontal edges) and a Verma module (vertical edges), with parameter s encoding its highest weight; we call s the *spin parameter*.

2B. ***The q-Hahn specialization and a limit relation.*** The complicated expression (2-6) can simplify at special values of parameters; we have already seen this in the case of L^{stoch} and M^{stoch}. Another such specialization that allows L and M to remain generic is the following (see [Kuniba et al. 2016, Proposition 7]

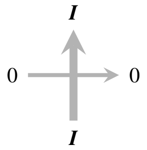
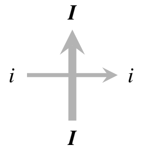
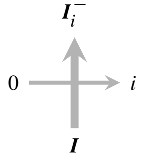
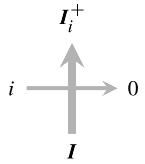
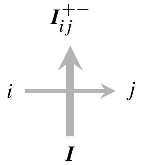
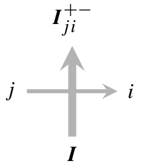
 $\dfrac{1-sxq^{\boldsymbol{I}_{[1,n]}}}{1-sx}$	 $\dfrac{(x-sq^{I_i})q^{\boldsymbol{I}_{[i+1,n]}}}{1-sx}$	 $\dfrac{x(1-q^{I_i})q^{\boldsymbol{I}_{[i+1,n]}}}{1-sx}$
 $\dfrac{1-s^2q^{\boldsymbol{I}_{[1,n]}}}{1-sx}$	 $\boldsymbol{I}_{ij}^{+-}$ i j $\boldsymbol{I}$ $\dfrac{x(1-q^{I_j})q^{\boldsymbol{I}_{[j+1,n]}}}{1-sx}$	 $\boldsymbol{I}_{ji}^{+-}$ j i $\boldsymbol{I}$ $\dfrac{s(1-q^{I_i})q^{\boldsymbol{I}_{[i+1,n]}}}{1-sx}$

Table 3. The explicit values of the weights (2-13).

and also [Bosnjak and Mangazeev 2016, (7.13)]): Assuming that $\mathsf{L} \le \mathsf{M}$, we have

$$\begin{aligned} W_{\mathsf{L},\mathsf{M}}\left(1;q;\ \boldsymbol{B}\overset{\boldsymbol{C}}{\underset{\boldsymbol{A}}{\dashrightarrow}}\boldsymbol{D}\right) &= q^{(\mathsf{L}-\mathsf{M})|\boldsymbol{D}|}\,\frac{(q^{\mathsf{L}-\mathsf{M}};q)_{|\boldsymbol{A}|-|\boldsymbol{D}|}(q^{-\mathsf{L}};q)_{|\boldsymbol{D}|}}{(q^{-\mathsf{M}};q)_{|\boldsymbol{A}|}}\, q^{\sum_{i<j} D_i(A_j-D_j)} \prod_{i=1}^{n} \binom{A_i}{A_i-D_i}_q \\ &= \Phi(\boldsymbol{D},\boldsymbol{A};q^{-\mathsf{L}},q^{-\mathsf{M}}). \end{aligned} \tag{2-16}$$

The proof follows from the fact that setting $x=1$ restricts the sum in (2-6) to a single term with $\boldsymbol{P}=\boldsymbol{B}$. This specialization is often referred to as the *q-Hahn point* because for $n=1$ it reproduces the orthogonality weights for the classical q-Hahn orthogonal polynomials.

Note that replacing $q^{-\mathsf{M}}$ by a generic complex parameter s^2 as was done for $W_{1,\mathsf{M}}$ in (2-11) can also be performed for $W_{\mathsf{L},\mathsf{M}}$ for $\mathsf{L}>1$ either in (2-6) or in (2-16), because those are weights are still manifestly rational in q^{M}. Alternatively, the result of such a replacement could be seen as a stochastic fusion of the weights L^{stoch} in the spirit of (2-10), where only the outer sum over j_*'s and ℓ_*'s is present.

We will also need the following limiting relation for the weights (2-6); see [Borodin et al. 2019, Lemma 6.8]:

Assume that $\boldsymbol{B}=(0,\ldots,0,\mathsf{L})$. Then

$$\lim_{q^{-\mathsf{M}}=s^2\to 0} W_{\mathsf{L},\mathsf{M}}\left(zq^{\mathsf{M}};q;\ \boldsymbol{B}\overset{\boldsymbol{C}}{\underset{\boldsymbol{A}}{\dashrightarrow}}\boldsymbol{D}\right) = \begin{cases} \dfrac{(zq^{\mathsf{L}};q)_\infty}{(z;q)_\infty}\dfrac{(q^{-\mathsf{L}};q)_d}{(q;q)_d}(zq^{\mathsf{L}})^d & \text{if } \boldsymbol{D}=(0,\ldots,0,d),\ d\ge 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2-17}$$

2C. *Row operators and rational functions.* Let V be an infinite-dimensional vector space obtained by taking the linear span of all n-tuples of nonnegative integers:

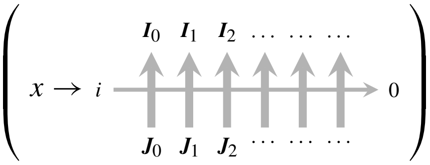
$$V=\operatorname{Span}\{|\boldsymbol{I}\rangle\}=\operatorname{Span}\{|i_1,\ldots,i_n\rangle\}_{i_1,\ldots,i_n\in\mathbb{N}}. \tag{2-18}$$

It is convenient to consider an infinite tensor product of such spaces,

$$\mathbb{V} = V_0 \otimes V_1 \otimes V_2 \otimes \cdots,$$

where each V_i denotes a copy of V. Let $\bigotimes_{k=0}^{\infty} |\boldsymbol{I}_k\rangle_k$ be a finite state in $\mathbb{V}$, i.e., assume that there exists $N \in \mathbb{N}$ such that $\boldsymbol{I}_k = \boldsymbol{0}$ for all $k > N$; in what follows only such states are considered. We define two families of linear operators acting on the finite states:

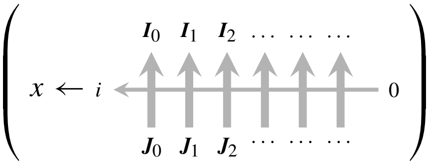
$$\mathcal{C}_i(x) : \bigotimes_{k=0}^{\infty} |\boldsymbol{I}_k\rangle_k \mapsto \sum_{\boldsymbol{J}_0,\boldsymbol{J}_1,\ldots\in\mathbb{N}^n}$$



$\boldsymbol{I}_0$ $\boldsymbol{I}_1$ $\boldsymbol{I}_2$ ⋯ ⋯ ⋯
$x \rightarrow i$
0
$\boldsymbol{J}_0$ $\boldsymbol{J}_1$ $\boldsymbol{J}_2$ ⋯ ⋯ ⋯

$$\bigotimes_{k=0}^{\infty} |\boldsymbol{J}_k\rangle_k, \quad 0 \le i \le n, \tag{2-19}$$

$$\mathcal{B}_i(x) : \bigotimes_{k=0}^{\infty} |\boldsymbol{I}_k\rangle_k \mapsto \sum_{\boldsymbol{J}_0,\boldsymbol{J}_1,\ldots\in\mathbb{N}^n}$$



$\boldsymbol{I}_0$ $\boldsymbol{I}_1$ $\boldsymbol{I}_2$ ⋯ ⋯ ⋯
$x \leftarrow i$
0
$\boldsymbol{J}_0$ $\boldsymbol{J}_1$ $\boldsymbol{J}_2$ ⋯ ⋯ ⋯

$$\bigotimes_{k=0}^{\infty} |\boldsymbol{J}_k\rangle_k, \quad 0 \le i \le n, \tag{2-20}$$

where the vertex weights are those from (2-13) and (2-14), respectively. Note that the sums above are always finite due to path conservation, and the infinite number of empty vertices far to the right all have weight 1.

A direct corollary of the Yang–Baxter equation (2-15) is the fact that the row operators $\mathcal{B}_i$ and $\mathcal{C}_i$ satisfy certain explicit quadratic commutation relations; see [Borodin and Wheeler 2018, Section 3.2]. The simplest ones state that for a fixed i, $0 \le i \le n$, the operators $\mathcal{B}_i(x)$ commute between themselves for different values of x, and, similarly, $\mathcal{C}_i(x)$ also commute for different values of x. Slightly more complicated are the following relations between $\mathcal{B}$- and $\mathcal{C}$-operators; see [Borodin and Wheeler 2018, Theorem 3.2.3]:

Fix two nonnegative integers i, j such that $0 \le i, j \le n$, and complex parameters x, y such that

$$\left| \frac{x-s}{1-sx} \cdot \frac{y-s}{1-sy} \right| < 1. \tag{2-21}$$

Then the row operators (2-19) and (2-20) obey the following commutation relations ($0 \le i, j \le n$):

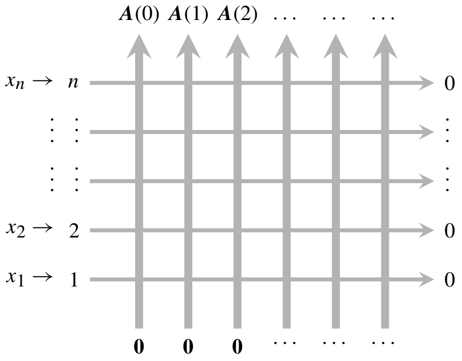
$$\begin{gathered}
\mathcal{C}_i(x)\mathcal{B}_j(y) = \frac{1-qxy}{1-xy}\mathcal{B}_j(y)\mathcal{C}_i(x), \quad i<j, \qquad q\,\mathcal{C}_i(x)\mathcal{B}_j(y) = \frac{1-qxy}{1-xy}\mathcal{B}_j(y)\mathcal{C}_i(x), \quad i>j, \\
\mathcal{C}_i(x)\mathcal{B}_i(y) = \frac{1-q^{-1}}{1-xy}\sum_{k<i}\mathcal{B}_k(y)\mathcal{C}_k(x) + \mathcal{B}_i(y)\mathcal{C}_i(x) - \frac{(1-q)xy}{1-xy}\sum_{k>i}\mathcal{B}_k(y)\mathcal{C}_k(x).
\end{gathered} \tag{2-22}$$

Note that matrix elements of the left-hand sides in (2-22) are given by infinite sums, and (2-21) ensures that those sums converge. It is thanks to the infinite range of the lattice in (2-19) and (2-20) that relations (2-22) are simpler than the usual commutation relations in the Yang–Baxter algebra.

We are now in position to introduce certain rational functions that will play a central role in what follows; they were called *nonsymmetric spin Hall–Littlewood functions* in [Borodin and Wheeler 2018].

For a composition μ and n complex parameters $x_1, \ldots, x_n$, define a rational function $f_\mu(x_1, \ldots, x_n)$ as a partition function depicted below (vertex weights (2-13) are being used):

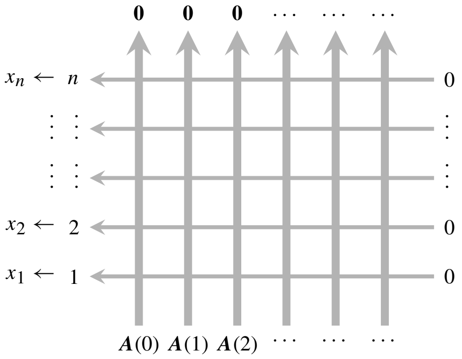
$$f_\mu(x_1, \ldots, x_n) = \tag{2-23}$$


$\boldsymbol{A}(0)$ $\boldsymbol{A}(1)$ $\boldsymbol{A}(2)$ ⋯ ⋯ ⋯
$x_n \to$ n
$x_2 \to$ 2
$x_1 \to$ 1
0
0
0
0 **0** **0** ⋯ ⋯ ⋯

with $\boldsymbol{A}(k) = \sum_{j=1}^{n} \mathbf{1}_{\mu_j = k} \boldsymbol{e}_j$. These are certain matrix elements of the operator $\mathcal{C}_1(x_1) \cdots \mathcal{C}_n(x_n)$ that can be symbolically written in the form $\langle \varnothing | \mathcal{C}_1(x_1) \cdots \mathcal{C}_n(x_n) | \mu \rangle$. A more general definition of the f_μ's, as described in [Borodin and Wheeler 2018, Section 3.4], involves some of the operators $\mathcal{C}_i$ being repeated with different arguments (equivalently, some of the paths entering the partition function (2-23) from the left being of the same colour); we will meet such functions in Section 2H.

Similarly to the f_μ's, one defines dual functions

$$g_\mu(x_1, \ldots, x_n) = \tag{2-24}$$



as matrix elements $\langle \mu | \mathcal{B}_1(x_1) \cdots \mathcal{B}_n(x_n) | \varnothing \rangle$ of $\mathcal{B}_1(x_1) \cdots \mathcal{B}_n(x_n)$. One proves (see [Borodin and Wheeler 2018, Proposition 5.6.1]) that these two families of functions are closely related:

$$g_{\tilde{\mu}}(x_n^{-1}, \ldots, x_1^{-1}; q^{-1}, s^{-1}) = c_\mu(q, s) \prod_{i=1}^{n} x_i \cdot f_\mu(x_1, \ldots, x_n; q, s), \qquad \tilde{\mu} = (\mu_n, \ldots, \mu_1), \tag{2-25}$$

where the multiplicative constant $c_\mu(q, s)$ is given by

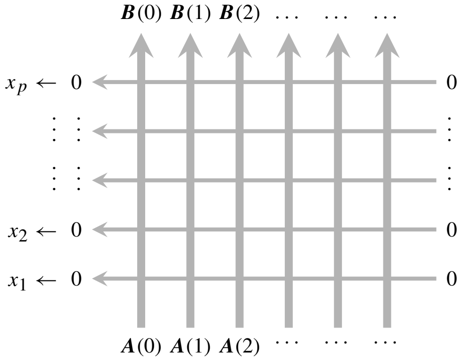
$$c_\mu(q, s) = \frac{s^n (q-1)^n q^{\#\{i<j : \mu_i \le \mu_j\}}}{\prod_{j \ge 0} (s^2; q)_{m_j(\mu)}}, \qquad m_j(\mu) := \#\{1 \le k \le n : \mu_k = j\}, \quad j \ge 0. \tag{2-26}$$

It will be convenient for us to use a slightly different normalization of the dual functions:

$$g^*_\mu(x_1, \ldots, x_n) := q^{n(n+1)/2} (q-1)^{-n} \cdot g_\mu(x_1, \ldots, x_n).$$

Finally, let us introduce a third family of symmetric rational functions parametrized by a pair of compositions μ, ν, or rather by a skew composition μ/ν, by

$$G_{\mu/\nu}(x_1,\dots,x_p) = \tag{2-27}$$


$\boldsymbol{B}(0)$ $\boldsymbol{B}(1)$ $\boldsymbol{B}(2)$ … … …
$x_p \leftarrow$ 0
0
$x_2 \leftarrow$ 0
0
$x_1 \leftarrow$ 0
0
$\boldsymbol{A}(0)$ $\boldsymbol{A}(1)$ $\boldsymbol{A}(2)$ ⋯ ⋯ ⋯

(where $p = 1, 2, \dots$ is arbitrary) with $\boldsymbol{A}(k) = \sum_{j=1}^{n} \mathbf{1}_{\mu_j = k} \boldsymbol{e}_j$, $\boldsymbol{B}(k) = \sum_{j=1}^{n} \mathbf{1}_{\nu_j = k} \boldsymbol{e}_j$, for all $k \in \mathbb{Z}_{\geq 0}$. These are matrix elements $\langle \mu | \mathcal{B}_0(x_1) \cdots \mathcal{B}_0(x_p) | \nu \rangle$ of the operator $\mathcal{B}_0(x_1) \cdots \mathcal{B}_0(x_p)$, and their symmetry with respect to the x_i's is a direct consequence of the commutativity of $\mathcal{B}_0(x_i)$'s noted after (2-20).

2D. ***Colour-blindness.*** For any integer $k \in \{0, 1, \dots, n\}$ define its colour-blind projection

$$\theta(k) := \mathbf{1}_{k \geq 1} = \begin{cases} 0, & k = 0, \\ 1, & k \geq 1. \end{cases}$$

Also, denote the set of compositions of a fixed length (number of entries) with a given weight (sum of entries) k as

$$\mathcal{W}(k) := \{\boldsymbol{K} \in \mathbb{Z}^n_{\geq 0} : |\boldsymbol{K}| = k\}.$$

Then one proves (see [Borodin and Wheeler 2018, Proposition 2.4.2]) that

$$\sum_{\boldsymbol{K} \in \mathcal{W}(k)} L_x(\boldsymbol{I}, j; \boldsymbol{K}, 0) = L_x^{(1)}(|\boldsymbol{I}|, \theta(j); k, 0), \qquad \sum_{\boldsymbol{K} \in \mathcal{W}(k)} \sum_{1 \leq \ell \leq n} L_x(\boldsymbol{I}, j; \boldsymbol{K}, \ell) = L_x^{(1)}(|\boldsymbol{I}|, \theta(j); k, 1), \tag{2-28}$$

where $L^{(1)}$ refers to the weights (2-13) with $n = 1$. Similarly, one has

$$\sum_{\boldsymbol{K} \in \mathcal{W}(k)} M_x(\boldsymbol{I}, j; \boldsymbol{K}, 0) = M_x^{(1)}(|\boldsymbol{I}|, \theta(j); k, 0), \qquad \sum_{\boldsymbol{K} \in \mathcal{W}(k)} \sum_{1 \leq \ell \leq n} M_x(\boldsymbol{I}, j; \boldsymbol{K}, \ell) = M_x^{(1)}(|\boldsymbol{I}|, \theta(j); k, 1).$$

This has been observed in a number of earlier publications [Foda and Wheeler 2013; Garbali et al. 2017; Kuan 2018], and used to different effects within those works. This property allows certain linear combinations of higher-rank partition functions to be computable as rank-1, or *colour-blind* partition functions.

As an example, one derives the symmetrization identities (see [Borodin and Wheeler 2018, Propositions 3.4.4 and 4.4.3])

$$\sum_{\mu : \mu^+ = \nu} f_\mu(x_1, \dots, x_n) = \mathsf{F}^{\mathsf{c}}_\nu(x_1, \dots, x_n), \qquad \sum_{\mu : \mu^+ = \kappa} G_{\nu/\mu}(x_1, \dots, x_p) = q^{-np} \mathsf{G}_{\nu^+/\kappa}(x_1, \dots, x_p), \tag{2-29}$$

with the sums taken over all compositions with a given dominant reordering denoted by the superscript "+", and where the symmetric rational functions F^{c} and G are colour-blind objects that were considered at length in [Borodin 2017; Borodin and Petrov 2017; 2018a].

2E. ***Recursive relations.*** While explicit formulas representing the functions f_μ as sums of monomials are rather involved (see Chapters 6 and 7 of [Borodin and Wheeler 2018] for two different versions) there exist concise recursive relations for them.

First, for *anti-dominant* compositions $\delta = (\delta_1 \le \cdots \le \delta_n)$, the functions f_δ are completely factorized (see [Borodin and Wheeler 2018, Proposition 5.1.1]):

$$f_\delta(x_1, \dots, x_n) = \frac{\prod_{j \ge 0}(s^2; q)_{m_j(\delta)}}{\prod_{i=1}^n (1 - s x_i)} \prod_{i=1}^n \left(\frac{x_i - s}{1 - s x_i} \right)^{\delta_i}, \qquad m_j(\delta) = \#\{1 \le k \le n : \delta_k = j\},\ j \ge 0. \quad (2\text{-}30)$$

This happens because the partition function (2-23) has only one configuration of paths that contributes nontrivially.

Second, the following recursion allows one to move step by step from antidominant compositions to the dominant ones; see [Borodin and Wheeler 2018, Theorem 5.3.1]:

Let $\mu = (\mu_1, \dots, \mu_n)$ be a composition with $\mu_i < \mu_{i+1}$ for some $1 \le i \le n-1$. Then

$$T_i \cdot f_\mu(x_1, \dots, x_n) = f_{(\mu_1, \dots, \mu_{i+1}, \mu_i, \dots, \mu_n)}(x_1, \dots, x_n), \quad (2\text{-}31)$$

where

$$T_i \equiv q - \frac{x_i - q x_{i+1}}{x_i - x_{i+1}}(1 - \mathfrak{s}_i), \qquad 1 \le i \le n-1, \quad (2\text{-}32)$$

with elementary transpositions $\mathfrak{s}_i \cdot h(x_1, \dots, x_n) := h(x_1, \dots, x_{i+1}, x_i, \dots, x_n)$, are the Demazure–Lusztig operators of the polynomial representation of the Hecke algebra of type A_{n-1}.

2F. ***Summation identities.*** The functions f_μ, g_μ, and $G_{\mu/\nu}$ satisfy several summation identities that can be found in [Borodin and Wheeler 2018, Section 4]; in what follows we will need a couple of them. The first one bears a certain similarity to a summation identity proved by Sahi [1996] and Mimachi and Noumi [1998] for nonsymmetric Macdonald polynomials [Mimachi and Noumi 1998]; it was proved as [Borodin and Wheeler 2018, Theorem 4.3.1].

Let $(x_1, \dots, x_n)$ and $(y_1, \dots, y_n)$ be two sets of complex parameters such that (see (2-21))

$$\left| \frac{x_i - s}{1 - s x_i} \cdot \frac{y_j - s}{1 - s y_j} \right| < 1 \qquad \text{for all } 1 \le i, j \le n. \quad (2\text{-}33)$$

Then

$$\sum_\mu f_\mu(x_1, \dots, x_n) g^*_\mu(y_1, \dots, y_n) = \prod_{i=1}^n \frac{1}{1 - x_i y_i} \prod_{n \ge i > j \ge 1} \frac{1 - q x_i y_j}{1 - x_i y_j}, \quad (2\text{-}34)$$

where the summation is over all compositions μ (with nonnegative coordinates).

This identity is proved by evaluating the matrix element $\langle \varnothing | \mathcal{C}_1(x_1) \cdots \mathcal{C}_n(x_n) \mathcal{B}_1(y_1) \cdots \mathcal{B}_n(y_n) | \varnothing \rangle$ in two different ways, by inserting the partition of unity $\sum_\mu |\mu\rangle\langle\mu|$ between the groups of $\mathcal{B}$- and $\mathcal{C}$-operators, and by using the commutation relations (2-22).

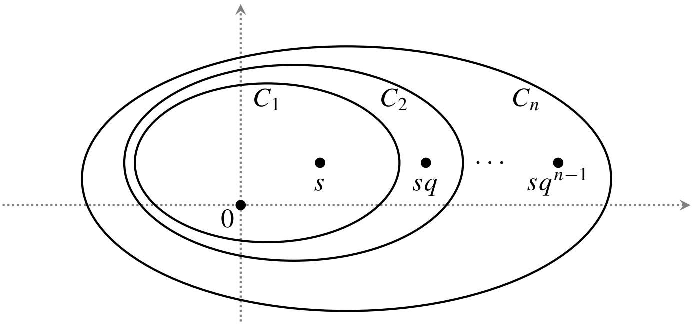


Figure 2. Admissible contours $\{C_1, \ldots, C_n\}$ with respect to (q, s).

A similar argument applied to the matrix element $\langle \varnothing | \mathcal{C}_1(x_1) \ldots \mathcal{C}_n(x_n) \mathcal{B}_0(y_1) \ldots \mathcal{B}_0(y_p) | \nu \rangle$ leads to the following summation identity that is reminiscent of the skew-Cauchy identities known in the theory of symmetric functions (see [Borodin and Wheeler 2018, Proposition 4.5.1] for a detailed proof):

Let $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_p)$ be two sets of complex parameters satisfying the constraints (2-33), and fix a composition $\nu = (\nu_1, \ldots, \nu_n)$. Then one has the identity

$$\sum_{\mu} f_{\mu}(x_1, \ldots, x_n) G_{\mu/\nu}(y_1, \ldots, y_p) = \prod_{i=1}^{n} \prod_{j=1}^{p} \frac{1 - q x_i y_j}{q(1 - x_i y_j)} \cdot f_{\nu}(x_1, \ldots, x_n), \tag{2-35}$$

where the summation is taken over all length-n compositions $\mu = (\mu_1, \ldots, \mu_n)$.

2G. ***Orthogonality and integral representations.*** Let $\{C_1, \ldots, C_n\}$ be a collection of contours in the complex plane. We say that the set $\{C_1, \ldots, C_n\}$ is admissible with respect to a pair of complex parameters (q, s) if the following conditions are met:

- The contours $\{C_1, \ldots, C_n\}$ are closed, positively oriented and pairwise nonintersecting.
- The contours C_i and $q \cdot C_i$ are both contained within contour C_{i+1} for all $1 \le i \le n-1$, where $q \cdot C_i$ denotes the image of C_i under multiplication by q.
- All contours surround the point s.

An illustration of such admissible contours is given in Figure 2.

Often when we integrate rational functions over $\{C_1, \ldots, C_n\}$, the integrals can also be computed as sums of residues of the integrand inside the contours. Such sums also make sense for values of parameters that prevent admissible contours from existing, and thus the integrals could also be defined via the residue sums. Therefore, we will tacitly assume that we perform such a replacement should the admissible contours not exist, and we will also use a similar convention for other contour integrals below; see Remark 5.3.

Let $\mu = (\mu_1, \ldots, \mu_n)$ and $\nu = (\nu_1, \ldots, \nu_n)$ be two compositions. We have the following orthonormality of nonsymmetric spin Hall–Littlewood functions (see [Borodin and Wheeler 2018, Theorem 8.2.1]):

$$\frac{1}{(2\pi\sqrt{-1})^n} \oint_{C_1} \frac{dx_1}{x_1} \cdots \oint_{C_n} \frac{dx_n}{x_n} \prod_{1 \le i < j \le n} \frac{x_j - x_i}{x_j - q x_i} f_{\mu}(x_1^{-1}, \ldots, x_n^{-1}) g_{\nu}^{*}(x_1, \ldots, x_n) = \mathbf{1}_{\mu=\nu}. \tag{2-36}$$

Coupled with (2-34) (which is easily shown to converge uniformly provided that the left-hand sides of (2-33) are bounded by a uniform constant < 1), this leads to the integral formula

$$f_\mu(x_1,\dots,x_n)=\frac{1}{(2\pi\sqrt{-1})^n}\oint_{C_1}\frac{dy_1}{y_1}\cdots\oint_{C_n}\frac{dy_n}{y_n}\prod_{1\le i<j\le n}\frac{y_j-y_i}{y_j-qy_i}\,f_\mu(y_1^{-1},\dots,y_n^{-1}) \\ \times\prod_{i=1}^{n}\frac{1}{1-x_iy_i}\prod_{n\ge i>j\ge 1}\frac{1-qx_iy_j}{1-x_iy_j}. \tag{2-37}$$

A similar integral representation for $G_{\mu/\nu}$ originating from (2-35) can be found in [Borodin and Wheeler 2018, Section 9.5].

2H. *Coloured compositions.* This section closely follows [Borodin and Wheeler 2018, Sections 3.3–3.4], where a more detailed exposition can be found.

Let $\lambda=(\lambda_1,\dots,\lambda_n)$ be a composition of length n and weight m: $|\lambda|=\sum_{i=1}^n\lambda_i=m$. Denote the partial sums of λ by $\sum_{i=1}^k\lambda_i=\ell_k$. We introduce the set $\mathcal{S}_\lambda$ of λ-coloured compositions as follows:

$$\mathcal{S}_\lambda=\bigl\{\mu=(\mu_1\ge\cdots\ge\mu_{\ell_1}|\mu_{\ell_1+1}\ge\cdots\ge\mu_{\ell_2}|\cdots|\mu_{\ell_{n-1}+1}\ge\cdots\ge\mu_{\ell_n})\bigr\}. \tag{2-38}$$

That is, the elements of $\mathcal{S}_\lambda$ are length-m compositions μ, which have been subdivided into blocks of length λ_k, $1\le k\le n$. These blocks demarcate the colouring of μ. Within any given block, the parts of μ have the same colouring and are weakly decreasing.

Two special cases of λ-coloured compositions play special roles. The first is when $\lambda=(n,0,\dots,0)$, when compositions $\mu\in\mathcal{S}_\lambda$ consist of a single block whose parts are weakly decreasing; i.e., one simply recovers partitions. As was noted in (2-29), reducing to this case recovers the symmetric rational functions F_*, G_* from [Borodin 2017; Borodin and Petrov 2017; Borodin and Petrov 2018a].

The second one is when $\lambda=(1,1,\dots,1)=(1^n)$. Then compositions $\mu\in\mathcal{S}_\lambda$ consist of n blocks, each of a different colour. Thus, the parts of μ are not bound by any inequalities; accordingly, one recovers the set of all length-n compositions. We will refer to these as *rainbow compositions*, or as composition in the *rainbow sector*. The functions f_μ, g_μ and $G_{\mu/\nu}$ introduced above were all defined under the assumption that the participating compositions were in the rainbow sector.

Let $\mu\in\mathcal{S}_\lambda$ be a λ-coloured composition, with ℓ_k denoting the partial sums of λ, as above. We associate to each such μ a vector $|\mu\rangle_\lambda\in\mathbb{V}$, defined as

$$|\mu\rangle_\lambda:=\bigotimes_{k=0}^{\infty}|\boldsymbol{A}(k)\rangle_k,\qquad \boldsymbol{A}(k)=\sum_{j=1}^{n}A_j(k)\boldsymbol{e}_j,\qquad A_j(k)=\#\{i:\mu_i=k,\ \ell_{j-1}+1\le i\le\ell_j\}, \tag{2-39}$$

where by agreement $\ell_0=0$. In other words, the component $A_j(k)$ enumerates the number of parts in the j-th block of μ (these are the parts of colour j) which are equal to k. Further, we define vector subspaces $\mathbb{V}(\lambda)$ of $\mathbb{V}$ which provide a natural grading of $\mathbb{V}$:

$$\mathbb{V}=\bigoplus_{m=0}^{\infty}\bigoplus_{|\lambda|=m}\mathbb{V}(\lambda),\qquad \mathbb{V}(\lambda):=\mathrm{Span}_{\mathbb{C}}\{|\mu\rangle_\lambda\}_{\mu\in\mathcal{S}_\lambda}. \tag{2-40}$$

The grading (2-40) splits $\mathbb{V}$ into subspaces with fixed particle content: $\mathbb{V}(\lambda)$ is the linear span of all states consisting of λ_i particles of colour i, for all $i \geq 1$. We refer to these subspaces as *sectors* of $\mathbb{V}$, thus generalizing the "rainbow sector" terminology.

The definitions of the rational functions f_μ, g_ν, and $G_{\mu/\nu}$ naturally lift to coloured composition labels. Concretely, let $\lambda = (\lambda_1, \dots, \lambda_n)$ be a composition of weight m with partial sums ℓ_k as above, and fix a λ-coloured composition $\mu = (\mu_1, \dots, \mu_m) \in \mathcal{S}_\lambda$. In analogy with (2-23), set

$$f_\mu(\lambda; x_1, \dots, x_m) := \langle \varnothing | \left(\prod_{i=1}^{\ell_1} \mathcal{C}_1(x_i) \right) \left(\prod_{i=\ell_1+1}^{\ell_2} \mathcal{C}_2(x_i) \right) \cdots \left(\prod_{i=\ell_{n-1}+1}^{\ell_n} \mathcal{C}_n(x_i) \right) | \mu \rangle_\lambda, \tag{2-41}$$

where $|\mu\rangle_\lambda \in \mathbb{V}(\lambda)$ is given by (2-39), and $\langle \varnothing | \in \mathbb{V}^*$ denotes the (dual) vacuum state $\langle \varnothing | = \bigotimes_{k=0}^{\infty} \langle \mathbf{0} |_k$, which is completely devoid of particles, and $\mathcal{C}_i$'s are the row-operators (2-19). Graphically, this is the partition function of the form (2-23), with the incoming paths in the bottom ℓ_1 rows having colour 1, having colour 2 in the ℓ_2 rows right above those, etc.

One similarly defines, in analogy with (2-24) and using the row operators (2-20),

$$g_\mu(\lambda; x_1, \dots, x_m) := \langle \mu |_\lambda \left(\prod_{i=1}^{\ell_1} \mathcal{B}_1(x_i) \right) \left(\prod_{i=\ell_1+1}^{\ell_2} \mathcal{B}_2(x_i) \right) \cdots \left(\prod_{i=\ell_{n-1}+1}^{\ell_n} \mathcal{B}_n(x_i) \right) | \varnothing \rangle, \tag{2-42}$$

where $\langle \mu |_\lambda \in \mathbb{V}^*(\lambda)$ is the dual of the vector (2-39), and $|\varnothing\rangle \in \mathbb{V}$ denotes the vacuum state $|\varnothing\rangle = \bigotimes_{k=0}^{\infty} |\mathbf{0}\rangle_k$. Again, here the exiting paths in the bottom ℓ_1 rows have colour 1, in the next ℓ_2 rows they have colour 2, etc.

Finally, in analogy with (2-27), for $\mu, \nu \in \mathcal{S}_\lambda$ we define

$$G_{\mu/\nu}(\lambda; x_1, \dots, x_p) = \langle \mu |_\lambda \mathcal{B}_0(x_1) \cdots \mathcal{B}_0(x_p) | \nu \rangle_\lambda, \tag{2-43}$$

with

$$\langle \mu |_\lambda := \bigotimes_{k=0}^{\infty} \langle \boldsymbol{A}(k) |_k, \quad \boldsymbol{A}(k) = \sum_{j=1}^{n} A_j(k) \boldsymbol{e}_j, \qquad |\nu\rangle_\lambda := \bigotimes_{k=0}^{\infty} \langle \boldsymbol{B}(k) |_k, \quad \boldsymbol{B}(k) = \sum_{j=1}^{n} B_j(k) \boldsymbol{e}_j,$$
$$A_j(k) = \#\{i : \mu_i = k,\ \ell_{j-1} + 1 \leq i \leq \ell_j\}, \qquad B_j(k) = \#\{i : \nu_i = k,\ \ell_{j-1} + 1 \leq i \leq \ell_j\},$$

and the graphical depiction (2-27) does not require any modifications. Since the labelling composition λ only participates in this definition as a record of the colours in the boundary states μ and ν, we will often omit it from the notation of $G_{\mu/\nu}$.

3. Extensions

In this section we provide a few straightforward extensions of the results from Section 2, most of which have very similar proofs.

3A. ***Column inhomogeneities.*** In (2-23) we defined the functions f_μ by utilizing the weights L_x of (2-13) with $x = x_i$ and a fixed spin parameter s for all the vertices in the i-th row, $1 \leq i \leq n$. It is meaningful, however, to extend the definition when we take $x = x_i \xi_j$ and $s = s_j$ for the vertex in the i-th row and j-th column, where $1 \leq i \leq n$, $0 \leq j < +\infty$, and $\{\xi_j\}_{j \geq 0}$ and $\{s_j\}_{j \geq 0}$ are two infinite sequences of complex parameters. Correspondingly, in the definitions (2-24) and (2-27) of g_μ and $G_{\mu/\nu}$ one needs to use the weights M_x of (2-14) with $x = x_i\xi_j^{-1}$ and $s = s_j$ for the vertex in the i-th row and j-th column, with the same sequences $\{\xi_j\}_{j\geq 0}$ and $\{s_j\}_{j\geq 0}$.

Many of the results cited in Section 2 and their proofs extend to such an inhomogeneous setup almost *verbatim*. In the colour-blind case of $n = 1$ this can be seen by comparing [Borodin and Petrov 2018a] and [Borodin and Petrov 2017].

The basic reason for such an easy extension lies in the fact that the Yang–Baxter equation (2-15) has a suitable extension. More exactly, the needed deformed equation has the form

$$\sum_{0\leq k_1,k_3\leq n}\ \sum_{\boldsymbol{K}\in\mathbb{N}^n} L_{x\xi}(\boldsymbol{I}, i_1; \boldsymbol{K}, k_1) R_{(qxz)^{-1}}(i_3, k_1; k_3, j_1) M_{z\xi^{-1}}(\boldsymbol{K}, k_3; \boldsymbol{J}, j_3)$$
$$= \sum_{0\leq k_1,k_3\leq n}\ \sum_{\boldsymbol{K}\in\mathbb{N}^n} M_{z\xi^{-1}}(\boldsymbol{I}, i_3; \boldsymbol{K}, k_3) R_{(qxz)^{-1}}(k_3, i_1; j_3, k_1) L_{x\xi}(\boldsymbol{K}, k_1; \boldsymbol{J}, j_1), \tag{3-1}$$

where the important thing to notice is that the parameters of the R-weights in the middle remain independent of ξ, and they are also independent of s by their definition (see Table 2).[6]

Let us quickly go through inhomogeneous analogues of the results from Section 2 that we will need.

The commutation relations (2-22) remain unchanged, as they are related to the R-matrix in (3-1), but the convergence condition (2-21) needs to be modified to

$$\lim_{L\to\infty} \prod_{j=0}^{L} \left| \frac{\xi_j x - s_j}{1 - s_j\xi_j x} \cdot \frac{y - s_j\xi_j}{\xi_j - s_j y} \right| = 0;$$

see [Borodin and Petrov 2018a, Proposition 4.8] for an explanation in the $n = 1$ case. In what follows, we will only need the situation of finitely many (in fact, exactly one) ξ_j's different from 1 and s_j's different from a certain fixed s, in which case, we can clearly continue using (2-21).

Correspondingly, the summation identities (2-34) and (2-35) also remain unchanged, modulo a similar comment about the convergence condition (2-33).

The relation (2-25) between f's and g's remains valid with the following modifications — the inversion of s for the g in the left-hand side needs to be applied to all the s_j's, and the multiplicative constant c_μ needs to be read as

$$\frac{\prod_{j\geq 0} s_j \cdot (q-1)^n q^{\#\{i<j:\mu_i\leq\mu_j\}}}{\prod_{j\geq 0}(s_j^2; q)_{m_j(\mu)}}.$$

The proof remains the same.

The colour-blindness results (2-29) remain in place with identical proofs and inhomogeneous F and G understood as in [Borodin and Petrov 2018a].

The factorization (2-30) for the antidominant compositions looks very similar (and the same argument works):

$$f_\delta(x_1, \ldots, x_n) = \prod_{j\geq 0}(s_j^2; q)_{m_j(\delta)} \cdot \prod_{i=1}^{n}\left(\frac{1}{1 - s_{\delta_i}\xi_{\delta_i}x_i} \prod_{j=0}^{\delta_i - 1} \frac{\xi_j x_i - s_j}{1 - s_j\xi_j x_i} \right). \tag{3-2}$$

[6] This version of the Yang–Baxter equation is also a consequence of the "master" Yang–Baxter equation (2-3).

The recurrence relation (2-31) remains unchanged, with the action of T_i's given by the same Demazure–Lusztig operators (2-32). Its proof was based on the commutation relations between $\mathcal{C}_i$'s given in [Borodin and Wheeler 2018, Theorem 3.2.1] and on [Borodin and Wheeler 2018, Proposition 5.3.3], both of which remain intact in the inhomogeneous setup.

Finally, the orthogonality relation (2-36) remains literally the same, with the contours surrounding all the points $\{\xi_j s_j\}_{j\geq 0}$ instead of just s in the homogeneous case. It is a bit more difficult to convince oneself that this is so because this proof relies on many ingredients. In addition to the facts already mentioned above, one needs monomial expansions of [Borodin and Wheeler 2018, Chapter 6] and a certain explicit contour integral computation. The latter in the inhomogeneous setup is exactly [Borodin and Petrov 2018a, Lemma 7.1], while the proof of the former carries over to the inhomogeneous case in the same spirit as all the other above-mentioned facts.

The principal reason for our carrying the column inhomogeneity of the model throughout Sections 3-6 is to be able to perform a limit transition in column 1 in Section 6, which will then give us access to averages of observables in the fused models and, as a consequence discussed in Section 7, in integrable models of directed random polymers.

3B. ***A simplifying specialization of $G_{\nu/\mu}$.*** Our next goal is to prove the following statement.

Proposition 3.1. *Fix generic complex parameters q and s, a composition λ, and two λ-coloured compositions μ, ν of length $n\geq 1$. Then for $\mathsf{L}\in\mathbb{N}$ and generic $\epsilon\in\mathbb{C}$, $G_{\nu/\mu}(\lambda;\epsilon,q\epsilon,\dots,q^{\mathsf{L}-1}\epsilon)$, with G defined as in (2-43), is a rational function in q^{L}, and there exists a limit*

$$\mathcal{G}_{\nu/\mu} := \lim_{\epsilon\to 0}(q^{n\mathsf{L}}\cdot G_{\nu/\mu}(\lambda;\epsilon,q\epsilon,\dots,q^{\mathsf{L}-1}\epsilon))|_{q^{\mathsf{L}}=(s\epsilon)^{-1}}, \tag{3-3}$$

where in the right-hand side we substituted a particular value into a rational function. Explicitly, $\mathcal{G}_{\nu/\mu}$ has the form

$$\frac{(-s)^{|\mu|}}{(-s)^{|\nu|}}\,\mathcal{G}_{\nu/\mu} = \begin{cases} s^{-2n}\prod_{x=0}^{\infty}\left((s^2;q)_{|\boldsymbol{m}^{(x)}|}q^{-\sum_{i>j}m_i^{(x)}m_j^{(x)}}\prod_{i=1}^{n}q^{m_i^{(x)}H_{>i}^{\nu/\mu}(x+1)}\binom{H_i^{\nu/\mu}(x+1)}{m_i^{(x)}}_q\right), & \textit{if all } \nu_i\neq 0,\\ 0, & \textit{otherwise,}\end{cases} \tag{3-4}$$

where we defined, for each $x\geq 0$, an n-component vector $\boldsymbol{m}^{(x)}$ whose i-th component $m_i^{(x)}$, $1\leq i\leq n$, is equal to the number of parts of μ of colour i that are equal to x (symbolically $\mu=0^{\boldsymbol{m}^{(0)}}1^{\boldsymbol{m}^{(1)}}2^{\boldsymbol{m}^{(2)}}\cdots$), and also ***coloured height functions***

$$H_i^{\varkappa}(x) = \#\{j : \mathrm{colour}(\varkappa_j)=i,\ \varkappa_j\geq x\},\qquad H_{>i}^{\varkappa} = \sum_{k>i}H_k^{\varkappa}(x),\qquad H_*^{\nu/\mu}\equiv H_*^{\nu}-H_*^{\mu}. \tag{3-5}$$

The same conclusion holds in the inhomogeneous setting of Section 3A under the condition that for each column $x\geq 0$ such that there exists a part of μ equal to x (i.e., $|m^{(x)}|>0$), the column rapidity and spin *parameter in that column are still equal to* 1 *and* s, *respectively, as in the homogeneous case, and the factor* $(-s)^{|\mu|-|\nu|}$ *in the left-hand side of* (3-4) *is replaced by* $\prod_{i=1}^{n}\prod_{j=\mu_i}^{\nu_i-1}(-s_j)^{-1}$.

Remark 3.2. We tacitly follow the convention that the q-binomial coefficients vanish unless their arguments are nonnegative, and the top one is at least as large as the bottom one. Thus, the top line of (3-4) can also produce a zero outcome.

Proof. Let us first switch from using the weights M_x of (2-14) in the partition function of the form (2-27) to using the weights M_x^{stoch} of (2-12) instead. The product of the correcting factors $(-s)^{\mathbf{1}_{j\geq 1}}$ in (2-14) over all vertices gives $(-s)^{|\nu|-|\mu|}$ (the exponent counts the number of horizontal steps of all the paths, which is exactly $|\nu|-|\mu|$). Note that this matches the inverse of $(-s)^{|\mu|-|\nu|}$ in the left-hand side of (3-4), and in the inhomogeneous setup, the product of these correcting factors gives

$$\prod_{i=1}^{n}\prod_{j=\mu_i}^{\nu_i-1}(-s_j)$$

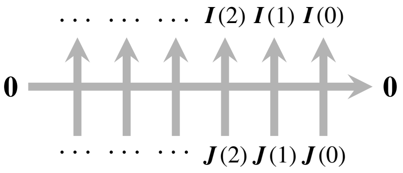
instead.

With the weights M_x^{stoch}, it is a bit more convenient to reflect the partition function with respect to a vertical axis so that paths move to the right horizontally. Then, noticing that the left and right boundary conditions of (2-27) consist of having no entering or exiting paths, we recognize that using the geometric progression $(\epsilon, q\epsilon, \dots, q^{\mathsf{L}-1}\epsilon)$ for horizontal rapidities is equivalent to performing the stochastic fusion in the vertical direction (corresponding to the outer sum in (2-10)). Hence, we obtain that $(-s)^{|\mu|-|\nu|}G_{\nu/\mu}(\lambda; \epsilon, q\epsilon, \dots, q^{\mathsf{L}-1}\epsilon)$ is a one-row partition function with the incoming paths from the bottom parametrizing ν, outgoing paths on the top parametrizing μ, no paths entering or exiting on either side, and the vertex weights given by

$$W_{\mathsf{L},\mathsf{M}}\left(\frac{s}{\epsilon};\ \frac{1}{q};\ \boldsymbol{B}\overset{\boldsymbol{C}}{\underset{\boldsymbol{A}}{\longrightarrow}}\boldsymbol{D}\right)\Bigg|_{q^{\mathsf{M}}\to s^{-2}}. \tag{3-6}$$

Here s and ϵ need to be replaced by s_j and $\xi_j^{-1}\epsilon$, if the vertex is in a column j that carries inhomogeneities (s_j, ξ_j). Pictorially,

$$(-s)^{|\mu|-|\nu|}G_{\nu/\mu}(\lambda; \epsilon, q\epsilon, \dots, q^{\mathsf{L}-1}\epsilon) = \tag{3-7}$$


. . . . $\boldsymbol{I}(2)$ $\boldsymbol{I}(1)$ $\boldsymbol{I}(0)$
$\mathbf{0}$
$\mathbf{0}$
. . . . $\boldsymbol{J}(2)$ $\boldsymbol{J}(1)$ $\boldsymbol{J}(0)$

with $\boldsymbol{I}(k)=\sum_{i=1}^{n}\mathbf{1}_{\mu_i=k}\boldsymbol{e}_i$, $\boldsymbol{J}(k)=\sum_{j=1}^{n}\mathbf{1}_{\nu_j=k}\boldsymbol{e}_j$, for all $k\in\mathbb{Z}_{\geq 0}$. This partition function is a rational function in q^{L} because every vertex weight (3-6) is; see (2-6).

Assume we are in the homogeneous setting first. Observe that if in the right-hand side of (2-6) we have $q^{\mathsf{L}-\mathsf{M}}x=1$, then the only nontrivially contributing term comes from $\boldsymbol{P}=\boldsymbol{C}$, for which the first Φ-factor turns into 1. In terms of our weights (3-6) this corresponds to $q^{\mathsf{M}-\mathsf{L}}s/\epsilon=1$, which is realized by setting $q^{\mathsf{L}}=(s\epsilon)^{-1}$ (recall that $q^{\mathsf{M}}=s^{-2}$).

Writing out the term with $\boldsymbol{P}=\boldsymbol{C}$ we obtain

$$\mathbf{1}_{\boldsymbol{A}+\boldsymbol{B}=\boldsymbol{C}+\boldsymbol{D}}\cdot\left(\frac{s}{\epsilon}\right)^{|\boldsymbol{D}|-|\boldsymbol{B}|}(s\epsilon)^{|\boldsymbol{A}|}s^{-2|\boldsymbol{D}|} \\ \times\frac{(s^{-2};q^{-1})_{|\boldsymbol{C}|}(s\epsilon^{-1};q^{-1})_{|\boldsymbol{B}|-|\boldsymbol{C}|}}{((s\epsilon)^{-1};q^{-1})_{|\boldsymbol{B}|}}\left(\frac{s}{\epsilon}\right)^{|\boldsymbol{C}|}q^{-\sum_{i<j}(B_i-C_i)C_j}\prod_{i=1}^{n}\binom{B_i}{C_i}_{q^{-1}}. \tag{3-8}$$

There is an additional factor of $q^{n\mathsf{L}}=(s\epsilon)^{-n}$ in (3-3). Since our compositions have n parts, there is a total of n paths exiting (3-7) from the top, and we can distribute $(s\epsilon)^{-n}$ as $(s\epsilon)^{-|\boldsymbol{C}|}$ over all vertices in (3-7) (with $\boldsymbol{C}$ corresponding to the paths exiting the vertex at the top). Hence, we need to take the limit as $\epsilon\to 0$ of (3-8) multiplied by $(s\epsilon)^{-|\boldsymbol{C}|}$. This is a straightforward calculation which yields

$$\mathbf{1}_{\boldsymbol{A}+\boldsymbol{B}=\boldsymbol{C}+\boldsymbol{D}}\cdot s^{-2|\boldsymbol{C}|}(s^2;q)_{|\boldsymbol{C}|}\cdot q^{\sum_{i>j}(B_i-C_j)C_j}\prod_{i=1}^{n}\binom{B_i}{C_i}_{q}. \tag{3-9}$$

For the inhomogeneous setting, our hypothesis implies that for any vertex in a column with inhomogeneity parameters $(\tilde{s},\xi)$ we must have $\boldsymbol{C}=\boldsymbol{0}$. This turns the sum in (2-6) into a single term with $\boldsymbol{P}=\boldsymbol{0}$, which leads to the replacement of (3-8) by

$$\mathbf{1}_{\boldsymbol{A}+\boldsymbol{B}=\boldsymbol{C}+\boldsymbol{D}}\cdot\left(\xi^{-1}\frac{s}{\epsilon}\right)^{|\boldsymbol{D}|-|\boldsymbol{B}|}(s\epsilon)^{|\boldsymbol{A}|}\tilde{s}^{-2|\boldsymbol{D}|}\frac{((s\epsilon)^{-1};q^{-1})_{|\boldsymbol{D}|}}{(\xi^{-1}\tilde{s}^{-2}s\epsilon^{-1};q^{-1})_{|\boldsymbol{D}|}}\frac{(\xi^{-1}s\epsilon^{-1};q^{-1})_{|\boldsymbol{B}|}}{((s\epsilon)^{-1};q^{-1})_{|\boldsymbol{B}|}}. \tag{3-10}$$

This expression has the exact same $\epsilon\to 0$ asymptotics as (3-8) with $\boldsymbol{C}=\boldsymbol{0}$, thus confirming (3-9) for the inhomogeneous setup as well.

It remains to interpret $\boldsymbol{B}$ and $\boldsymbol{C}$ in (3-9) in terms of the height functions. Assume that our vertex is located in the column with coordinate $x\geq 0$ (which means, in particular, that $\boldsymbol{A}=\boldsymbol{J}(x)$, $\boldsymbol{C}=\boldsymbol{I}(x)$ in the notation of (3-7)). Then B_i is the number of paths of colour i in (3-7) that enter from the bottom at a location strictly to the left of x minus the number of paths of colour i that exit through the top at a location strictly to the left of x, and this equals $H_i^{\nu/\mu}(x+1)$ as defined in (3-5). Furthermore, C_i is exactly $m_i^{(x)}$. Hence, the part of (3-9) past the indicator function takes the form

$$s^{-2|\boldsymbol{m}^{(x)}|}(s^2;q)_{|\boldsymbol{m}^{(x)}|}\cdot q^{\sum_{j\geq 1}m_j^{(x)}H_{>j}^{\nu/\mu}(x+1)-\sum_{i>j}m_i^{(x)}m_j^{(x)}}\prod_{i=1}^{n}\binom{H_i^{\nu/\mu}(x+1)}{m_i^{(x)}}_{q}, \tag{3-11}$$

and the product of these expressions over $x\geq 0$ gives the first line of the right-hand side of (3-4).

On the other hand, the role of the indicator function $\mathbf{1}_{\boldsymbol{A}+\boldsymbol{B}=\boldsymbol{C}+\boldsymbol{D}}$ in (3-9) is in providing a recipe for uniquely assigning the paths along the horizontal edges of (3-7) inductively: $\boldsymbol{D}$ is assigned the value of $\boldsymbol{A}+\boldsymbol{B}-\boldsymbol{C}$, starting from the far left where no paths are present. This works smoothly until column 0, in which $\boldsymbol{D}$ must be equal to $\boldsymbol{0}$, enforcing $\boldsymbol{A}+\boldsymbol{B}=\boldsymbol{C}$. Since the q-binomial coefficients require $B_i\geq C_i$ for a nonzero outcome, we conclude that $\boldsymbol{A}$ must be $\boldsymbol{0}$ in the 0-th column, giving us the second line of the right-hand side of (3-4). □

An important special case of $\mathcal{G}_{\nu/\mu}$ is when μ has only zero parts. A direct inspection of (3-4) leads to the following:

Corollary 3.3. *For $\mu = 0^n = (0, 0, \dots, 0)$, with notation $\mathcal{G}_\nu := \mathcal{G}_{\nu/0^n}$, the limit (3-3) takes the form*

$$(-s)^{-|\nu|}\,\mathcal{G}_\nu = \begin{cases} s^{-2n}(s^2; q)_n, & \text{if all } \nu_i \neq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3-12}$$

In the inhomogeneous setting, assuming that $(s_0, \xi_0) = (s, 1)$, a similar formula holds, where one needs to replace $(-s)^{-|\nu|}$ in the left-hand side by $\prod_{i=1}^{n}\prod_{j=0}^{\nu_i-1}(-s_j)^{-1}$.

The second symmetrization relation of (2-29) implies that $\mathcal{G}_\nu$ has to equal

$$\lim_{\epsilon\to 0} \mathsf{G}_{\nu^+}(\epsilon, q\epsilon, \dots, q^{\mathsf{L}-1}\epsilon)|_{q^{\mathsf{L}}=(s\epsilon)^{-1}}$$

with ν^+ denoting the dominant reordering of ν, and G as in [Borodin and Petrov 2018a]. This limit was evaluated in [Borodin and Petrov 2018a, Proposition 6.7], and was shown to be equivalent to (3-12).[7]

Let us also record for the future what happens to (3-4) when μ and ν are rainbow compositions (equivalently, $\lambda = (1, 1, \dots, 1)$).

Corollary 3.4. *Under the assumption that μ and ν are rainbow compositions, (3-4) takes the form*

$$\frac{(-s)^{|\mu|}}{(-s)^{|\nu|}}\,\mathcal{G}_{\nu/\mu} = \begin{cases} s^{-2n}\prod_{x=0}^{\infty}\Bigl((s^2; q)_{|\boldsymbol{m}^{(x)}|}\,q^{-\binom{|\boldsymbol{m}^{(x)}|}{2}}\prod_{1\le i\le n;\, m_i^{(x)}=1}\mathbf{1}_{H_i^{\nu/\mu}(x+1)=1}\,q^{H_{>i}^{\nu/\mu}(x+1)}\Bigr), & \text{if all } \nu_i \neq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3-13}$$

The same formula holds in the inhomogeneous setting of Section 3A under the condition that for each column $x \ge 0$ such that there exists a part of μ equal to x (i.e., $|\boldsymbol{m}^{(x)}| > 0$), the column rapidity and spin parameter in that column are still equal to 1 and s, respectively, and the factor $(-s)^{|\mu|-|\nu|}$ in the left-hand side is replaced by $\prod_{i=1}^{n}\prod_{j=\mu_i}^{\nu_i-1}(-s_j)^{-1}$.

Proof. Direct inspection of (3-4). Note that in the rainbow sector, for any given i, $1 \le i \le n$, the i-th colour height function H_i^* defined as in (3-5), can only take values 0 or 1. □

3C. *Colour merging.*

Different versions of colour merging properties of vertex weights have been previously observed and studied in several works including [Foda and Wheeler 2013; Garbali et al. 2017; Kuan 2018; Borodin and Wheeler 2018; Borodin et al. 2019]. We use this section to formulate the statements we need in suitable notation.

Let n_1, n_2 be two positive integers, and let $\theta : \{1, \dots, n_1\} \to \{1, \dots, n_2\}$ be an arbitrary monotone map. It induces a map θ_* that turns a n_1-dimensional vector into an n_2-dimensional one as follows:

$$\theta_* : \boldsymbol{I} = (I_1, \dots, I_{n_1}) \mapsto \boldsymbol{J} = (J_1, \dots, J_{n_2}) \qquad \text{if } J_j = \sum_{i\in\theta^{-1}(j)} I_i. \tag{3-14}$$

In other words, we sum the coordinates of $\boldsymbol{I}$ that have the same θ-image and turn the result into a coordinate of $\boldsymbol{J}$ whose index is that image. Empty sums are interpreted as having value 0.

[7] In fact, [Borodin and Petrov 2018a] provided two different evaluations for this limit, neither of which coincides with the color-blind version of the proof of Proposition 3.1.

Proposition 3.5. *Denote the weights $W_{\mathsf{L},\mathsf{M}}$ given by (2-6) with n-dimensional vector arguments as $W^{(n)}_{\mathsf{L},\mathsf{M}}$. Then for any $n_1, n_2 \geq 1$ and a map θ as above, we have the following colour merging relation: For any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{Z}^{n_1}_{\geq 0}$ and $\widetilde{\boldsymbol{C}}, \widetilde{\boldsymbol{D}} \in \mathbb{Z}^{n_2}_{\geq 0}$ with $|\boldsymbol{A}|, |\widetilde{\boldsymbol{C}}| \leq \mathsf{M}$, $|\boldsymbol{B}|, |\widetilde{\boldsymbol{D}}| \leq \mathsf{L}$,*

$$\sum_{\substack{\boldsymbol{C} \in \mathbb{Z}^{n_1}_{\geq 0} : \theta_*(\boldsymbol{C}) = \widetilde{\boldsymbol{C}} \\ \boldsymbol{D} \in \mathbb{Z}^{n_1}_{\geq 0} : \theta_*(\boldsymbol{D}) = \widetilde{\boldsymbol{D}}}} W^{(n_1)}_{\mathsf{L},\mathsf{M}}\left(x; q; \begin{array}{ccc} & \boldsymbol{C} & \\ \boldsymbol{B} & \overset{\uparrow}{\rightarrow} & \boldsymbol{D} \\ & \boldsymbol{A} & \end{array}\right) = W^{(n_2)}_{\mathsf{L},\mathsf{M}}\left(x; q; \begin{array}{ccc} & \widetilde{\boldsymbol{C}} & \\ \theta_*(\boldsymbol{B}) & \overset{\uparrow}{\rightarrow} & \widetilde{\boldsymbol{D}} \\ & \theta_*(\boldsymbol{A}) & \end{array}\right). \tag{3-15}$$

Proof. For $\mathsf{L} = \mathsf{M} = 1$, when the W-weights turn into matrix elements of the R-matrix (see (2-7)) the statement coincides with [Borodin et al. 2019, Proposition 4.3] (and it is also easy to check directly from the formulas given in Table 2 for the weights). For general $\mathsf{L}, \mathsf{M} \geq 1$, (3-15) readily follows from the $\mathsf{L} = \mathsf{M} = 1$ case and the stochastic fusion (2-10). □

Corollary 3.6. *Colour merging statements similar to Proposition 3.5 hold for the vertex weights L_x, M_x, L^{stoch}_x, M^{stoch}_x and q-Hahn weights (2-16), as well as for the vertex weights defined by (3-9).*

Proof. This follows from the fact that all these weights are obtained from $W_{\mathsf{L},\mathsf{M}}$ by specializations, analytic continuation, multiplication by factors that give the same contribution to the two sides of the merging relation (3-15), and a limit transition $\epsilon \to 0$ in (3-6) in the case of (3-9). □

The colour-blindness statements of Section 2D correspond, in the notation of Proposition 3.5, to $n_2 = 1$. In particular, applying the colour-blindness relation to either the q-Hahn weights (2-16) or to the weights (3-9) and removing common factors on the two sides, one obtains the q-identity (3-16).

Corollary 3.7. *For any $Q \in \mathbb{C}$, $(\alpha_1, \dots, \alpha_n) \in \mathbb{Z}^n_{\geq 0}$, and a fixed $|\beta| \in \mathbb{Z}_{\geq 0}$, $|\beta| \leq |\alpha|$, one has*

$$\sum_{\substack{0 \leq \beta_i \leq \alpha_i,\, 1 \leq i \leq n, \\ \beta_1 + \cdots + \beta_n = |\beta|}} Q^{\sum_{1 \leq i < j \leq n} \beta_i(\alpha_j - \beta_j)} \prod_{i=1}^{n} \binom{\alpha_i}{\alpha_i - \beta_i}_Q = \binom{|\alpha|}{|\alpha| - |\beta|}_Q. \tag{3-16}$$

Since compositions are vectors, the map θ_* from (3-14) sends any composition of length n_1 to a composition of length n_2. This can be naturally extended to coloured compositions as follows.

Let λ be a composition with $\ell(\lambda) = n_1$, weight $|\lambda| = m$, and partial sums $\ell_k = \sum_{i=1}^{k} \lambda_i$. Then $\rho := \theta_*(\lambda)$ is a composition with $\ell(\rho) = n_2$, same weight $|\rho| = m$, and partial sums that we denote as $r_k = \sum_{i=1}^{k} \rho_i$. Further, let μ be a λ-coloured composition of length m; see Section 2H for a definition. As in (2-39), we can encode μ by a sequence of n_1-dimensional vectors $\{\boldsymbol{A}(k)\}_{k \geq 0}$:

$$\boldsymbol{A}(k) = \sum_{j=1}^{n_1} A_j(k) \boldsymbol{e}_j, \qquad A_j(k) = \#\{i : \mu_i = k,\ \ell_{j-1} + 1 \leq i \leq \ell_j\}.$$

It is not difficult to see that the sequence of n_2-dimensional vectors $\boldsymbol{B}(k) := \theta_*(\boldsymbol{A}(k))$, $k \geq 0$, corresponds, via

$$\theta_*(\boldsymbol{A}(k)) = \boldsymbol{B}(k) = \sum_{j=1}^{n_2} B_j(k) \boldsymbol{e}_j, \qquad B_j(k) = \#\{i : \nu_i = k,\ r_{j-1} + 1 \leq i \leq r_j\},$$

to a ρ-coloured composition ν of length m that we will define to be the image of μ under θ_*:

$$\theta^*(\mu) = \nu. \tag{3-17}$$

In less formal terms, if we view a coloured composition μ as positions of finitely many paths coloured by $\{1, \dots, n_1\}$, then $\nu = \theta_*(\mu)$ represents positions of the same paths that have been recoloured according to the map θ. Note that we required θ to be monotone, which means that the order of colours is being preserved.

Proposition 3.8. *Let λ be a composition with $\ell(\lambda) = n_1$ of weight $|\lambda| = m$, ρ be a composition with $\ell(\rho) = n_2$ and same weight $|\rho| = m$ such that $\theta_*(\lambda) = \rho$, ν' by a λ-coloured composition, and ν'' be a ρ-coloured composition (both of length m). Then for any $p \geq 1$ and complex parameters $x_1, x_2, \dots,$ we have*

$$\sum_{\mu : \theta_*(\mu) = \nu''} f_\mu(\lambda; x_1, \dots, x_m) = f_{\nu''}(\rho; x_1, \dots, x_m), \tag{3-18}$$

$$\sum_{\mu : \theta_*(\mu) = \nu''} G_{\nu'/\mu}(\lambda; x_1, \dots, x_p) = G_{\theta_*(\nu')/\nu''}(\rho; x_1, \dots, x_p). \tag{3-19}$$

Remark 3.9. When $\lambda = (1, 1, \dots, 1)$ and $n_2 = 1$, one recovers the symmetrization formulas (2-29).

Proof. In complete analogy with proofs of (2-29) in [Borodin and Wheeler 2018, Propositions 3.4.4 and 4.4.3], the argument consists in multiple applications of Proposition 3.5, or rather its version for the weights L_x and M_x (Corollary 3.6) used in the definitions of f's and G's. In the case of (3-18), one starts with the top-right nontrivial vertex of the partition function (2-23) (with appropriate, not necessarily rainbow colours of entering paths on the left), while in the case of (3-19) one starts with the top-left nontrivial vertex of the partition function (2-27), and then moves step by step into the bulk of the partition function. Once the colour-merging summation has been performed for all nontrivial vertices, one recovers the right-hand sides of (3-18) and (3-19). □

4. Cauchy identities

The goal of this section is to show how the skew-Cauchy identity (2-35) leads to formulas for averages of certain observables for stochastic vertex models.

For this section let us assume that we are either in the column-homogeneous situation, or, slightly more generally, the number of columns in which inhomogeneity parameters (s_j, ξ_j) of Section 3A are different from $(s, 1)$ is finite. We start by extending the skew-Cauchy identity (2-35) to limiting versions of the G-functions from Section 3B.

Proposition 4.1. *Let μ be a rainbow composition of length $n \geq 1$, and $x_1, \dots, x_n$ be complex parameters satisfying*

$$\left| \frac{s(x_i - s)}{1 - s x_i} \right| < 1, \qquad 1 \leq i \leq n. \tag{4-1}$$

Then we have

$$\sum_{\nu} f_\nu(x_1,\dots,x_n)\,\mathcal{G}_{\nu/\mu} = \prod_{i=1}^{n}\left(1-\frac{x_i}{s}\right)\cdot f_\mu(x_1,\dots,x_n), \tag{4-2}$$

where $\mathcal{G}_{\nu/\mu}$ is as in Corollary 3.4, and f_*'s are as in (2-23).

Proof. We start with the summation identity (2-35), multiply both sides by q^{np}, and substitute $p=\mathsf{L}$, $(y_1,\dots,y_p)=(\epsilon,q\epsilon,\dots,q^{\mathsf{L}-1}\epsilon)$ with $\epsilon=s^{-1}q^{-\mathsf{L}}$ to eventually match with (3-3). The identity takes the form

$$\sum_{\nu} f_\nu(x_1,\dots,x_n)\cdot q^{n\mathsf{L}}G_{\nu/\mu}(\epsilon,q\epsilon,\dots,q^{\mathsf{L}-1}\epsilon) = \prod_{i=1}^{n}\frac{1-x_i/s}{1-x_i/(sq^{\mathsf{L}})}\cdot f_\mu(x_1,\dots,x_n), \tag{4-3}$$

where the convergence conditions (2-33) read

$$\left|\frac{x_i-s}{1-sx_i}\cdot\frac{y_j-s}{1-sy_j}\right| = \left|\frac{x_i-s}{1-sx_i}\cdot\frac{s^{-1}q^{j-\mathsf{L}-1}-s}{1-q^{j-\mathsf{L}-1}}\right| = \left|\frac{x_i-s}{1-sx_i}\cdot\frac{1-s^2q^{\mathsf{L}-j+1}}{s(q^{\mathsf{L}-j+1}-1)}\right| < 1$$

for $1\le i\le n$ and $1\le j\le \mathsf{L}$. They will hold for all $\mathsf{L}=1,2,\dots$ as long as the x_i's range over a sufficiently small neighbourhood of s (we assume $|q|<1$).

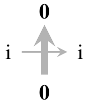
As in the proof of Proposition 3.1, $q^{n\mathsf{L}}G_{\nu/\mu}(\epsilon,q\epsilon,\dots,q^{\mathsf{L}-1}\epsilon)$ can be represented as a one-row partition function of the form (3-7) with fused weights (3-6) multiplied by $(s\epsilon)^{-|\boldsymbol{C}|}$. Denoting $\mathfrak{l}:=q^{\mathsf{L}}$, we can rewrite those vertex weights as follows (see (3-8) and recalling that $\epsilon=(s\mathfrak{l})^{-1}$),

$$\mathbf{1}_{\boldsymbol{A}+\boldsymbol{B}=\boldsymbol{C}+\boldsymbol{D}}\cdot\frac{s^{-2|\boldsymbol{B}|-|\boldsymbol{C}|}(s^{-2};q^{-1})_{|\boldsymbol{C}|}(s^2\mathfrak{l};q^{-1})_{|\boldsymbol{B}|-|\boldsymbol{C}|}}{(\mathfrak{l};q^{-1})_{|\boldsymbol{B}|}}\left(s^2\mathfrak{l}\right)^{|\boldsymbol{C}|}q^{-\sum_{i<j}(B_i-C_i)C_j}\prod_{i=1}^{n}\binom{B_i}{C_i}_{q^{-1}}. \tag{4-4}$$

Using the fact that $|\boldsymbol{B}|,|\boldsymbol{C}|\le n$, one readily sees that this expression remains uniformly bounded as $\mathfrak{l}$ varies in the extended complex plane while staying uniformly bounded away from the potential poles at $\{1,q,\dots,q^{n-1}\}$. Hence, we obtain an estimate of the form $|q^{n\mathsf{L}}G_{\nu/\mu}(\epsilon,q\epsilon,\dots,q^{\mathsf{L}-1}\epsilon)|\le \mathrm{const}^{|\nu|}$ for such $\mathfrak{l}$. On the other hand, we also have

$$|f_\nu(x_1,\dots,x_n)|\le \mathrm{const}\prod_{i=1}^{n}\left|\frac{x_i-s}{1-sx_i}\right|^{|\nu|}, \tag{4-5}$$

which follows from the fact that in the partition function (2-23) only vertices of the form



have the number of appearances that is not a priori bounded, and the weight of such a vertex in the j-th row, according to (2-13), is $(x_j-s)/(1-sx_j)$ (at least sufficiently far to the right, even if there are finitely many column inhomogeneities). We conclude that for the x_i's in a sufficiently small neighbourhood of s and $\mathfrak{l}$ staying away from $\{1,q,\dots,q^{n-1}\}$, the series in the left-hand side of (4-3) converges uniformly, yielding an analytic function of $\mathfrak{l}$. Clearly, the right-hand side of (4-3) is also analytic in $\mathfrak{l}$, which implies that (4-3) holds under the same conditions on the x_i's and $\mathfrak{l}$. Substituting $\mathfrak{l}=\infty$ (equivalently, $\epsilon=0$) into (4-3) leads to (4-2).

Once (4-2) is proved for x_i's sufficiently close to s and $|q|<1$, we can relax the assumptions by further analytic continuation in these parameters (although we will not need $|q|\ge 1$ below). In particular, since the explicit formula (3-4) for $\mathcal{G}_{\nu/\mu}$ shows $\mathrm{const}\cdot|s|^{|\nu|}$ behaviour of $\mathcal{G}_{\nu/\mu}$ for large ν, using (4-5) we can extend the equality to x_i's satisfying (4-1). □

Taking $\mu=(0,0,\dots,0)$ in Proposition 4.1, we obtain the following:

Corollary 4.2. *For $x_1,\dots,x_n\in\mathbb{C}$ satisfying (4-1), we have*

$$\prod_{i=1}^{n}\frac{1-sx_i}{s(s-x_i)}\cdot\sum_{\nu:\,\text{all }\nu_i>0}(-s)^{|\nu|}f_\nu(x_1,\dots,x_n)=1, \tag{4-6}$$

where the summation is over all compositions ν of length n with no zero parts. In the inhomogeneous setting of Section 3A, assuming that $(s_0,\xi_0)=(s,1)$, a similar formula holds, where one needs to replace $(-s)^{|\nu|}$ in the left-hand side by $\prod_{i=1}^{n}\prod_{j=0}^{\nu_i-1}(-s_j)$.

Proof. This is given by straightforward substitution of $\mu=(0,\dots,0)$ and (3-12) into (4-2), with the evaluation

$$f_{(0,\dots,0)}(x_1,\dots,x_n)=\prod_{i=1}^{n}\frac{1-s^2q^{i-1}}{1-sx_i}, \tag{4-7}$$

which follows from the fact that the corresponding partition function (2-23) is the product of L-weights (2-13) for the unique path configuration that gives a nonzero contribution. □

It is not difficult to extend Proposition 4.1 and Corollary 4.2 to partially merged colours.

Proposition 4.3. *Let λ be a composition with $|\lambda|=n$, μ be a λ-coloured composition, and $x_1,\dots,x_n\in\mathbb{C}$ satisfy (4-1). Then*

$$\sum_{\nu\text{ is }\lambda\text{-coloured}}f_\nu(\lambda;x_1,\dots,x_n)\,\mathcal{G}_{\nu/\mu}=\prod_{i=1}^{n}\left(1-\frac{x_i}{s}\right)\cdot f_\mu(\lambda;x_1,\dots,x_n), \tag{4-8}$$

which in the case of μ having only zero parts, can be rewritten as

$$\prod_{i=1}^{n}\frac{1-sx_i}{s(s-x_i)}\cdot\sum_{\substack{\nu\text{ is }\lambda\text{-coloured;}\\ \text{all }\nu_i>0}}(-s)^{|\nu|}f_\nu(\lambda;x_1,\dots,x_n)=1. \tag{4-9}$$

In the inhomogeneous setting of Section 3A, relation (4-9) also holds under assumption that $(s_0,\xi_0)=(s,1)$ and with $(-s)^{|\nu|}$ in the left-hand side replaced by $\prod_{i=1}^{n}\prod_{j=0}^{\nu_i-1}(-s_j)$.

Proof. Let $\theta:\{1,\dots,n\}\to\{1,\dots,\ell(\lambda)\}$ be the unique colour merging monotone map such that $\theta_*((1,\dots,1))=\lambda$; see Section 3C. Then we can sum (4-2) over rainbow μ with a given image $\theta_*(\mu)$. The right-hand side is immediately computed via (3-18), and in the left-hand side we use (3-19) and subsequently perform, using (3-18), a partial summation over ν's with the same image $\theta_*(\nu)$. The result is (4-8), with $\theta_*(\mu)$ and $\theta_*(\nu)$ replaced back by μ and ν. The second relation (4-9) follows from (4-8) in the same way as in the rainbow case of Corollary 4.2. □

Definition 4.4. For any $n \geq 1$, composition λ with $|\lambda| = n$, and $q, s, x_1, \dots, x_n \in \mathbb{C}$ satisfying (4-1), we define a (generally speaking, complex valued) probability measure on the subset of $\mathcal{S}_\lambda$ (defined in (2-38)) consisting of λ-coloured compositions ν with no zero parts by

$$\boldsymbol{P}(\{\nu\}) = \prod_{i=1}^{n} \frac{1 - s x_i}{s(s - x_i)} \cdot (-s)^{|\nu|} f_\nu(\lambda; x_1, \dots, x_n). \tag{4-10}$$

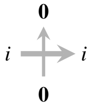
Graphically, the weight of this measure could be seen as partition functions of the form (2-23) with incoming colours on the left partially identified according to λ, with vertex weights L^{stoch} given by (2-13), and conditioned to have no exiting paths in the 0-th column (i.e., $\boldsymbol{A}(0) = \boldsymbol{0}$ in terms of (2-23)).

In what follows we will also use the notation

$$f_\nu^{\text{stoch}}(\lambda; x_1, \dots, x_n) = (-s)^{|\nu|} f_\nu(\lambda; x_1, \dots, x_n). \tag{4-11}$$

In the inhomogeneous setting of Section 3A, we will assume that $(s_0, \xi_0) = (s, 1)$, and in the right-hand sides of (4-10) and (4-11) replace $(-s)^{|\nu|}$ by $\prod_{i=1}^{n} \prod_{j=0}^{\nu_i - 1} (-s_j)$.

The graphical interpretation is based on the observation that the prefactor of the sums in (4-6) and (4-9) is exactly the inverse of the product of L^{stoch}-weights of vertices



in the 0-th column of a partition function of the form (2-23) with no turns in the 0-th column, and (4-11) is the result of computing the partition function for f_ν with L-weights replaced by the L^{stoch}-weights. Together with the stochasticity of the L^{stoch}-weights, this also implies (4-6) and (4-9).

Definition 4.5. In the context of Definition 4.4, for any λ-coloured composition $\mu = 0^{\boldsymbol{m}^{(0)}} 1^{\boldsymbol{m}^{(1)}} 2^{\boldsymbol{m}^{(2)}} \cdots$, introduce an observable $\mathcal{O}_\mu$, whose values on λ-coloured compositions ν with no zero parts are given by

$$\mathcal{O}_\mu(\nu) = \prod_{x \geq 1} \prod_{i \geq 1} q^{m_i^{(x)} H_{>i}^{\nu/\mu}(x+1)} \binom{H_i^{\nu/\mu}(x+1)}{m_i^{(x)}}_q, \tag{4-12}$$

where we use the coloured height functions (3-5). Note that only nonzero parts of μ play a role in this definition.

For rainbow compositions, i.e., when $\lambda = (1, \dots, 1)$, the observables take a simpler form; see Corollary 3.4:

$$\begin{aligned} \mathcal{O}_\mu^{\text{rainbow}}(\nu) &= \prod_{\substack{x \geq 1,\, i \geq 1 \\ m_i^{(x)} = 1}} \mathbf{1}_{H_i^{\nu/\mu}(x+1) = 1} \, q^{H_{>i}^{\nu/\mu}(x+1)} \\ &= \prod_{\substack{x \geq 1,\, i \geq 1 \\ m_i^{(x)} = 1}} \frac{q^{H_{>i+1}^{\nu/\mu}(x+1)} - q^{H_{>i}^{\nu/\mu}(x+1)}}{q - 1}. \end{aligned} \tag{4-13}$$

Example 4.6. In the colour-blind case $\lambda = (n)$, thinking of the coordinates of μ as being ordered: $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$, we can rewrite (4-12) as a shifted q-moment of the colourless height function:

$$\mathcal{O}_\mu^{\text{colour-blind}}(\nu) = \frac{(1-q^{H^\nu(\mu_1+1)})(1-q^{H^\nu(\mu_2+1)-1})\cdots(1-q^{H^\nu(\mu_k+1)-k+1})}{\prod_{j\geq 1}(q;q)_{\mathrm{mult}_j(\mu)}}, \tag{4-14}$$

where $\mathrm{mult}_j(\mu) = \#\{i : \mu_i = j\}$ and $k = \max\{j : \mu_j > 0\}$.

We are now in position to formulate the main result of this section.

Theorem 4.7. *With notation of Definitions 4.4 and 4.5 above, we have*

$$\mathbb{E}\,\mathcal{O}_\mu = \mathbb{E}_\nu\left[\mathcal{O}_\mu(\nu)\right] = \frac{q^{\sum_{x\geq 1}\sum_{i>j} m_i^{(x)} m_j^{(x)}}}{\prod_{x\geq 0}(s^2;q)_{|\boldsymbol{m}^{(x)}|}} \prod_{i=1}^{n}(1-sx_i)\cdot f_\mu^{\text{stoch}}(\lambda; x_1,\ldots,x_n), \tag{4-15}$$

where the expectation is taken with respect to the weights (4-10), and the observables are given by (4-12) in the general case, or by (4-13) in the rainbow case.

The formula (4-15) also holds in the inhomogeneous setting of Section 3A under the assumption that $(s_x, \xi_x) = (s, 1)$ for any x such that $|\boldsymbol{m}^{(x)}| > 0$, and for $x = 0$.

Proof. Let us take the ratio of two skew-Cauchy identities (4-8) with μ and with $\mu = (0,\ldots,0)$. This yields

$$\frac{\sum_\nu f_\nu(\lambda; x_1,\ldots,x_n)\,\mathcal{G}_{\nu/\mu}}{\sum_\nu f_\nu(\lambda; x_1,\ldots,x_n)\,\mathcal{G}_\nu} = \frac{f_\mu(\lambda; x_1,\ldots,x_n)}{f_{(0,\ldots,0)}(\lambda; x_1,\ldots,x_n)},$$

which can be rewritten, via (4-11) and (4-7) (stated in the rainbow case, but also holding in the non-rainbow one for the same reasons), as

$$\frac{\sum_\nu f_\nu(\lambda; x_1,\ldots,x_n)\,\mathcal{G}_\nu\cdot(\mathcal{G}_{\nu/\mu}/\mathcal{G}_\nu)}{\sum_\nu f_\nu(\lambda; x_1,\ldots,x_n)\,\mathcal{G}_\nu} = \frac{\prod_{i=1}^n(1-sx_i)}{(-s)^{|\mu|}(s^2;q)_n} f_\mu^{\text{stoch}}(\lambda; x_1,\ldots,x_n). \tag{4-16}$$

The summations are taken over λ-coloured compositions ν with no zero parts.

Observe that the left-hand side of (4-16) is exactly $\mathbb{E}_\nu(\mathcal{G}_{\nu/\mu}/\mathcal{G}_\nu)$; see Definition 4.4. The expression (3-4) for $\mathcal{G}_{\nu/\mu}$ (which is also the source of our assumption in the inhomogeneous setting) implies

$$\frac{\mathcal{G}_{\nu/\mu}}{\mathcal{G}_\nu} = \frac{(s^2;q)_{|\boldsymbol{m}^{(0)}|}}{(-s)^{|\mu|}(s^2;q)_n} q^{\sum_{i=1}^n\left(m_i^{(0)}\sum_{n\geq j>i}(H_j^{\nu/\mu}(1)-m_j^{(0)})\right)} \prod_{i\geq 1}\binom{H_i^{\nu/\mu}(1)}{m_i^{(0)}}_q \\ \times\prod_{x\geq 1}(s^2;q)_{|\boldsymbol{m}^{(x)}|} q^{-\sum_{i>j} m_i^{(x)} m_j^{(x)}}\cdot\mathcal{O}_\mu(\nu).$$

The only ν-dependent part of this expression is $\mathcal{O}_\mu(\nu)$, and it stays under the expectation; the other factors can be moved to the right-hand side of (4-16). It only remains to notice that for $\mathcal{G}_{\nu/\mu}$ not to vanish, we must have $H_i^{\nu/\mu}(1) = m_i^{(0)}$ for all colours $i \geq 1$. To see that, it might be easiest to return to the expression (3-9) for the vertex weights in the one-row representation (3-7) of $\mathcal{G}_{\nu/\mu}$, and note that in the 0-th column we have $\boldsymbol{D} = \boldsymbol{0}$, which together with $\boldsymbol{A}+\boldsymbol{B} = \boldsymbol{C}+\boldsymbol{D}$ and $\boldsymbol{B} \geq \boldsymbol{C}$ (enforced by the q-binomial coefficients) implies $\boldsymbol{B} = \boldsymbol{C}$, which is exactly the statement we are making, as $B_i = H_i^{\nu/\mu}(1)$ and $C_i = m_i^{(0)}$. Taking these equalities into account and cancelling out common factors gives (4-15). □

5. Integral representations for f_μ

The result of averaging the observable $\mathcal{O}_\mu$ in Theorem 4.7 is the function f_μ, up to simple prefactors. The coloured composition μ here has the same dimension (i.e., number of parts) as the coloured compositions ν over which we are averaging (see Definition 4.4), and thus does not offer much reduction in complexity. The goal of this section is to show that f_μ has an integral representation of the dimension equal to the number of *nonzero* parts of μ. Hence, by choosing μ consisting of mostly zeros, we will be able to obtain tangible formulas for averages over ν's of a growing dimension.

Definition 5.1. For a coloured composition μ (defined as in Section 2H), we will denote by $\mu^{\geq 1}$ the coloured composition obtained from μ by removing all of its zero parts. The colouring of $\mu^{\geq 1}$ will be the one naturally inherited from μ, and we will denote it by $\lambda^{\geq 1}$ if μ is λ-coloured, or by colour$(\mu^{\geq 1})$ if λ is not explicit.

Our first goal is to prove the following integral representation of f_μ for rainbow compositions.

Proposition 5.2. *Let $\mu = 0^{\boldsymbol{m}^{(0)}} 1^{\boldsymbol{m}^{(1)}} 2^{\boldsymbol{m}^{(2)}} \cdots$ be a rainbow composition of length $n \geq 1$ not consisting entirely of zeros, let $\mu^{\geq 1}$ be as in Definition 5.1 with $m := n - |\boldsymbol{m}^{(0)}| \geq 1$ being the length of $\mu^{\geq 1}$, and let $c_1 < \cdots < c_m$ be the colours of the (necessarily nonzero) parts of $\mu^{\geq 1}$. Then*

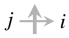
$$f_\mu(x_1, \ldots, x_n) = \frac{(s^2; q)_{|\boldsymbol{m}^{(0)}|}}{(1 - s x_1) \cdots (1 - s x_n)} \frac{(-1)^m}{(2\pi\sqrt{-1})^m} \oint \cdots \oint_{\text{around}\{x_j^{-1}\}} \prod_{1 \leq i < j \leq m} \frac{y_j - y_i}{y_j - q y_i} \\ \times f_{\mu^{\geq 1}}(y_1^{-1}, \ldots, y_m^{-1}) \prod_{j=1}^{m} \left(\frac{y_j - s}{y_j} \frac{x_{c_j}}{1 - x_{c_j} y_j} \prod_{i > c_j}^{n} \frac{1 - q x_i y_j}{1 - x_i y_j} \, dy_j \right), \tag{5-1}$$

*where (positively oriented) integration contours are chosen to encircle all points $\{x_j^{-1}\}_{j=1}^n$ and no other singularities of the integrand, or as q-**nested** closed simple curves with y_i-contour containing $q^{-1} \cdot (y_j\text{-contour})$ for all $i < j$, and all of the contours encircling $\{x_j^{-1}\}_{j=1}^n$.*

The formula also holds in the column inhomogeneous setting under the assumption that in the 0-th column $(s_0, \xi_0) = (s, 1)$.

Remark 5.3. Contour integration around a specific set of singularities can be viewed formally as the sum of residues at those singularities, and such a sum would make sense if the parameters are such that required contours are not possible to construct. This gives a slightly different way of interpreting (5-1), as well as all the other integral representations below.

Remark 5.4. According to Definition 5.1, $\mu^{\geq 1}$ is a coloured composition, and thus $f_{\mu^{\geq 1}}(y_1^{-1}, \ldots, y_m^{-1})$ should really be written as $f_{\mu^{\geq 1}}(\lambda; y_1^{-1}, \ldots, y_m^{-1})$, where λ is the colouring composition for $\mu^{\geq 1}$. However, since all parts of $\mu^{\geq 1}$ have different colours, and our vertex weights always depend on the colours only through their ordering, we could replace the colours $c_1 < \cdots < c_m$ represented in $\mu^{\geq 1}$ by $1, 2, \ldots, m$, and denoting the resulting rainbow composition by $\tilde{\mu}^{\geq 1}$, we would have $f_{\mu^{\geq 1}}(\lambda, y_1^{-1}, \ldots, y_m^{-1}) = f_{\tilde{\mu}^{\geq 1}}(y_1^{-1}, \ldots, y_m^{-1})$. It is this function that we denoted as $f_{\mu^{\geq 1}}(y_1^{-1}, \ldots, y_m^{-1})$, thus slightly abusing the notation.

Remark 5.5. In the case $s_0 = 0$, there exists a fairly straightforward argument leading to a formula similar to (5-1). Namely, in that case vertices of the form  with $j > i$ have weight zero (in the 0-th column) due to vanishing of the bottom right entry in (2-13). This means that the paths of colours $c_1, \dots, c_m$ are not allowed make any vertical steps in column 0, as otherwise they would have no way of exiting this column to the right (because of the ordering of entering colours along the left boundary). This completely determines the configuration of paths in column 0, and the contribution of the remaining columns can be encoded as the matrix element

$$\langle \varnothing | \mathcal{C}_0(x_1) \cdots \mathcal{C}_0(x_{c_1-1}) \mathcal{C}_{c_1}(x_{c_1}) \mathcal{C}_0(x_{c_1+1}) \cdots \mathcal{C}_0(x_{c_m-1}) \mathcal{C}_{c_m}(x_{c_m}) \mathcal{C}_0(x_{c_m+1}) \cdots \mathcal{C}_0(x_n) | \varkappa \rangle,$$

with $\varkappa = \mu^{\geq 1} - 1^m$ and $\mathcal{C}$-row operators from Section 2C. The summation over $\varkappa$ of such expressions multiplied by $g^*_{\varkappa}(y_1, \dots, y_m)$ is evaluated as an explicit product similar to (2-34) via the commutation relations (2-22), and then the coefficients of $g^*_{\varkappa}$ are extracted by the orthogonality (2-36).

It is not clear, however, how to extend this argument to $s_0 \neq 0$, and we need to employ a different idea in the proof below.

Proof of Proposition 5.2. Our argument consists of two steps: We will first prove the formula for $(c_1, \dots, c_m) = (n-m+1, \dots, n)$, and then show that the formula continues to hold when we reduce one of the c_j's by 1. Iterations of such reductions will cover all possible choices of $1 \leq c_1 < \cdots < c_m \leq n$.

Both steps work identically in the column homogeneous and inhomogeneous settings, and we will give the argument in the homogeneous case. The inhomogeneous analogs of several statements from Section 2 used below can be found in Section 3A.

Assume that $(c_1, \dots, c_m) = (n-m+1, \dots, n)$. This means that paths that enter the partition function of the form (2-23) from the left in rows $1, \dots, n-m$ must immediately turn up and exit on top, while paths that enter in rows $n-m+1, \dots, n$ must move to the right across the 0-th column (recall that no horizontal edge can carry more than one path). Hence, the configuration of paths in the 0-th column is completely determined, and the product of the L-weights (2-13) in this column gives

$$\frac{(s^2; q)_{|\boldsymbol{m}^{(0)}|}}{\prod_{i=1}^{n-m}(1 - s x_i)} \prod_{j=n-m+1}^{n} \frac{x_j - s}{1 - s x_j}.$$

On the other hand, the contribution of the remaining columns can be written as

$$\prod_{j=n-m+1}^{n} \frac{1 - s x_i}{x_i - s} \cdot f_{\mu^{\geq 1}}(x_{n-m+1}, \dots, x_n),$$

where the prefactor is responsible for the fact that the partition function for $f_{\mu^{\geq 1}}(x_{n-m+1}, \dots, x_n)$ starts with column 0, and we only had weights of vertices in columns ≥ 1 remaining. We now need to show that the product of the last two expressions agrees with the right-hand side of (5-1). With our choice of c_i's, the integrand is independent of $x_1, \dots, x_{n-m}$, and its only poles are at $y_i = x_j^{-1}$ with $n-m+1 \leq j \leq n$. The number of integration variables thus coincides with the number of potential pole locations, and no two variables can have nonvanishing residues at the same location because of $\prod_{i<j}(y_i - y_j)$ in the integrand. Hence, the residue locations cover all points $x_{n-m+1}^{-1}, \dots, x_n^{-1}$ exactly once, and $\prod_{j=1}^{m}(y_j - s)x_{c_j}$ in the

integrand necessarily evaluates to $\prod_{j=n-m+1}^{n}(1-sx_j)$ in every set of nontrivial residues taken. Moving this factor out of the integral, we are led to the following desired equality:

$$f_{\mu^{\geq 1}}(x_{n-m+1},\dots,x_n)=\frac{(-1)^m}{(2\pi\sqrt{-1})^m}\oint\cdots\oint_{\text{around }\{x_j^{-1}\}}\prod_{1\leq i<j\leq m}\frac{y_j-y_i}{y_j-qy_i}\\ \times f_{\mu^{\geq 1}}(y_1^{-1},\dots,y_m^{-1})\prod_{j=n-m+1}^{n}\left(\frac{1}{1-x_jy_j}\prod_{i>j}^{n}\frac{1-qx_iy_j}{1-x_iy_j}\,\frac{dy_j}{y_j}\right). \tag{5-2}$$

Comparing to (2-37), we see that the difference is in $(-1)^m$ and the choice of contours, which both come from the same origin. Namely, let us take the right-hand side of (2-37) (with n replaced by m and $(x_1,\dots,x_n)$ replaced by $(x_{n-m+1},\dots,x_n)$), and deform the outermost y_m-contour in the outside direction, moving it through ∞ and closing around $\{x_j^{-1}\}$. The recursive construction of (rainbow) f_μ's from Section 2E readily shows that $(y_1\cdots y_m)^{-1}f_{\mu^{\geq 1}}(y_1^{-1},\dots,y_m^{-1})$ is a ratio of two polynomials in y's with the denominator consisting only of factors of the form (y_i-s), and $f_{\mu^{\geq 1}}(y_1^{-1},\dots,y_m^{-1})$ viewed as such a ratio has the numerator degree 1 less than the denominator degree in each of the variables y_i, $1\leq i\leq m$. Hence, our deformation of the y_m collects no residues along the way and yields the factor (-1) for changing the contour direction.

Next, we do the same deformation with the y_{m-1}-contour. Following the same reasoning, there is only one possible pole along the way: $y_{m-1}=q^{-1}y_m$. If we leave this potential singularity inside the y_{m-1}-contour, and proceed similarly for subsequent deformations, then we end up with two q-nested contours surrounding $\{x_j^{-1}\}$. It turns out that the q-nestedness is not necessary. Indeed, a direct inspection of the integrand shows that the only possible pole of y_m inside its current contour is x_m^{-1}, and then the factor $(1-qx_my_{m-1})$ makes the residue at $y_{m-1}=q^{-1}y_m$ vanish. Hence, we can close the y_{m-1}-contour around $\{x_j^{-1}\}$, orient it positively, and acquire another (-1).

Continuing with this procedure for $y_{m-2}, y_{m-3},\dots,y_1$-contours (in this order), we turn (2-37) into (5-2), thus completing the first step of the proof.

Let us now see why lowering of the c_j's in (5-1) keeps the formula intact. From the point of view of the (rainbow) composition μ, replacing $c_j\mapsto c_j-1$ (assuming there is no $i\neq j$ such that $c_i=c_j-1$) is equivalent to swapping $\mu_{c_j-1}=0$ and $\mu_{c_j}>0$. This can be done with the help of the exchange relations (2-31) by acting on the right-hand side of (5-1) by

$$T_{c_j-1}=q-\frac{x_{c_j-1}-qx_{c_j}}{x_{c_j-1}-x_{c_j}}(1-\mathfrak{s}_{c_j-1});$$

see (2-32). The only part of the right-hand side of (5-1) that is not symmetric in (x_{c_j-1},x_{c_j}) is the factor $x_{c_j}/(1-x_{c_j}y_j)$ (note that the contours are symmetric too), and applying T_{c_j-1} to it we read

$$\frac{qx_{c_j}}{1-x_{c_j}y_j}-\frac{x_{c_j-1}-qx_{c_j}}{x_{c_j-1}-x_{c_j}}\left(\frac{x_{c_j}}{1-x_{c_j}y_j}-\frac{x_{c_j-1}}{1-x_{c_j-1}y_j}\right)=\frac{x_{c_j-1}}{1-x_{c_j-1}y_j}\,\frac{1-qx_{c_j}y_j}{1-x_{c_j}y_j}.$$

This recovers the integrand of (5-1) with $c_j\mapsto c_j-1$ and completes the proof. □

Remark 5.6. While the integral representation (2-37) allowed for moving the contours through ∞, this is no longer true for (5-1), as the added factors $(y_i - s)$ create poles at ∞.

The main result of this section is a generalization of Proposition 5.2 to coloured nonrainbow compositions.

Theorem 5.7. *Let λ be a composition of weight $|\lambda| = n$ with partial sums $\sum_{i=1}^{k} \lambda_i = \ell_k,\ k \geq 1;\ \ell_0 := 0$. Further, let $\mu = 0^{\boldsymbol{m}^{(0)}} 1^{\boldsymbol{m}^{(1)}} 2^{\boldsymbol{m}^{(2)}} \cdots$ be a λ-coloured composition of length $\ell(\mu) = |\lambda| = n$ not consisting entirely of zeros, let $\mu^{\geq 1}$ be as in Definition 5.1 with inherited colouring $\lambda^{\geq 1}$ and $m := n - |\boldsymbol{m}^{(0)}|$ being the length of $\mu^{\geq 1}$, and let $c_1 < \cdots < c_\alpha$ be the colours of the (necessarily nonzero) parts of $\mu^{\geq 1}$. Finally, let $\mathfrak{m}_1, \dots, \mathfrak{m}_\alpha \geq 1$ be the number of parts of $\mu^{\geq 1}$ of colours $c_1, \dots, c_\alpha$, respectively, and denote $\mathfrak{m}[a, b] = \mathfrak{m}_a + \mathfrak{m}_{a+1} + \cdots + \mathfrak{m}_b$. Then*

$$
\begin{aligned}
&f_\mu(\lambda; x_1, \dots, x_n) = \\
&\quad \frac{(s^2; q)_{|\boldsymbol{m}^{(0)}|}}{(1 - s x_1) \cdots (1 - s x_n)} \frac{1}{(2\pi\sqrt{-1})^m} \\
&\qquad \times \oint \cdots \oint_{\text{around}\{x_j^{-1}\}} \prod_{1 \leq i < j \leq m} \frac{y_j - y_i}{y_j - q y_i} \\
&\qquad\quad \times \prod_{k=1}^{\alpha} \Biggl(\sum_{j=0}^{\mathfrak{m}_k} \frac{(-1)^j q^{\binom{\mathfrak{m}_k - j}{2}}}{(q;q)_j (q;q)_{\mathfrak{m}_k - j}} \prod_{p > \mathfrak{m}[1,k-1]}^{j + \mathfrak{m}[1,k-1]} \prod_{a > \ell_{c_k - 1}}^{n} \frac{1 - q x_a y_p}{1 - x_a y_p} \prod_{r > j + \mathfrak{m}[1,k-1]}^{\mathfrak{m}[1,k]} \prod_{b > \ell_{c_k}}^{n} \frac{1 - q x_b y_r}{1 - x_b y_r} \Biggr) \\
&\qquad\qquad \times f_{\mu^{\geq 1}}(\lambda^{\geq 1}; y_1^{-1}, \dots, y_m^{-1}) \prod_{i=1}^{m} \frac{(y_i - s) d y_i}{y_i^2}, \qquad (5\text{-}3)
\end{aligned}
$$

where (positively oriented) integration contours are chosen to encircle all points $\{x_j^{-1}\}_{j=1}^{n}$ and no other singularities of the integrand, or as q-nested closed simple curves with y_i-contour containing $q^{-1} \cdot (y_j\text{-contour})$ for all $i < j$, and all of the contours encircling $\{x_j^{-1}\}_{j=1}^{n}$. The contours can also be chosen to either encircle or not encircle the point 0.

The formula also holds in the column inhomogeneous setting under the assumption that in the 0*-th column* $(s_0, \xi_0) = (s, 1)$.

Remark 5.8. For $\lambda = (1, \dots, 1)$ and μ a rainbow composition, we have $\ell_k \equiv k,\ \alpha = m,\ \mathfrak{m}_1 = \cdots \mathfrak{m}_\alpha = 1$, $\mathfrak{m}[1, k] \equiv k$, and the middle line of (5-3) evaluates to

$$
(-1)\left(\frac{1}{1-q} - \frac{1}{1-q} \frac{1 - q x_{c_k} y_k}{1 - x_{c_k} y_k} \right) \prod_{b > c_k} \frac{1 - q x_b y_k}{1 - x_b y_k} = \frac{x_{c_k} y_k}{1 - x_{c_k} y_k} \prod_{b > c_k} \frac{1 - q x_b y_k}{1 - x_b y_k},
$$

thus reproducing (5-1).

Proof. The central role in the argument is played by the following:

Lemma 5.9. *For any positive integers $\mathfrak{k}$, $\mathfrak{l}$ and $\mathfrak{m}$ which satisfy $\mathfrak{k} + \mathfrak{m} < \mathfrak{l}$, and any symmetric function $\Phi(z_1, \dots, z_{\mathfrak{m}})$ that is holomorphic in a neighbourhood of the domain encircled by the integration contours,* *one has*

$$\sum_{\mathfrak{l}-1\ge i_1>\cdots>i_{\mathfrak{m}}\ge\mathfrak{k}} \oint\cdots\oint_{\mathrm{around}\{\mathfrak{x}_j\}} \prod_{i<j}\frac{z_j-z_i}{z_j-qz_i}\,\Phi(z_1,\dots,z_{\mathfrak{m}})\prod_{p=1}^{\mathfrak{m}}\biggl(\frac{1}{z_p-\mathfrak{x}_{i_p+1}}\prod_{j=1}^{i_p}\frac{qz_p-\mathfrak{x}_j}{z_p-\mathfrak{x}_j}dz_p\biggr)$$
$$=\oint\cdots\oint_{\mathrm{around}\{\mathfrak{x}_j\}} \prod_{i<j}\frac{z_j-z_i}{z_j-qz_i}\,\Phi(z_1,\dots,z_{\mathfrak{m}})$$
$$\times\biggl(\sum_{j=0}^{\mathfrak{m}}\frac{(-1)^j q^{\binom{\mathfrak{m}-j}{2}}}{(q;q)_j(q;q)_{\mathfrak{m}-j}}\prod_{p>0}^{j}\prod_{a=1}^{\mathfrak{l}}\frac{qz_p-\mathfrak{x}_a}{z_p-\mathfrak{x}_a}\prod_{r>j}^{\mathfrak{m}}\prod_{b=1}^{\mathfrak{k}}\frac{qz_r-\mathfrak{x}_b}{z_r-\mathfrak{x}_b}\biggr)\prod_{p=1}^{\mathfrak{m}}\frac{dz_p}{z_p}, \tag{5-4}$$

where (*positively oriented*) *integration contours are chosen* (*independently in the two sides of* (5-4)) *to encircle all points* $\{\mathfrak{x}_j\}_{j=1}^{\mathfrak{l}}\subset\mathbb{C}$ *and no other singularities of the integrand*, *or as* q*-nested closed simple curves with* z_i*-contour containing* $q^{-1}\cdot(z_j\text{-contour})$ *for all* $i<j$, *and all of the contours encircling* $\{\mathfrak{x}_j\}_{j=1}^{\mathfrak{l}}$. *The contours can also be chosen to either encircle or not encircle the point* 0.

Let us postpone the proof of Lemma 5.9 and use it for the proof of Theorem 5.7 first.

Let $\theta:\{1,\dots,n\}\to\{1,\dots,\ell(\lambda)\}$ be the unique monotone map such that

$$\theta^{-1}(k)=\{\ell_{k-1}+1,\dots,\ell_k\}\quad\text{for all}\quad 1\le k\le\ell(\lambda). \tag{5-5}$$

We can then use the colour merging relation (3-18) to obtain a formula for λ-coloured μ's from the formula (5-1) of Proposition 5.2 for the rainbow ones. This means that we need to sum the right-hand sides of (5-1), written for a rainbow composition $\tilde{\mu}$, over all $\tilde{\mu}$ with $\theta_*(\tilde{\mu})=\mu$.

Such a summation can be performed in two steps. At the first step we choose, for each colour c_k represented in $\mu^{\ge1}$, $1\le k\le\alpha$, the $\mathfrak{m}_k$ colours in $\theta^{-1}(c_k)$ that are represented in $\tilde{\mu}^{\ge1}$. At the second step we choose, for each k, $1\le k\le\alpha$, different assignments of the chosen $\mathfrak{m}_k$ colours in $\theta^{-1}(c_k)$ to the $\mathfrak{m}_k$ parts of $\mu^{\ge1}$ that have colour c_k. The summation of the second step is exactly the colour merging applied to $f_{\tilde{\mu}^{\ge1}}$ (no other factors of the integrand of (5-1) depend on the choices in the second step), and (3-18) shows that the second step summation results in replacing $f_{\tilde{\mu}^{\ge1}}$ in the integrand by $f_{\mu^{\ge1}}$.

Returning to the first step summation, we see that $f_{\mu^{\ge1}}$ in the integrand is now independent of the choices involved, and we can focus on the rest of the integrand. For each colour c_k, $1\le k\le\alpha$, we are choosing $c_k^{(1)}<\cdots<c_k^{(\mathfrak{m}_k)}$ in $\theta^{-1}(c_k)$, or, according to (5-5), $\ell_{c_k-1}<c_k^{(1)}<\cdots<c_k^{(\mathfrak{m}_k)}\le\ell_{c_k}$. Now the sum over such $\mathfrak{m}_k$-tuple of indices can be computed, for each $1\le k\le\alpha$, using Lemma 5.9, where one needs to make substitutions

$$\mathfrak{m}\mapsto\mathfrak{m}_k,\quad \mathfrak{k}\mapsto(n-\ell_{c_k}),\quad \mathfrak{l}\mapsto(n-\ell_{c_k-1}),\quad (z_1,\dots,z_{\mathfrak{m}})\mapsto(y_{\mathfrak{m}[1,k-1]+1},\dots,y_{\mathfrak{m}[1,k]}),$$
$$(\mathfrak{x}_1,\mathfrak{x}_2,\dots\mathfrak{x}_{\mathfrak{l}})\mapsto(x_n^{-1},x_{n-1}^{-1},\dots,x_{\ell_{c_k-1}+1}^{-1}),\quad (\mathfrak{x}_1,\mathfrak{x}_2,\dots\mathfrak{x}_{\mathfrak{k}})\mapsto(x_n^{-1},x_{n-1}^{-1},\dots,x_{\ell_{c_k}+1}^{-1}),$$
$$(\mathfrak{l}>i_1>\cdots>i_{\mathfrak{m}}\ge\mathfrak{k})\mapsto(n-\ell_{c_k-1}>n-c_k^{(1)}>\cdots>n-c_k^{(\mathfrak{m}_k)}\ge n-\ell_{c_k}),$$

and collect the factors that do not correspond to the ones in the left-hand side of (5-4) into a $\Phi(z_1,\dots,z_{\mathfrak{m}})$. The holomorphicity of $\Phi(z_1,\dots,z_{\mathfrak{m}})$ in a neighbourhood of $\{\mathfrak{x}_j\}_{j=1}^{\mathfrak{l}}$ is readily visible; and the fact that it is symmetric in $z_1, \dots, z_{\mathfrak{m}}$ follows from the fact that $f_{\mu^{\geq 1}}(y_1^{-1}, \dots, y_m^{-1})$, which enters $\Phi(z_1, \dots, z_{\mathfrak{m}})$ as the only factor that is not manifestly symmetric, is symmetric in

$$(y_{\mathfrak{m}[1,k-1]+1}, \dots, y_{\mathfrak{m}[1,k]})$$

thanks to the commutativity of the $\mathcal{C}_i(x)$ operators (for a fixed i and varying x) in the definition (2-41) of f-functions for coloured compositions.

Comparing the right-hand sides of (5-3) and (5-4), we see that this completes the proof of Theorem 5.7 modulo the proof of Lemma 5.9. □

Proof of Lemma 5.9. While a direct book-keeping of residues of the two sides of (5-4) might be possible, we will use the theory of Hall–Littlewood processes as a shortcut, with the work [Borodin et al. 2016a, Section 2] as our main reference; a more detailed description can be found in [Borodin and Corwin 2014, Section 2]. See also Remark 5.10 on the origin of the argument given below.

Consider an ascending Hall–Littlewood process with weights on sequences of partitions $\lambda^{(1)}, \dots, \lambda^{(\mathfrak{n})}$ proportional to

$$P_{\lambda(1)}(\mathfrak{x}_1)P_{\lambda(2)/\lambda(1)}(\mathfrak{x}_2)P_{\lambda(\mathfrak{n})/\lambda(\mathfrak{n}-1)}(\mathfrak{x}_{\mathfrak{n}})Q_{\lambda(\mathfrak{n})}(\rho), \qquad \ell(\lambda(k)) \leq k, \quad 1 \leq k \leq \mathfrak{n}, \tag{5-6}$$

with, generally speaking, complex parameters $\{\mathfrak{x}_i\}_{i=1}^{\mathfrak{n}}$, ρ being the specialization of the algebra of symmetric functions into a sequence of variables $(b_1, b_2, \cdots)$, and P_* and Q_* being the Hall–Littlewood symmetric functions.

Our argument is based on [Borodin et al. 2016a, Proposition 2.2], see also [Borodin and Corwin 2014, Proposition 2.2.14], which says that for any $\mathfrak{n} \geq m_1 \geq \ldots \geq m_n \geq 1$, one has

$$\begin{aligned}\mathbb{E}_{HL}&(q^{m_1-\ell(\lambda(m_1))} \cdots q^{m_n-\ell(\lambda(m_n))}) \\ &= \frac{q^{\binom{n}{2}}}{(2\pi\sqrt{-1})^n} \oint \cdots \oint \prod_{i<j} \frac{z_j - z_i}{z_j - qz_i} \prod_{l=1}^{n} \biggl(\prod_{j \geq 1} \frac{1 - z_l b_j}{1 - qz_l b_j} \prod_{i=1}^{m_l} \frac{qz_l - \mathfrak{x}_i}{z_l - \mathfrak{x}_i} \frac{dz_l}{z_l} \biggr),\end{aligned} \tag{5-7}$$

where (positively oriented) z_j-contours are such that they surround $\{\mathfrak{x}_j\}_{j=1}^{m_1}$ and 0, and they are also q-nested in the sense that z_i-contour contains $q \cdot (z_j\text{-contour})$ for all $i < j$; no other poles are taken into account.

Let us fix $1 \leq \mathfrak{k} < \mathfrak{l} \leq \mathfrak{n}$, and consider the sum

$$\begin{aligned}\sum_{\mathfrak{k} \leq i_{\mathfrak{m}} < i_{\mathfrak{m}-1} < \cdots < i_1 < \mathfrak{l}} \prod_{p=1}^{m} &\frac{q^{i_p - \ell(\lambda(i_p))} - q^{i_p+1-\ell(\lambda(i_p+1))}}{1-q} \\ &= q^{\binom{\mathfrak{m}}{2}} \cdot q^{m(\mathfrak{k}-\ell(\lambda(\mathfrak{k})))} \cdot \binom{\mathfrak{l} - \ell(\lambda(\mathfrak{l})) - \mathfrak{k} + \ell(\lambda(\mathfrak{k}))}{\mathfrak{m}}_q .\end{aligned} \tag{5-8}$$

The equality between the two sides of (5-8) is a (nonobvious) special case of Corollary 3.7. More exactly, choosing $Q = q$, $\alpha_i \equiv 1$, $|\beta| = \mathfrak{m}$, and reversing the order of indices of α_i's and β_i's, turns (3-16) into

$$q^{-\binom{\mathfrak{m}}{2}} \cdot \sum_{\substack{\beta_1, \dots, \beta_n \in \{0,1\} \\ \beta_1 + \cdots + \beta_n = \mathfrak{m}}} \prod_{i=1}^{n} q^{\beta_i \cdot (i-1)} = \binom{n}{\mathfrak{m}}_q . \tag{5-9}$$

Let $1 \le k_{\mathfrak{m}} < k_{\mathfrak{m}-1} < \cdots < k_1 \le n$ be the set of index values k for which $\beta_k = 1$ in (5-9). Observe that

$$\frac{q^{j-\ell(\lambda(j))} - q^{j+1-\ell(\lambda(j+1))}}{1-q} = \begin{cases} q^{j-\ell(\lambda(j))} = q^{\mathfrak{k}-\ell(\lambda(\mathfrak{k}))} \cdot q^{j-\ell(\lambda(j))-\mathfrak{k}+\ell(\lambda(\mathfrak{k}))}, & \ell(\lambda(j)) = \ell(\lambda(j+1)), \\ 0, & \text{otherwise}, \end{cases}$$

where “otherwise” refers to the only possible alternative $\ell(\lambda(j)) + 1 = \ell(\lambda(j+1))$.

Set $n = \mathfrak{l} - \ell(\lambda(\mathfrak{l})) - \mathfrak{k} + \ell(\lambda(\mathfrak{k}))$, and let $\mathfrak{k} \le j_1 < j_2 < \cdots < j_n < \mathfrak{l}$ be all the values of $j \in [\mathfrak{k}, \mathfrak{l})$ such that $\ell(\lambda(j)) = \ell(\lambda(j+1))$; equivalently $(j+1-\ell(\lambda(j+1))) - (j - \ell(\lambda(j))$ equals 1 rather than 0. These are all possible values that summation indices $i_1 > \cdots > i_{\mathfrak{m}}$ in the left-hand side of (5-8) can take to produce a nonzero term; let $j_{l_1} > \cdots > j_{l_{\mathfrak{m}}}$ be the corresponding choices. Then matching $(k_1, \ldots, k_{\mathfrak{m}}) \equiv (l_1, \ldots, l_{\mathfrak{m}})$ establishes the equivalence of (5-8) and (5-9), thus proving (5-8).

Our next step is to compute the averages, with respect to the ascending Hall–Littlewood process, of both sides of (5-8) using (5-7).

The left-hand side of (5-8) is a simple linear combination of those from (5-7). Moving that linear combination inside the integrand and observing that

$$\frac{1}{1-q}\left(\prod_{i=1}^{i_p} \frac{qz_p - \mathfrak{x}_i}{z_p - \mathfrak{x}_i} - \prod_{i=1}^{i_p+1} \frac{qz_p - \mathfrak{x}_i}{z_p - \mathfrak{x}_i}\right) = \frac{z_p}{z_p - \mathfrak{x}_{i_p+1}} \prod_{i=1}^{i_p} \frac{qz_p - \mathfrak{x}_i}{z_p - \mathfrak{x}_i},$$

we obtain

$$\mathbb{E}_{HL}\left(\sum_{\mathfrak{k} \le i_{\mathfrak{m}} < i_{\mathfrak{m}-1} < \cdots < i_1 < \mathfrak{l}} \prod_{p=1}^{m} \frac{q^{i_p - \ell(\lambda(i_p))} - q^{i_p+1-\ell(\lambda(i_p+1))}}{1-q}\right)$$

$$= \sum_{\mathfrak{k} \le i_{\mathfrak{m}} < i_{\mathfrak{m}-1} < \cdots < i_1 < \mathfrak{l}} \frac{q^{\binom{\mathfrak{m}}{2}}}{(2\pi\sqrt{-1})^{\mathfrak{m}}} \oint \cdots \oint \prod_{i<j} \frac{z_j - z_i}{z_j - qz_i}$$

$$\times \prod_{p=1}^{\mathfrak{m}} \left(\prod_{j \ge 1} \frac{1 - z_p b_j}{1 - qz_p b_j} \frac{1}{z_p - \mathfrak{x}_{i_p+1}} \prod_{i=1}^{i_p} \frac{qz_p - \mathfrak{x}_i}{z_p - \mathfrak{x}_i}\, dz_p \right), \tag{5-10}$$

with the same integration contours as in (5-7).

Note that 0 is no longer a potential singularity of the integrand; thus, the contour may or may not contain it. Let us also explain why the presence or absence of the potential poles at $z_i = q^{-1} z_j$ does not affect the value of the integral. Generally speaking, the integral with q-nested contours is equal to the sum of residues at $z_p = q^{-k_p} \mathfrak{x}_{l_p}$, $1 \le p \le \mathfrak{m}$, for certain values of $k_p \ge 0$ and $l_p \ge 1$, that arises by sequential evaluation of residues inside the $z_{\mathfrak{m}}, z_{\mathfrak{m}-1}, \ldots, z_1$-contours in that order. Let p^* be the maximal index such that $k_{p^*} > 0$. Since the $z_{\mathfrak{m}}$-contour encircles only the poles at $\mathfrak{x}_j$'s, we must have $p^* \le \mathfrak{m} - 1$. Also, since this pole must have come from a denominator factor $z_{j^*} - qz_{p^*}$ with $j^* > p^*$, due to the maximality of p^*, the pole must be at $z_{p^*} = q^{-1}\mathfrak{x}_{l^*}$ with $1 \le l^* \le i_{j^*}$. But the integrand contains the factor $(qz_{p^*} - \mathfrak{x}_l^*)$, which will turn the residue into 0 (note that we need the fact that $i_{p^*} \ge i_{j^*}$ to guarantee the presence of this factor). We conclude that the q-nestedness is irrelevant for the value of the integral.

Let us proceed to computing the Hall–Littlewood expectation of the right-hand side of (5-8).

Using the q-binomial theorem

$$\sum_{j\ge 0}\frac{(a;q)_j}{(q;q)_j}z^j=\frac{(az;q)_\infty}{(z;q)_\infty}=\left(1-\frac{z}{q}\right)\cdots\left(1-\frac{z}{q^{\mathfrak{m}}}\right)\qquad\text{for } a=q^{-\mathfrak{m}},$$

we obtain

$$\binom{A}{\mathfrak{m}}_q=\frac{(1-q^A)(1-q^{A-1})\cdots(1-q^{A-\mathfrak{m}+1})}{(q;q)_{\mathfrak{m}}}=\sum_{j=0}^{\mathfrak{m}}\frac{(q^{-\mathfrak{m}};q)_j}{(q;q)_j(q;q)_{\mathfrak{m}}}q^{(A+1)j}=\sum_{j=0}^{\mathfrak{m}}\frac{(-1)^jq^{(A+1)j-\mathfrak{m}j+\binom{j}{2}}}{(q;q)_j(q;q)_{\mathfrak{m}-j}}.$$

Hence, the right-hand side of (5-8) can be written as

$$q^{\binom{\mathfrak{m}}{2}}\cdot q^{m(\mathfrak{k}-\ell(\lambda(\mathfrak{k})))}\cdot\binom{\mathfrak{l}-\ell(\lambda(\mathfrak{l}))-\mathfrak{k}+\ell(\lambda(\mathfrak{k}))}{\mathfrak{m}}_q=\sum_{j=0}^{\mathfrak{m}}\frac{q^{\binom{\mathfrak{m}-j}{2}}}{(q;q)_j(q;q)_{\mathfrak{m}-j}}q^{j(\mathfrak{l}-\ell(\lambda(\mathfrak{l})))+(\mathfrak{m}-j)(\mathfrak{k}-\ell(\mathfrak{k}))},$$

where for powers of q we used $\binom{\mathfrak{m}}{2}-\mathfrak{m}j+j+\binom{j}{2}=\binom{\mathfrak{m}-j}{2}$. Employing (5-7), we now obtain

$$\begin{aligned}\mathbb{E}_{HL}\left(q^{\binom{\mathfrak{m}}{2}}\cdot q^{m(\mathfrak{k}-\ell(\lambda(\mathfrak{k})))}\cdot\binom{\mathfrak{l}-\ell(\lambda(\mathfrak{l}))-\mathfrak{k}+\ell(\lambda(\mathfrak{k}))}{\mathfrak{m}}_q\right)&\\ =\sum_{j=0}^{\mathfrak{m}}\frac{q^{\binom{\mathfrak{m}-j}{2}}}{(q;q)_j(q;q)_{\mathfrak{m}-j}}\frac{q^{\binom{\mathfrak{m}}{2}}}{(2\pi\sqrt{-1})^{\mathfrak{m}}}\oint\cdots\oint\prod_{i<j}\frac{z_j-z_i}{z_j-qz_i}&\\ \times\prod_{p>0}^{j}\prod_{a=1}^{\mathfrak{l}}\frac{qz_p-\mathfrak{x}_a}{z_p-\mathfrak{x}_a}\prod_{r>j}^{\mathfrak{m}}\prod_{b=1}^{\mathfrak{k}}\frac{qz_r-\mathfrak{x}_b}{z_r-\mathfrak{x}_b}\prod_{p=1}^{\mathfrak{m}}\prod_{j\ge 1}\frac{1-z_pb_j}{1-qz_pb_j}\frac{dz_p}{z_p},&\qquad(5\text{-}11)\end{aligned}$$

where the integration contours are as for (5-7).

Since the two of sides of (5-8) are equal, the right-hand sides of (5-10) and (5-11) are also equal. This is literally the desired statement of Lemma 5.9, equation (5-4), with a specific choice of

$$\Phi(z_1,\dots,z_{\mathfrak{m}})=\phi(z_1)\cdots\phi(z_{\mathfrak{m}}),\qquad\phi(z)=\text{const}\prod_{j\ge 1}\frac{1-zb_j}{1-qzb_j},\qquad(5\text{-}12)$$

and a specific choice of contours for the right-hand side. The equality (of the right-hand sides of (5-10) and (5-11)) has been thus proven for generic $\{\mathfrak{x}_i\}$ and $\{b_j\}$, although certain inequalities on them are needed to make sure that the weights (5-6) are summable (see [Borodin et al. 2016a, Section 2] for details). However, the equality itself is an identity of finite sums (of residues), and the restrictions on $\{\mathfrak{x}_i\}$ and $\{b_j\}$ can thus be removed by analytic continuation.

The two sides of (5-4) computed as (finite) sums of residues are linear combinations of values of the function $\Phi(z_1,\dots,z_{\mathfrak{m}})$ at $\mathfrak{m}$-tuples of distinct (because of $\prod_{i<j}(z_j-z_i)$ in the integrand) points from the list of possible singularities, consisting of $\{\mathfrak{x}_i\}$, their q-multiples, and 0. Values of $\Phi(z_1,\dots,z_{\mathfrak{m}})$ with permuted z_j's are equal (due to the symmetry of Φ) and can be grouped together; their coefficients are certain explicit rational functions of $\{\mathfrak{x}_i\}$ and q. Fix an $\mathfrak{m}$-tuple of distinct possible poles. Using the freedom in the choice of $\{b_j\}$ and the constant prefactor in (5-12), we can make $\phi(z)$ to be arbitrarily close to 1 at the points of the chosen $\mathfrak{m}$-tuple, and arbitrarily close to 0 at all the other potential singularities (we

are assuming that $\{\mathfrak{x}_i\}$ are generic, we only need to prove an identity between rational functions in them).[8] This will lead to Φ being close to 1 at the chosen $\mathfrak{m}$-tuple and its permutations, and close to 0 at all other possible $\mathfrak{m}$-tuples. Since we already proved the equality of the two sides of (5-4) for all Φ's of the form (5-12), this implies the equality of the coefficients of Φ evaluated at the chosen $\mathfrak{m}$-tuple (together its permutations) in residue expansions of the two sides of (5-4). This proves (5-4) for arbitrary Φ, but still with the particular choice of integration contours in the right-hand side, which are q-nested and include 0.

However, we already know that the declared freedom for the choice of contours is valid for the left-hand side; see the argument after (5-10). This implies that the values of Φ at $\mathfrak{m}$-tuples that include either 0 or a nontrivial q-multiple of a $\mathfrak{x}_i$ do not contribute to the left-hand side. Since we already know that such coefficients are the same on both sides, these values must not contribute to the right-hand side either, which means that we can use the same freedom of contours in the right-hand side without changing the value of the integral. □

Remark 5.10. Let us comment on the origin of our proof of Lemma 5.9 that might have looked somewhat cryptic. If, in the setting of Theorem 5.7, μ has no zero parts and only one colour, then f_μ is colour-blind, according to (2-29). Theorem 4.7 then implies that it is given by an average of a q-moment type observable (4-14) over a colour-blind stochastic vertex model. The height function of the colour-blind stochastic vertex models can be interpreted, along any down-right path in a quadrant (which includes horizontal lines), as lengths of partitions distributed according to Hall–Littlewood processes; this was the main result of [Borodin et al. 2016a]. Thus, we get an expression for f_μ in the form of an average over a Hall–Littlewood process.

On the other hand, the symmetrization of colours in (2-29) can also be taken after the computation of the expectation, with respect to a rainbow coloured model, in the left-hand side of (4-15). As was shown in [Borodin and Wheeler 2018, Chapter 10], the distribution of coloured height functions at a single observation point can also be described via lengths of partitions in a Hall–Littlewood process, and this is where the computation of that expectation can take place. Once the corresponding average over the Hall–Littlewood process is computed, one can perform its colour symmetrization.

The two resulting expressions must be the same - the operations of averaging over our measure and symmetrizing over colours commute. Understanding the reason for that in the language of the Hall–Littlewood processes is not too difficult, this is essentially (5-8). Rewriting what it means in terms of integral representations for averages of Hall–Littlewood observables results in a general identity for symmetric functions, which is exactly our Lemma 5.9.

6. Observables of stochastic lattice models

The purpose of this section is to combine the results of Sections 4-5 in order to obtain integral representations for averages of the observables introduced in Definition 4.5, as well as to explore corollaries thereof.

[8] Indeed, $\phi(z) = \mathrm{const} \cdot \psi(x)/\psi(qz)$ with ψ being an arbitrary polynomial that can be chosen to approximate any values at any finite set of points; the constant in front is needed to control $\phi(0)$.

6A. ***The main result.*** Recalling graphical interpretation of Definition 4.4, we consider a stochastic lattice model in the quadrant $\mathbb{Z}_{\geq 1} \times \mathbb{Z}_{\geq 1}$, with vertex weights given by $L^{\mathrm{stoch}}_{x_i}$ in (2-11) (see also (2-13) and Table 3) for the vertices in row i from the bottom. The boundary conditions are as follows: No paths (equivalently, only paths of colour 0) enter the quadrant through its bottom boundary; and along the left boundary we have a single path entering at every row, with the bottom λ_1 paths having colour 1, next λ_2 paths having colour 2, and so on. We focus on the state of the model between row n and row $n+1$; that is, we record the locations where the paths of colour $1, 2, \dots$ exit the n-th row upwards as a λ-coloured composition ν, where λ is the composition with parts $\lambda_1, \lambda_2, \dots$, and we truncate this sequence so that $|\lambda| = n$.[9] The partial sums of λ are denoted as $\sum_{i=1}^{k} \lambda_i = \ell_k,\ k \geq 1;\ \ell_0 := 0$.

The model is also allowed to have column inhomogeneities $\{s_j, \xi_j\}_{j \geq 1}$, which replace the (s, x_i) parameters in the L^{stoch}-weight of the vertex in row i and column j by $(s_j, \xi_j x_i)$; see Section 3A. For convenience, we assume that the number of column inhomogeneities is finite. (For our results this assumption is not restrictive as the state of the model far enough to the right will not play any role, and thus, due to stochasticity, the column inhomogeneities there can be chosen freely.)

Also recall the observables $\mathcal{O}_\mu$ of Definition 4.5, defined for any λ-coloured composition $\mu = 0^{\boldsymbol{m}^{(0)}} 1^{\boldsymbol{m}^{(1)}} 2^{\boldsymbol{m}^{(2)}} \cdots$, that take values

$$\mathcal{O}_\mu(\nu) = \prod_{x \geq 1} \prod_{i \geq 1} q^{m_i^{(x)} H_{>i}^{\nu/\mu}(x+1)} \binom{H_i^{\nu/\mu}(x+1)}{m_i^{(x)}}_q$$

$$= \prod_{x \geq 1} \prod_{i \geq 1} \frac{(q^{H_{>i}^{\nu/\mu}(x+1)} - q^{H_{\geq i}^{\nu/\mu}(x+1)})(q^{H_{>i}^{\nu/\mu}(x+1)} - q^{H_{\geq i}^{\nu/\mu}(x+1)-1}) \cdots (q^{H_{>i}^{\nu/\mu}(x+1)} - q^{H_{\geq i}^{\nu/\mu}(x+1)-m_i^{(x)}+1})}{(q;q)_{m_i^{(x)}}} \tag{6-1}$$

given in terms of the coloured height functions (3-5).

For any λ-coloured μ, we make a new coloured composition $\mu^{\geq 1}$ that consists of its nonzero parts coloured in the same way with the labelling composition of $\mu^{\geq 1}$ denoted by $\lambda^{\geq 1}$; see Definition 5.1. Clearly, given the colouring λ of μ, μ is uniquely reconstructed from $\mu^{\geq 1}$. We use the notation $m := n - |\boldsymbol{m}^{(0)}|$ for the length of $\mu^{\geq 1}$, denote by $c_1 < \cdots < c_\alpha$ the colours of parts of $\mu^{\geq 1}$, and denote by $\mathfrak{m}_1, \dots, \mathfrak{m}_\alpha \geq 1$ the number of parts of $\mu^{\geq 1}$ of colours $c_1, \dots, c_\alpha$, respectively. Set $\mathfrak{m}[a,b] = \mathfrak{m}_a + \mathfrak{m}_{a+1} + \cdots + \mathfrak{m}_b$.

We can now state the main result of this paper.

Theorem 6.1. *With the above notation, we have*

$$\mathbb{E}\,\mathcal{O}_\mu = \frac{q^{\sum_{u \geq 1} \sum_{i > j} m_i^{(u)} m_j^{(u)}}}{\prod_{j \geq 1} (s^2; q)_{|\boldsymbol{m}^{(j)}|}} \frac{1}{(2\pi\sqrt{-1})^m} \oint \cdots \oint_{\mathrm{around}\{x_j^{-1}\}} \prod_{1 \leq i < j \leq m} \frac{y_j - y_i}{y_j - q y_i}$$

$$\times \prod_{k=1}^{\alpha} \left(\sum_{j=0}^{\mathfrak{m}_k} \frac{(-1)^j q^{\binom{\mathfrak{m}_k - j}{2}}}{(q;q)_j (q;q)_{\mathfrak{m}_k - j}} \prod_{p > \mathfrak{m}[1,k-1]}^{j + \mathfrak{m}[1,k-1]} \prod_{a > \ell_{c_k - 1}}^{n} \frac{1 - q x_a y_p}{1 - x_a y_p} \prod_{r > j + \mathfrak{m}[1,k-1]}^{\mathfrak{m}[1,k]} \prod_{b > \ell_{c_k}}^{n} \frac{1 - q x_b y_r}{1 - x_b y_r} \right)$$

$$\times f^{\mathrm{stoch}}_{\mu^{\geq 1}}(\lambda^{\geq 1}; y_1^{-1}, \dots, y_m^{-1}) \prod_{i=1}^{m} \frac{(y_i - s)\,dy_i}{y_i^2}, \tag{6-2}$$

[9] Without loss of generality, for convenience of notation, we assume incoming paths in rows n and $n+1$ have different colours.

where (positively oriented) integration contours are chosen to encircle all points $\{x_j^{-1}\}_{j=1}^n$ *and no other singularities of the integrand, or as q-nested closed simple curves with y_i-contour containing $q^{-1}\cdot(y_j$-contour$)$ for all $i<j$, and all of the contours encircling* $\{x_j^{-1}\}_{j=1}^n$. *The contours can also be chosen to either encircle or not encircle the point* 0.

The formula also holds in the column inhomogeneous setting under the assumption that $(s_x,\xi_x)=(s,1)$ *for any* $x\geq 1$ *such that* $|\boldsymbol{m}^{(x)}|>0$.

Proof. If the parameters satisfy the inequalities (4-1), then the statement is a substitution of Theorem 5.7 into Theorem 4.7. Observe that $\mathcal{O}_\mu$ is independent of the state of the model to the right of the maximal coordinate of μ. Thus, both sides of (6-2) are actually rational functions in x_i's, and the extra assumption (4-1) can thus be removed by analytic continuation. □

Remark 6.2. The factor $\prod_{j\geq 1}(s^2;q)^{-1}_{|\boldsymbol{m}^{(j)}|}$ in the right-hand side of (6-2) does not make the expression singular in the finite spin situation $s^2=q^{-J}$, $J=1,2,\dots$, because the same factor appears in the numerator when one writes $f^{\text{stoch}}_{\mu^{\geq 1}}=(-s)^{|\mu|}f_{\mu^{\geq 1}}$ explicitly by taking the factorization (2-30), where this factor is manifest, and acting on it by difference operators (2-31)–(2-32) and colour merging (3-18) as needed.

6B. ***The colour-blind case.*** Let us see how Theorem 6.1 works in the colour-blind situation. The observables then simplify to (4-14), and we obtain:

Corollary 6.3. *In the colour-blind case* $\lambda=(n)$, *with ordered coordinates* $\theta_1\geq\ldots\geq\theta_m\geq 1$ *of* $\mu^{\geq 1}$ *and no column inhomogeneities, we have*

$$\mathbb{E}\Big[(1-q^{H^\nu(\theta_1+1)})(1-q^{H^\nu(\theta_2+1)-1})\cdots(1-q^{H^\nu(\theta_m+1)-m+1})\Big]$$
$$=\frac{(-1)^m(-s)^{|\theta|}}{(2\pi\sqrt{-1})^m}\oint\cdots\oint_{\text{around}\{x_j^{-1}\}}\prod_{1\leq i<j\leq m}\frac{y_i-y_j}{y_i-qy_j}\prod_{p=1}^{m}\left(\frac{1-sy_p}{y_p-s}\right)^{\theta_i}\prod_{a=1}^{n}\frac{1-qx_ay_p}{1-x_ay_p}\frac{dy_p}{y_p},\tag{6-3}$$

where (positively oriented) integration contours are chosen to encircle all points $\{x_j^{-1}\}_{j=1}^n$ *and no other singularities of the integrand, and* $H^\nu(x)=\#\{i\in\{1,\dots,n\}:\nu_i\geq x\}$.

Remark 6.4. Equation (6-3) is readily seen to coincide with [Borodin and Petrov 2017, Lemma 9.10], which is essentially equivalent to the integral representation of the most general multipoint moments for the colourless stochastic vertex model along a single line.

It is not hard to extend this result, with a very similar proof to the one given below, to the column inhomogeneous case under the condition that $(s_x,\xi_x)=(s,1)$ for any $x\geq 1$ such that no parts of μ are equal to x. However, [Borodin and Petrov 2018a, Lemma 9.11] gives such an extension to the fully column inhomogeneous situation. It remains unclear how to achieve this level of generality in the coloured case.

Remark 6.5. While the freedom in choosing the contours as q-nested is still there (it can be checked directly as in the proof of Lemma 5.9 or carried over from Theorem 6.1), the contours cannot include the point 0 anymore, and its inclusion (together with q-nestedness) would actually change the q-shifted moments in the left-hand side into unshifted moments of [Borodin and Petrov 2017, Theorem 9.8]; see [Borodin and Petrov 2017, Section 9] for a detailed explanation of that transition.

Proof of Corollary 6.3. In the colour-blind case $\lambda = (n)$, using (4-11), (2-29) and (4-14), we write (6-2) as

$$\mathbb{E}\left[\frac{(1-q^{H^\nu(\theta_1+1)})(1-q^{H^\nu(\theta_2+1)-1})\cdots(1-q^{H^\nu(\theta_m+1)-m+1})}{\prod_{j\geq 1}(q;q)_{\mathrm{mult}_j(\theta)}}\right]$$
$$= \frac{(-s)^{|\theta|}}{\prod_{j\geq 1}(s^2;q)_{\mathrm{mult}_j(\theta)}}\frac{1}{(2\pi\sqrt{-1})^m}\oint\cdots\oint_{\mathrm{around}\{x_j^{-1}\}}\prod_{1\leq i<j\leq m}\frac{y_j-y_i}{y_j-qy_i}$$
$$\times\sum_{j=0}^{m}\left(\frac{(-1)^j q^{\binom{m-j}{2}}}{(q;q)_j(q;q)_{m-j}}\prod_{p=1}^{j}\prod_{a=1}^{n}\frac{1-qx_ay_p}{1-x_ay_p}\right)\mathsf{F}^{\mathsf{c}}_\theta(y_1^{-1},\dots,y_m^{-1})\prod_{i=1}^{m}\frac{(y_i-s)dy_i}{y_i^2}, \tag{6-4}$$

with contours around $\{x_j^{-1}\}$ and no other singularities. Observe that if the summation index j in the integrand above takes any value $j < m$, then the integrand, viewed as a function in y_m, has no singularities at $\{x_*^{-1}\}$, and the integral vanishes. Hence, we can set $j = m$.[10]

Further we write

$$\prod_{i<j}\frac{y_j-y_i}{y_j-qy_i}=\prod_{i\neq j}\frac{y_j-y_i}{y_j-qy_i}\cdot\prod_{i>j}\frac{y_j-qy_i}{y_j-y_i},$$

note that the first factor in the right-hand side is symmetric in y_i's, and the same is true about all other parts of the integrand. Hence, we can sum over the second factor over all permutations of the y_i's, which yields $(q;q)_m/(1-q)^m$ in the integrand and multiplies the value of the integral by $m!$.

Finally, from [Borodin and Petrov 2017, Theorem 4.12] we read

$$\mathsf{F}^{\mathsf{c}}_\theta(y_1^{-1},\dots,y_m^{-1})=\frac{\prod_{j\geq 1}(s^2;q)_{\mathrm{mult}_j(\theta)}}{\prod_{j\geq 1}(q;q)_{\mathrm{mult}_j(\theta)}}\frac{(1-q)^m}{\prod_{i=1}^m(1-sy_i^{-1})}\sum_{\sigma\in\mathfrak{S}_m}\sigma\left(\prod_{i<j}\frac{y_i^{-1}-qy_j^{-1}}{y_i^{-1}-y_j^{-1}}\prod_{i=1}^{m}\left(\frac{y_i^{-1}-s}{1-sy_i^{-1}}\right)^{\theta_i}\right),$$

where the sum is over all permutations σ in the symmetric group $\mathfrak{S}_m$ on m symbols, and σ's permute the variables $y_1,\dots,y_m$ in the expression that they are applied to. Since the rest of the integrand is symmetric in the y_i's, we can remove the sum over permutation leaving only the term with $\sigma = \mathrm{id}$, and divide the integral by $m!$.

Implementing the above transformation and cancelling common factors yields (6-3). □

6C. ***Duality.*** One convenient feature of the integral representation (6-2) for $\mathbb{E}\mathcal{O}_\mu$ is that the dependence on values of coordinates of μ is concentrated in the factor $f_{\mu^{\geq 1}}(\lambda^{\geq 1}; y_1^{-1},\dots,y_m^{-1})$ of the integrand. This allows us to easily derive certain difference equations that $\mathbb{E}\mathcal{O}_\mu$ must satisfy. Let us see how this works.

The skew-Cauchy identity (2-35) was stated for the rainbow case but is readily extended to the colour-merged case with the help of Proposition 3.8 (in fact, this colour merging argument was already used in the proof of Proposition 4.3 for G_*'s replaced by their limits $\mathcal{G}_*$'s). In particular, it implies

$$\sum_{\mu^{\geq 1}\text{ is }\lambda^{\geq 1}\text{ coloured}} f_{\mu^{\geq 1}}(\lambda^{\geq 1}; y_1^{-1},\dots,y_m^{-1})G_{\mu^{\geq 1}/\varkappa}((qX)^{-1})=\prod_{i=1}^{m}\frac{1-y_iX}{1-qy_iX}\cdot f_\varkappa(\lambda^{\geq 1}; y_1^{-1},\dots,y_m^{-1}),$$

[10] It is at this moment that we lose the freedom to have 0 inside the contours, as the terms that we remove may have nontrivial residues at 0; see Remark 6.5.

where $\varkappa$ is an arbitrary $\lambda^{\geq 1}$-coloured composition, X is a complex parameter. Multiplying both sides by $(-s)^{|\varkappa|} = (-s)^{|\varkappa|-|\mu|}(-s)^{|\mu|}$, we can also replace f by f^{stoch} and G by G^{stoch} in this identity, where $G^{\text{stoch}}_{\mu^{\geq 1}/\varkappa} = (-s)^{|\varkappa|-|\mu|} G_{\mu^{\geq 1}/\varkappa}$ is a stochastic kernel with matrix elements built from stochastic M^{stoch}-weights (2-12).

Observe that the prefactor is of the form of the exact inverse of the factors in the integrand of (6-2) that depend on x_i's. Hence, if we compute the expectation $\mathbb{E}_{n+1}\mathcal{O}_\mu$ along the $(n+1)$-st row of the stochastic vertex model, with left entering colour $(n+1)$ and rapidity X used in that row, and subsequently sum against the stochastic kernel $G^{\text{stoch}}_{\mu^{\geq 1}/\varkappa}((qX)^{-1})$, the result will coincide (subject to certain convergence conditions that we are ignoring here) with the expectation of $\mathbb{E}_n\mathcal{O}_\mu$ computed along the n-th row:

$$\sum_{\mu^{\geq 1}} \mathbb{E}_{n+1}\mathcal{O}_\mu \cdot G^{\text{stoch}}_{\mu^{\geq 1}/\varkappa}((qX)^{-1}) = \mathbb{E}_n\mathcal{O}_\varkappa. \tag{6-5}$$

As the transition from row n to row $n+1$ is also realized by a stochastic kernel, we observe a *duality* of the action of two stochastic operators on the *duality functional* $\mathcal{O}_\mu$.

Of course, the above arguments only verify the duality relation (6-5) for a specific class of distributions on row n. However, this class is sufficiently general ($n \geq m$ and parameters $x_1, \ldots, x_n$ are arbitrary), and it is quite plausible that the duality relation will hold for generic distributions on coloured compositions $\mu^{\geq 1}$ on the n-th row, and also for the n-th row being the whole lattice $\mathbb{Z}$, rather than $\mathbb{Z}_{\geq 1}$ with a path entering from the left.

It would be very interesting to see an independent verification of (6-5), possibly by a reduction to duality functional constructed in [Kuan 2018]. Given that spectral decomposition of stochastic kernels G^{stoch} are known (see [Borodin and Wheeler 2018, Section 9.5]) this would likely lead to an alternative proof of Theorem 6.1, apart from other possible applications.

For a colour-blind version of the above discussion see [Borodin and Petrov 2017, Section 8.5] and [Borodin and Petrov 2018a, Section 8.5].

6D. *The rainbow case.* Let us now focus on the rainbow sector, with the labelling composition λ being $(1, 1, \ldots, 1)$. The simplification of the observables $\mathcal{O}_\mu$ in this case was given in (4-13), and this leads us to:

Corollary 6.6. *In the notation of Section 6A, assume that* $\lambda = (1, 1, \ldots, 1)$. *Then*

$$\mathbb{E} \prod_{\substack{i,j \geq 1 \\ m_i^{(j)}=1}} \frac{q^{H^{\nu/\mu}_{\geq i}(j+1)} - q^{H^{\nu/\mu}_{> i}(j+1)}}{q-1}$$
$$= \frac{q^{\sum_{j\geq 1}\binom{|\boldsymbol{m}^{(j)}|}{2}}}{\prod_{j\geq 1}(s^2;q)_{|\boldsymbol{m}^{(j)}|}} \frac{(-1)^m}{(2\pi\sqrt{-1})^m} \oint \cdots \oint_{\text{incl.}\{x_j^{-1}\}} \prod_{1\leq i<j\leq m} \frac{y_j - y_i}{y_j - q y_i}$$
$$\times f^{\text{stoch}}_{\mu^{\geq 1}}(y_1^{-1}, \ldots, y_m^{-1}) \prod_{j=1}^{m} \left(\frac{y_j - s}{y_j} \frac{x_{c_j}}{1 - x_{c_j} y_j} \prod_{i>c_j}^{n} \frac{1 - q x_i y_j}{1 - x_i y_j} \, dy_j \right), \tag{6-6}$$

where (positively oriented) integration contours are chosen to encircle all points $\{x_j^{-1}\}_{j=1}^n$ *and no* *other singularities of the integrand, or as q-nested closed simple curves with y_i-contour containing $q^{-1}\cdot(y_j$-contour) for all $i<j$, and all of the contours encircling $\{x_j^{-1}\}_{j=1}^n$.*

The formula also holds in the column inhomogeneous setting under the assumption that $(s_x,\xi_x)=(s,1)$ for any $x\geq 1$ such that $|\boldsymbol{m}^{(x)}|>0$.

Proof. This can be obtained by either a direct substitution $\lambda=(1,\dots,1)$ into Theorem 6.1, or a substitution of Proposition 5.2 into the rainbow case of Theorem 4.7. □

Remark 6.7. Denote the parts of $\mu^{\geq 1}$ by $(\theta_1,\dots,\theta_m)$, where θ_1 carries the smallest colour c_1 of those represented in $\mu^{\geq 1}$, θ_2 carries the next smallest colour c_2, etc. Then in the antidominant case, when $\theta_1\leq\theta_2\leq\dots\leq\theta_m$, $f_\mu^{\text{stoch}}(y_1^{-1},\dots,y_m^{-1})$ in the integrand of (6-6) completely factorizes — and hence so does the whole integrand — as (in the column homogeneous case)

$$f_{\mu^{\geq 1}}^{\text{stoch}}(y_1^{-1},\dots,y_m^{-1})=\prod_{j\geq 1}(s^2;q)_{|\boldsymbol{m}^{(j)}|}\prod_{i=1}^{n}\frac{y_i}{y_i-s}\prod_{i=1}^{n}\left(\frac{1-sy_i}{y_i-s}\right)^{\theta_i}; \tag{6-7}$$

see (2-30), (3-2), and Remark 5.4. In this case there is also another path to (6-6). Namely, using the shift invariance property established in [Borodin et al. 2019], one can rewrite the left-hand side of (6-6) in terms of an average for a combination of q-moments of the height function for a colour-blind vertex on a quadrant taken at points along a down-right path in the quadrant. In their turn, such moments possess explicit integral representations, see [Borodin et al. 2016a] for details. See also Remark 7.10 below for a related observation.

Remark 6.8. The identity (6-6) admits two colour-position symmetries, one for each side.

For the left-hand side, [Borodin and Bufetov 2019, Theorem 7.3] shows that the joint distributions of the coloured height functions of the (rainbow) coloured stochastic vertex model along boundaries of a certain class of down-right domains are symmetric with respect to rotations of the domains by 180 degrees that also swap the roles of colour and position of the entering/exiting paths. In our case, the domain is the rectangle, and the application of this symmetry will swap colours and positions in the observables, as well as the roles of rows and columns. This will give integral formulas for averages of the new observables, as well as indicate that those observables are also likely to be duality functionals for the same reasons as those in Section 6C.

For the right-hand side, if we represent f_μ^{stoch} as a result of a sequence of applications of the difference operators T_i given by (2-31)–(2-32) to a factorized expression of the type (6-7), then we can use self-adjointness of the T_i's with respect to the integral scalar product with weight $\prod_{i<j}(y_j-y_i)/(y_j-qy_j)$ (see [Borodin and Wheeler 2018, Proposition 8.1.3]) to move the application of the T_i's to the fully factorized part of the integrand. Treating that part as an analogue of (2-30), or rather of (3-2), at the specific value of $s=q^{-1/2}$, we will see a new f-like function appearing in the integrand, that will utilize suitably permuted colours $c_1,\dots,c_m$ for its coordinates, and horizontal rapidities $x_1,\dots,x_n$ as its inhomogeneities. Thus, we will see a similar formula with positions and colours, as well as rows and columns, swapped.

It would be interesting to see if applying both symmetries, at least in the case $s = q^{-1/2}$, returns one to the original formula, but we will not pursue that here.

6E. ***Fusion.*** The goal of this section is to fuse rows of the stochastic vertex model from Section 6A and to see what Theorem 6.1 turns into in that situation.

First, one can fuse finitely many rows, which in terms of individual vertices corresponds to the outer sum in (2-10) (one can think of the inner sum in that relation as already performed, with our spin parameter s being $q^{-\mathsf{M}/2}$). This starts by replacing a single row of colour c with rapidity x by $\mathsf{L} \geq 1$ rows of the same colour c and rapidities forming a finite geometric progression $x, \dots, q^{\mathsf{L}-1}x$. Since our left boundary condition in these L rows consists of all incoming edges occupied by paths of the same colour, no summation along that boundary is necessary. Hence, the L rows of vertices with L^{stoch}-weights can be collapsed into a single row with the weight of a vertex in that row and column j being

$$W_{\mathsf{L},\mathsf{M}}(s_j^{-1}\xi_j x; q; *)|_{q^{-\mathsf{M}}=s_j^2},$$

where (s_j, ξ_j) are the column inhomogeneities as in Section 3A, and the appearance of s_j^{-1} in front of $\xi_j x_j$ is due to the argument x/s in the expression (2-11) of L^{stoch}-weights in terms of W-weights.

The right-hand side of (6-2) also behaves well with respect to such fusion. More exactly, it leads to a simple replacement of all factors of the form $(1 - qxy_k)/(1 - xy_k)$, for various k and $x = x_*$ being the rapidity of the fused row, by

$$\frac{1-qxy_k}{1-xy_k} \cdot \frac{1-q^2xy_k}{1-qxy_k} \cdots \frac{1-q^{\mathsf{L}}xy_k}{1-q^{\mathsf{L}-1}xy_k} = \frac{1-q^{\mathsf{L}}xy_k}{1-xy_k}. \tag{6-8}$$

Thus, the right-hand side can be immediately analytically continued in q^{L}. It is not clear, however, what that would mean on the side of the stochastic vertex model as the left boundary condition in the fused row consists of L paths.

To remedy this situation, we will perform a special limit transition with vertical inhomogeneity in column 1; this is parallel to what was done in [Borodin et al. 2019, Section 6]. More exactly, we will rely on the limiting relation (2-17). According to the left-hand side of that relation, we will take $\xi_1 = \zeta/s_1$ to turn the weight in the first column of the fused row into $W_{\mathsf{L},\mathsf{M}}(s_1^{-2}\zeta x; q; *)|_{q^{-\mathsf{M}}=s_1^2}$, and then take the limit $s_1 \to 0$. Here z is an additional parameter that remains finite in the limit $s_1 \to 0$; it regulates the strength of the left boundary. As we are about to make the rest of the model homogeneous, let us also set ζ to s, as this value will make the first column "blend in" with the other ones. Thus, our limit results, by virtue of (2-17), a random number of paths of colour c passing horizontally from column 1 to column 2 in the fused row, with the distribution of this random number given by

$$\mathrm{Prob}\{k\} = \frac{(sq^{\mathsf{L}}x; q)_\infty}{(sx; q)_\infty} \frac{(q^{-\mathsf{L}}; q)_k}{(q; q)_k} (sq^{\mathsf{L}}x)^k, \qquad k \geq 0.$$

Note that for $\mathsf{L} \in \mathbb{Z}_{\geq 1}$, this distribution is supported by $\{0, 1, \dots, \mathsf{L}\}$, as it should be, as an L-fused row cannot carry more than L paths. But this distribution is also suitable for analytic continuation in q^{L}, which we will use momentarily.

Let us now discuss vertices in the fused row and other columns. Their weights $W_{\mathsf{L},\mathsf{M}}(s_j^{-1}\xi_j x; q; *)$ are given by the right-hand side of (2-6), which is explicit but rather complicated. We will consider a simpler situation instead, governed by the q-Hahn specialization of Section 2B. Hence, we will specialize $(x, s_j, \xi_j) \mapsto (s, s, 1)$ to turn the weights into more tangible expressions as in the right-hand side of (2-16).

Finally, rather than performing fusion for a particular row of colour c as we have done above (replacing it by L columns first and then fusing them together), let us do it for every row of the stochastic vertex model of Section 6A.

Gathering all the pieces together, we obtain a stochastic vertex model in the quadrant $\mathbb{Z}_{\geq 2} \times \mathbb{Z}_{\geq 1}$, depending on three parameters q, $s(=q^{-\mathsf{M}/2})$, and $z(=q^{-\mathsf{L}/2})$ satisfying $|q|, |s|, |z|, |s/z| < 1$, defined as follows:

- Along the boundary of the quadrant, no paths enter the quadrant through its bottom boundary; and along the left boundary we have a random number of paths entering at every row, with the bottom λ_1 rows hosting entering paths of colour 1, next λ_2 rows hosting paths of colour 2, and so on; the sequence $\{\lambda_j\}_{j\geq 1}$ of nonnegative integers is given. As before, the partial sums of λ are denoted as $\sum_{i=1}^{k} \lambda_i = \ell_k,\ k \geq 1;\ \ell_0 := 0$. The distribution of the number of paths that enter in any row is given by
$$\operatorname{Prob}\{k\} = \frac{(s^2/z^2; q)_\infty}{(s^2; q)_\infty} \frac{(z^2; q)_k}{(q; q)_k} \left(\frac{s^2}{z^2}\right)^k, \qquad k \geq 0. \tag{6-9}$$

- The vertex weights for the model are given by
$$\text{weight}_{s,z}\left(\begin{array}{ccc} & \boldsymbol{C} & \\ \boldsymbol{B} & \overset{\uparrow}{\rightarrow} & \boldsymbol{D} \\ & \boldsymbol{A} & \end{array}\right) = \left(\frac{s^2}{z^2}\right)^{|\boldsymbol{D}|} \frac{(s^2/z^2; q)_{|\boldsymbol{A}|-|\boldsymbol{D}|}(z^2; q)_{|\boldsymbol{D}|}}{(s^2; q)_{|\boldsymbol{A}|}} q^{\sum_{i<j} D_i(A_j - D_j)} \prod_{i\geq 1} \binom{A_i}{A_i - D_i}_q. \tag{6-10}$$

This model is more general than the one in Section 6A in the sense that it can carry any number of paths along any edges, not just the vertical ones. However, it is less general in that there are no remaining row and column inhomogeneities.[11]

As before, we focus on the paths that cross upwards from row n to row $n+1$ for some $n \geq 1$.[12] Encoding colours and horizontal positions of these crossings by a coloured composition ν, we can define observables $\mathcal{O}_\mu$ on the set of possible ν's by the same formula (6-1), where $\mu = 2^{\boldsymbol{m}^{(2)}} 3^{\boldsymbol{m}^{(3)}} \cdots$ is an arbitrary coloured composition with no parts smaller than 2; this is what used to be $\mu^{\geq 1}$. Let us use the familiar notation $m = \ell(\mu)$ for the length of μ, $c_1 < \cdots < c_\alpha$ for the colours represented in μ, and $\mathfrak{m}_1, \ldots, \mathfrak{m}_\alpha \geq 1$ for the number of parts of μ of colours $c_1, \ldots, c_\alpha$, respectively. Finally, let colour(μ) be the composition that encodes the colouring of μ, and let $\mu - 1^m$ denote the composition obtained from μ by subtracting 1 from each part.

[11] In fact, we could have left s and z parameters column and row dependent, respectively, but chose not to do so for the sake of simplicity. On the other hand, we could not have left row and column rapidities x_* and ξ_* generic and still had the same factorized form of the weights.

[12] Since rows above n do not matter to us, we assume, as in Section 6A, that colours of left-entering arrows in rows n and $n+1$ are different.

Corollary 6.9. *With the above notation for the fused stochastic vertex model, we have*

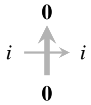
$$
\begin{aligned}
\mathbb{E}\,\mathcal{O}_\mu = {} & \frac{q^{\sum_{u\geq 2}\sum_{i>j} m_i^{(u)} m_j^{(u)}}}{\prod_{j\geq 2}(s^2;q)_{|\boldsymbol{m}^{(j)}|}} \frac{(-s)^m}{(2\pi\sqrt{-1})^m} \\
& \times \oint \cdots \oint \prod_{1\leq i<j\leq m} \frac{y_j-y_i}{y_j-qy_i} \\
& \times \prod_{k=1}^{\alpha} \Biggl(\sum_{j=0}^{\mathfrak{m}_k} \frac{(-1)^j q^{\binom{\mathfrak{m}_k-j}{2}}}{(q;q)_j (q;q)_{\mathfrak{m}_k-j}} \prod_{p>\mathfrak{m}[1,k-1]}^{j+\mathfrak{m}[1,k-1]} \left(\frac{1-z^{-2}sy_p}{1-sy_p}\right)^{n-\ell_{c_k-1}} \prod_{r>j+\mathfrak{m}[1,k-1]}^{\mathfrak{m}[1,k]} \left(\frac{1-z^{-2}sy_r}{1-sy_r}\right)^{n-\ell_{c_k}} \Biggr) \\
& \times f^{\mathrm{stoch}}_{\mu-1^m}(\mathrm{colour}(\mu); y_1^{-1},\ldots,y_m^{-1}) \prod_{i=1}^{m} \frac{dy_i}{y_i^2}, \qquad (6\text{-}11)
\end{aligned}
$$

where the integration contours are either positively oriented and q-nested around s^{-1} with y_i-contour containing $q^{-1}\cdot(y_j$-contour) for all $i<j$, or negatively oriented and q-nested around s, with y_j-contour containing $q\cdot(y_i -$ contour) for all $i<j$. The point 0 *can be either inside or outside the contours in either case.*

Proof. The starting point is Theorem 6.1 with μ from (6-11) being $\mu^{\geq 1}$ there.

The first step is to turn each row of the quadrant into L rows of the same colour, and fuse them as was described above. This does not affect the integral representation much, except for the change described around (6-8).

The next step is the limit transition in column 1 described above. To see how it affects the factor $f^{\mathrm{stoch}}_{\mu^{\geq 1}}(y_1^{-1},\cdots,y_m^{-1})$, recall the partition function definition (2-23) (and its coloured modification (2-41)) of the f's. Since all parts of $\mu^{\geq 1}$ are assumed to be at least 2, the first column contains only vertices of the form



that have L^{stoch}-weights

$$
\frac{(-s_1)(\xi_1 y_j^{-1}-s_1)}{1-s_1\xi_1 y_j^{-1}} = \frac{(-s_1)(\xi_1-s_1y_j)}{y_j-s_1\xi_1} = \frac{(-s_1)(s/s_1-s_1y_j)}{y_j-s} = \frac{s_1^2y_j-s}{y_j-s}, \qquad 1\leq j\leq m,
$$

where we used the value $\xi_1=s/s_1$ as above. The limit $s_1\to 0$ of this expression is $-s/(y_j-s)$, and the product over all $1\leq j\leq m$ gives $(-s)^m\prod_{j=1}^m(y_j-s)^{-1}$. Adding that to the integrand of (6-2) yields the integrand of (6-11), together with the replacement $\mu^{\geq 1}\mapsto\mu-1^m$ in the index of f^{stoch}, where the subtraction of 1^m is responsible for removing a column from the partition function representation of f^{stoch} that we just performed.

Let us now look at the contours. For the application of Theorem 6.1 we could choose them to q-nest around $\{x_i,qx_i,\ldots,q^{\mathsf{L}-1}x_i\}_{i=1}^n=\{s,qs,\ldots,q^{\mathsf{L}-1}s\}$ and either contain 0 or not. (We could not choose them to encircle $\{x_i,qx_i,\ldots,q^{\mathsf{L}-1}x_i\}_{i=1}^n$ and no other singularities as this choice of horizontal rapidities forces the inclusion of at least some singularities of $\prod_{i<j}(y_j - qy_i)^{-1}$ into the contours.) Since the integrand is manifestly nonsingular at $qs, \dots, q^{\mathsf{L}-1}s$, we can remove the condition of encircling those points. As the integrand is readily seen to not have poles at $y_* = \infty$, we can also move the contours through ∞ and have them q-nest around the only other singularity, which is at s (again, 0 can be either inside or outside).

Thus, we have now proved (6-11) for $\mathsf{L} = 1, 2, \dots$, and the final step consists in analytic continuation in $q^{\mathsf{L}} = z^{-2}$ from the set of points $\{q, q^2, q^3, \dots\}$ accumulating at 0. This, however, is straightforward: The only dependence on z of the right-hand side is through factors $(1 - z^{-2}y_*)$, and the left-hand side is readily seen to be given by uniformly convergent series with rational terms at least as long as $|q|, |s|, |sz^{-1}| < \text{const} < 1$. □

7. Limit to polymers

The goal of this section is to explore the consequences of Corollary 6.9 for a few models of directed polymers in (1+1) dimensions. The exposition of the limit transitions from fused coloured stochastic models to directed polymers follows [Borodin et al. 2019, Section 7].

7A. ***Continuum stochastic vertex model.*** Let us start by introducing a vertex model that will serve as a limiting object for the fused vertex model of Section 6E described by weights (6-9)–(6-10).

As in the fused case, the vertices of the continuum model will be parametrized by points of a quadrant, and to keep the notation parallel to that of Section 6E, we will use the quadrant $\mathbb{Z}_{\geq 2} \times \mathbb{Z}_{\geq 1}$.

Each vertex will have a certain *mass* of each colour ≥ 1 entering from the bottom and from the left, and exiting through the top and to the right. The mass is a real number in $[0, \infty)$, and for each vertex the total number of colours that have nonzero mass entering the vertex will always be finite. The mass of each colour passing through a vertex will always be preserved – the sum of incoming mass from the bottom and from the left must be equal to the sum of exiting mass to the right and through the top.

Let us denote the masses of colours $1, 2, \dots$ entering through the bottom of a vertex by $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots)$, entering from the left by $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots)$, exiting through the top by $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots)$, and exiting to the right by $\boldsymbol{\delta} = (\delta_1, \delta_2, \dots)$, respectively. The mass preservation means

$$|\boldsymbol{\alpha}| + |\boldsymbol{\beta}| = |\boldsymbol{\gamma}| + |\boldsymbol{\delta}|.$$

The notation is chosen to be in parallel with $(\boldsymbol{A}, \boldsymbol{B}; \boldsymbol{C}, \boldsymbol{D})$ notation for the vertex models, as in (2-1).

Recall that a random variable with values in $(0, 1)$ is said to be *Beta-distributed* with parameters $a, b > 0$ if it has a density, with respect to the Lebesgue measure, given by

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \qquad 0 < x < 1. \tag{7-1}$$

Given the coloured masses $\boldsymbol{\alpha}, \boldsymbol{\beta}$ entering a vertex of our continuum vertex model, the coloured masses $\boldsymbol{\gamma}, \boldsymbol{\delta}$ exiting the vertex are random and determined as follows. The procedure has two parameters $a, b > 0$, as in the Beta-distribution (7-1). If all coordinates $\boldsymbol{\alpha}$ are zero, i.e., $\boldsymbol{\alpha} = \boldsymbol{0}$, then we set $\boldsymbol{\gamma} = \boldsymbol{\beta}$ and $\boldsymbol{\delta} = \boldsymbol{0}$. If $\boldsymbol{\alpha} \neq \boldsymbol{0}$, let $n \geq 1$ be the maximal natural number such that $\alpha_n \neq 0$, and let ζ be an (a, b)-Beta distributed random variable. Then we set $\delta_{n+1}, \delta_{n+2}, \dots$ to 0 and define $\delta_n, \delta_{n-1}, \dots, \delta_1$ recursively via

$$\begin{aligned} \delta_n &= -\log(e^{-\alpha_n} + \zeta(1 - e^{-\alpha_n})), \\ \delta_{n-1} + \delta_n &= -\log(e^{-\alpha_{n-1}-\alpha_n} + \zeta(1 - e^{-\alpha_{n-1}-\alpha_n})), \\ &\;\vdots \\ \delta_1 + \cdots + \delta_{n-1} + \delta_n &= -\log(e^{-\alpha_1 - \cdots - \alpha_{n-1}-\alpha_n} + \zeta(1 - e^{-\alpha_1 - \cdots - \alpha_{n-1}-\alpha_n})). \end{aligned} \tag{7-2}$$

One can show that this implies $0 < \delta_j < \alpha_j$ for $1 \le j \le n$. Finally, we set $\boldsymbol{\gamma} = \boldsymbol{\beta} + (\boldsymbol{\alpha} - \boldsymbol{\delta})$, thus enforcing mass conservation.

In addition to defining what happens at the vertices of the quadrant, we need to specify boundary conditions. As before, we will assume that no mass enters the quadrant from the bottom, i.e., $\boldsymbol{\alpha} = \mathbf{0}$ for all vertices in the bottom row. On the other hand, along the left boundary we will assume that for the left-most vertex in row i, the left-entering coloured mass $\boldsymbol{\beta}$ has all but one coordinates equal to 0, with the exception of the i-th one, which is (a, b)-Beta distributed.

As usual, we think of the randomness as having no space dependency, which means that the Beta-distributed random variables at different vertices, as well as those used to define the left boundary condition, are independent.

The following statement was proved in [Borodin et al. 2019, Corollary 6.22].

Proposition 7.1. *Consider the fused coloured vertex model defined around (6-9)-(6-10) and set*

$$q = \exp(-\epsilon), \qquad s^2 = q^{\sigma}, \qquad z^2 = q^{\rho}, \tag{7-3}$$

for some $\sigma > \rho > 0$ and $\epsilon > 0$. Then as $\epsilon \to 0$, the fused coloured vertex models scaled by ϵ converges to the continuum vertex model defined above with parameters $(a, b) = (\sigma - \rho, \rho)$, in the sense that any finite collection of numbers of paths of arbitrary fixed colours entering/exiting any fixed set of vertices in fixed directions, when multiplied by ϵ, weakly converges to the collection of corresponding colour masses entering/exiting the corresponding vertices of the continuum model.

This immediately implies the convergence of the averages of the observables from Corollary 6.9 as well, but we will postpone the limiting statement until we reformulate the continuum vertex model as a directed random polymer in the next section.

7B. *Random Beta-polymer.*

The Beta-polymer was first introduced in [Barraquand and Corwin 2017]. In order to define it, let $\{\eta_{t,m}\}_{t,m\ge1}$ be a family of independent identically Beta-distributed random variables with parameters $a, b > 0$; see (7-1). The partition function $\mathfrak{Z}_{t,m}$ of the Beta-polymer, with $(t, m) \in \mathbb{Z}_{\ge0} \times \mathbb{Z}_{\ge1}$ and $t \ge m - 1$, is determined by the recurrence relation

$$\mathfrak{Z}_{t,m} = \eta_{t,m}\mathfrak{Z}_{t-1,m} + (1 - \eta_{t,m})\mathfrak{Z}_{t-1,m-1}$$

and boundary conditions

$$\mathfrak{Z}_{t,t+1} \equiv 1, \qquad \mathfrak{Z}_{t,1} = \eta_{1,1}\eta_{2,1}\cdots\eta_{t,1}.$$

Pictorially, $\mathfrak{Z}_{t,m}$ is a sum over all directed lattice paths with $(0, 1)$ and $(1, 1)$ steps that join $(0, 1)$ and (t, m), of products of edge weights that all have the form η_* or $1-\eta_*$; see [Barraquand and Corwin 2017; Borodin et al. 2019].

We also define *delayed* partition functions $\mathfrak{Z}^{(k)}_{t,m}$, $k \geq 1$, $t \geq m+k-1$, by the same recurrence relation

$$\mathfrak{Z}^{(k)}_{t,m} = \eta_{t,m}\mathfrak{Z}^{(k)}_{t-1,m} + (1-\eta_{t,m})\mathfrak{Z}^{(k)}_{t-1,m-1},$$

where we are using the same family of random variable $\{\eta_{t,m}\}_{t,m\geq 1}$ to evaluate the coefficients, and shifted boundary conditions

$$\mathfrak{Z}^{(k)}_{t,t-k+2} \equiv 1, \qquad \mathfrak{Z}^{(k)}_{t,1} = \eta_{k,1}\eta_{k+1,1}\cdots\eta_{t,1}.$$

Clearly, $\mathfrak{Z}^{(1)}_{t,m} \equiv \mathfrak{Z}^{(k)}_{t,m}$, and for any $k \geq 1$, $\mathfrak{Z}^{(k)}_{t,m}$ can be interpreted graphically in a similar way to $\mathfrak{Z}_{t,m}$, but with paths joining $(k-1, 1)$ and (t, m).

As was shown in [Borodin et al. 2019, Section 7.1], there is a way to identify the continuum vertex model of Section 7A and the family of Beta-polymer partition functions $\{\mathfrak{Z}^{(k)}_{t,m}\}$. In order to see the equivalence, let us introduce the coloured height functions $\{h^{(\geq k)}(x, y) \mid x \geq 2;\ y, k \geq 1\}$ that count the total mass of colours $\geq k$ that exit vertices $(x, 1), (x, 2), \ldots, (x, y)$, either upward or rightward, in the continuum model. Then one has the identification

$$-\log \mathfrak{Z}^{(k)}_{t,m} = h^{(\geq k)}(m+1, t), \qquad t, m, k \geq 1,\ t \geq m+k-1. \tag{7-4}$$

For $t, m, k \geq 1$ that do not satisfy the inequality $t \geq m+k-1$, the right-hand side of (7-4) is readily seen to vanish, and we set $\mathfrak{Z}^{(k)}_{t,m}$ to 1 for these values as well; then (7-4) holds for any $t, m, k \geq 1$.

Together with Proposition 7.1, this allows us to obtain a limiting version of Corollary 6.9 for the Beta-polymer. But in order to take the corresponding limit of the integral representation, we need to introduce limiting versions of the rational functions f_μ.

Lemma 7.2. *Take $q = \exp(-\epsilon)$, and $s^2 = q^\sigma$ (as in (7-3)), and fix a λ-coloured composition $\mu = 0^{\mathbf{m}^{(0)}}1^{\boldsymbol{m}^{(1)}}2^{\boldsymbol{m}^{(2)}}\cdots$ of length $m \geq 1$, where λ is a composition of weight $|\lambda| = m$. Then there exists a limit*

$$\mathfrak{f}_\mu(\lambda; u_1, \ldots, u_m) = \lim_{\epsilon\to 0} \epsilon^m \cdot \frac{f_\mu(\lambda; 1+\epsilon u_1, \ldots, 1+\epsilon u_m)}{\prod_{j\geq 0}(s^2; q)_{|\boldsymbol{m}^{(j)}|}}, \tag{7-5}$$

where the convergence is uniform for complex $u_1, \ldots, u_m$ varying in compact sets that do not include $\sigma/2$, and the rational function $\mathfrak{f}_\mu(\lambda; u_1, \ldots, u_m)$ can be characterized as follows.

(i) *For a rainbow μ, i.e., for $\lambda = (1, \ldots, 1)$, in the antidominant sector $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_m$ one has (omitting λ from the notation)*

$$\mathfrak{f}_\mu(u_1, \ldots, u_m) = \prod_{i=1}^{m} \frac{1}{\sigma/2 - u_i}\left(\frac{\sigma/2 + u_i}{\sigma/2 - u_i}\right)^{\mu_i}, \tag{7-6}$$

and in the case $\mu_i < \mu_{i+1}$ for some $1 \leq i \leq m-1$, one has

$$\mathfrak{T}_i \cdot \mathfrak{f}_\mu(u_1, \ldots, u_m) = \mathfrak{f}_{(\mu_1,\ldots,\mu_{i+1},\mu_i,\ldots,\mu_m)}(u_1, \ldots, u_m),$$
$$\mathfrak{T}_i \equiv 1 - \frac{u_i - u_{i+1} + 1}{u_i - u_{i+1}}(1 - \mathfrak{s}_i), \qquad 1 \leq i \leq m-1, \tag{7-7}$$

with elementary transpositions $\mathfrak{s}_i \cdot h(u_1, \ldots, u_m) := h(u_1, \ldots, u_{i+1}, u_i, \ldots, u_m)$; see (2-31)–(2-32).

(ii) *For a general labelling composition* λ, *let* $\theta : \{1, \dots, m\} \to \{1, \dots, n\}$ *be a monotone map* (*as in* Section 3C) *with* $\theta_*((1, \dots, 1)) = \lambda$. *Then*

$$\sum_{\text{rainbow } \nu \,:\, \theta_*(\nu)=\mu} \mathfrak{f}_\nu(u_1, \dots, u_m) = \mathfrak{f}_\mu(\lambda; u_1, \dots, u_m). \tag{7-8}$$

Proof. This limit is an immediate consequence of the recursive definition (2-30)–(2-32) of Section 2E and (3-18) of Proposition 3.8. □

We are now ready to take a $q \to 1$ limit in Corollary 6.9.

Proposition 7.3. *Fix* $\sigma > \rho > 0$, *and consider the partition functions* $\mathfrak{Z}^{(k)}_{t,m}$ *of the Beta-polymer as defined above with the parameters of* (7-1) *given by* $(a, b) = (\sigma - \rho, \rho)$. *Let* $\mu = 2^{\boldsymbol{m}^{(2)}} 3^{\boldsymbol{m}^{(3)}} \cdots$ *be a coloured composition with no parts smaller than* 2, $m = \ell(\mu)$ *be the length of* μ, $c_1 < \dots < c_\alpha$ *be the colours represented in* μ, *and* $\mathfrak{m}_1, \dots, \mathfrak{m}_\alpha \geq 1$ *be the number of parts of* μ *of colours* $c_1, \dots, c_\alpha$, *respectively. Also, let* colour(μ) *be the composition that encodes the colouring of* μ. *Then for any* $t \geq \max\{i : m^{(x)}_i > 0\}$ *we have*

$$\begin{aligned}
&\mathbb{E} \prod_{x\geq 1} \prod_{i\geq 1} \frac{(\mathfrak{Z}^{(i+1)}_{t,x} - \mathfrak{Z}^{(i)}_{t,x})^{m^{(x)}_i}}{m^{(x)}_i!} \\
&\quad = \frac{(-1)^{|\mu|}}{(2\pi\sqrt{-1})^m} \oint \cdots \oint \prod_{1\leq i<j\leq m} \frac{u_j - u_i}{u_j - u_i + 1} \\
&\qquad \times \prod_{k=1}^{\alpha} \Biggl(\sum_{j=0}^{\mathfrak{m}_k} \frac{(-1)^j}{j!(\mathfrak{m}_k - j)!} \prod_{p>\mathfrak{m}[1,k-1]}^{j+\mathfrak{m}[1,k-1]} \biggl(\frac{\sigma/2 - u_p - \rho}{\sigma/2 - u_p} \biggr)^{t-c_k+1} \prod_{r>j+\mathfrak{m}[1,k-1]}^{\mathfrak{m}[1,k]} \biggl(\frac{\sigma/2 - u_r - \rho}{\sigma/2 - u_r} \biggr)^{t-c_k} \Biggr) \\
&\qquad\qquad \times \mathfrak{f}_{\mu-1^m}(\text{colour}(\mu); -u_1, \dots, -u_m) \prod_{i=1}^{m} du_i,
\end{aligned} \tag{7-9}$$

with the integration contours either positively oriented and nested around $\sigma/2$ *with* u_i*-contour containing* $(u_j$*-contour*$)+1$ *for all* $i < j$, *or negatively oriented and nested around* $-\sigma/2$, *with* u_j*-contour containing* $(u_i$*-contour*$) - 1$ *for all* $i < j$.

Proof. We start with (6-11) and $(\lambda_1, \lambda_2, \dots) = (1, 1, \dots)$; equivalently, $\ell_j = j$. Let us make the substitution (7-3) and look at the asymptotics of both sides.

On the left-hand side we have averages of the observables $\mathcal{O}_\mu$ given by (6-1). The denominators are deterministic and asymptotically give $(q; q)^{-1}_{m^{(x)}_i} \sim \epsilon^{-m^{(x)}_i} m^{(x)}_i!$ as $\epsilon \to 0$. For the numerators, according to Proposition 7.1, we obtain, along row $t = n$, $\epsilon^{-1} H^{\nu}_{\geq i}(x + 1) \to h^{(\geq i)}(x + 1, t)$, and, changing ν to ν/μ with ϵ-independent μ, $\epsilon^{-1} H^{\nu/\mu}_{\geq k}(x + 1) \to h^{(\geq k)}(x + 1, t)$ weakly as $\epsilon \to 0$, where h denotes the coloured height functions of the Beta-polymer. Using (7-4) and the fact that all the observables are bounded, we see that $\mathbb{E}\mathcal{O}_\mu$ is asymptotically equivalent to ϵ^{-m} times the left-hand side of (7-9).

For the right-hand side of (6-11), we change the variables $y_i = 1 + \epsilon u_i$, use Lemma 7.2, and also $(q; q)^{-1}_j (q; q)^{-1}_{\mathfrak{m}_k - j} \sim \epsilon^{-\mathfrak{m}_k} j!^{-1} (\mathfrak{m}_k - j)!^{-1}$.

The powers of ϵ from the q-symbols on both sides cancel out, the powers of ϵ from the changes of variables and $f \to \mathfrak{f}$ limit also cancel out, and the prefactor $(-s)^m$ in the right-hand side of (6-11) together with $(-s)^{|\mu|-m}$ required to convert $f^{\text{stoch}}_{\mu-1^m}$ to $f_{\mu-1^m}$ give $(-1)^{|\mu|}$ in the limit. This concludes the proof. □

7C. ***Strict-weak polymer.*** The strict-weak or gamma polymer was first introduced in [Corwin et al. 2015] and [O'Connell and Ortmann 2015]. Its partition functions are determined by a very similar recurrence as those for the Beta-polymer. Namely, let us define $Z^{(k)}_{t,m}$ for $(t,m,k) \in \mathbb{Z}_{\geq 0} \times \mathbb{Z}_{\geq 1} \times \mathbb{Z}_{\geq 1}$, $t \geq m+k-1$, by

$$Z^{(k)}_{t,m} = \eta_{t,m} Z^{(k)}_{t-1,m} + Z^{(k)}_{t-1,m-1}$$

with boundary conditions

$$Z^{(k)}_{t,t-k+2} \equiv 1, \qquad Z^{(k)}_{t,1} = \eta_{k,1}\eta_{k+1,1}\cdots\eta_{t,1},$$

where $\{\eta_{t,m}\}_{t,m\geq 1}$ is a family of independent identically distributed random variables with a Gamma distribution that has density

$$\frac{1}{\Gamma(\kappa)} x^{\kappa-1}\exp(-x), \qquad x > 0,$$

with respect to the Lebesgue measure. Here $\kappa > 0$ is a parameter.

The strict-weak polymer is a limiting instance of the Beta-polymer of Section 7B, because a Beta-distributed random variable with density (7-1) and parameters $(a,b) = (\kappa, \epsilon^{-1})$, when multiplied by ϵ^{-1}, converges to a Gamma-distributed random variable with parameter κ as $\epsilon \to 0$, both in distribution and with all moments.

In order to argue the convergence of the partition functions and their moments, we will appeal to the following:

Lemma 7.4. *Let $\{X_n\}_{n\geq 1}$ and $\{Y_n\}_{n\geq 1}$ be two sequences of nonnegative random variables such that $\{(X_n, Y_n)\}_{n\geq 1}$ weakly converges to a two-dimensional random variable (X, Y) with finite moments and a continuous joint distribution function.*[13] *Furthermore, assume that coordinate moments converge*:

$$\lim_{n\to\infty} \mathbb{E}X_n^k = \mathbb{E}X^k, \qquad \lim_{n\to\infty} \mathbb{E}Y_n^k = \mathbb{E}Y^k, \qquad k \geq 1.$$

Then the joint moments also converge:

$$\lim_{n\to\infty} \mathbb{E}(X_n^k Y_n^l) = \mathbb{E}\big(X^k Y^l\big), \qquad k, l \geq 1.$$

Proof. Fix $C > 0$ and write

$$\mathbb{E}(X_n^k Y_n^l) = \mathbb{E}(X_n^k Y_n^l \cdot \mathbf{1}\{X_n < C, Y_n < C\}) + \mathbb{E}(X_n^k Y_n^l \cdot \mathbf{1}\{X_n \geq C \text{ or } Y_n \geq C\}),$$

where we use $\mathbf{1}\{A\}$ to denote the indicator function of an event A. In the first term, we have a bounded functional under the expectation, which converges to $\mathbb{E}(X^k Y^l \cdot \mathbf{1}\{X < C, Y < C\})$ by the distributional convergence of (X_n, Y_n). (If the distribution of (X, Y) were not continuous, we would have needed to choose C as its continuity point.)

[13] This requirement of continuity can be easily removed by an extra step in the proof.

Further, let us show that the second term converges to 0 as $C \to \infty$ uniformly in n. We have

$$\begin{aligned}\mathbb{E}(X_n^k Y_n^l \cdot \mathbf{1}\{X_n \geq C \text{ or } Y_n \geq C\}) \leq \mathbb{E}(X_n^k Y_n^l \cdot (X_n/C + Y_n/C)) = C^{-1}\mathbb{E}(X_n^k Y_n^l \cdot (X_n + Y_n))\\ \leq 2C^{-1}(\mathbb{E}X_n^{k+l+1} + \mathbb{E}Y_n^{k+l+1}),\end{aligned}$$

where we used the inequality $x^a y^b \leq x^{a+b} + y^{a+b}$ that holds for $x, y, a, b > 0$. Since the moments of X_n and Y_n are bounded (because they converge by the hypothesis), the final expression tends to 0 as $C \to \infty$ uniformly in n.

As $\lim_{C\to\infty} \mathbb{E}(X^k Y^l \cdot \mathbf{1}\{X < C, Y < C\}) = \mathbb{E}(X^k Y^l)$, the proof is complete.[14] □

Lemma 7.4 implies, in particular, that by choosing $(\sigma - \rho, \rho) = (\kappa, \epsilon^{-1})$, we can ensure the convergence

$$\lim_{\epsilon \to 0} \epsilon^{m+k-t-2} \cdot \mathfrak{z}_{t,m}^{(k)} = Z_{t,m}^{(k)}, \qquad t \geq m+k-2,$$

together with all the joint moments. This will allow us to take such a limit in Proposition 7.3 momentarily, after the following analogue of Lemma 7.2.

Lemma 7.5. *Fix a λ-coloured composition $\mu = 0^{\boldsymbol{m}^{(0)}} 1^{\boldsymbol{m}^{(1)}} 2^{\boldsymbol{m}^{(2)}} \cdots$ of length $m \geq 1$, where λ is a composition of weight $|\lambda| = m$. Then there exists a limit*

$$\mathfrak{p}_\mu(\lambda; v_1, \dots, v_m) = \lim_{\sigma \to \infty} (-1)^{|\mu|+m} \sigma^{-|\mu|} \cdot \mathfrak{f}_\mu(\lambda; \sigma/2 + v_1, \dots, \sigma/2 + v_m) \tag{7-10}$$

where the convergence is uniform for complex $v_1, \dots, v_m$ varying in compact sets that do not include 0, *and the function $\mathfrak{p}_\mu(\lambda; u_1, \dots, u_m)$ is a polynomial in $v_1^{-1}, \dots, v_m^{-1}$ that can be characterized as follows.*

(i) *For a rainbow μ, in the antidominant sector $\mu_1 \leq \mu_2 \leq \dots \leq \mu_m$, one has* (*omitting λ from the notation*)

$$\mathfrak{p}_\mu(v_1, \dots, v_m) = \prod_{i=1}^{m} v_i^{-\mu_i - 1}, \tag{7-11}$$

and in the case $\mu_i < \mu_{i+1}$ for some $1 \leq i \leq m-1$ one uses the exchange relations (7-7) *with* $(\mathfrak{p}, v_*)$ *instead of* $(\mathfrak{f}, u_*)$.

(ii) *For a general labelling composition λ, let $\theta : \{1, \dots, m\} \to \{1, \dots, n\}$ be a monotone map with $\theta_*((1, \dots, 1)) = \lambda$. Then*

$$\sum_{\text{rainbow } \nu \,:\, \theta_*(\nu) = \mu} \mathfrak{p}_\nu(u_1, \dots, u_m) = \mathfrak{p}_\mu(\lambda; u_1, \dots, u_m). \tag{7-12}$$

The proof of this lemma is straightforward.

Proposition 7.6. *Fix $\kappa > 0$, and consider the partition functions $Z_{t,m}^{(k)}$ of the strict-weak polymer as defined above with κ being the parameter of the Gamma distribution. Let μ and the associated notation*

[14] We are very grateful to Vadim Gorin for providing this argument.

be as in Proposition 7.3. Then for any $t \geq m + \max\{i : m_i^{(x)} > 0\} - 1$, *we have*

$$\mathbb{E} \prod_{x\geq 1} \prod_{i=1}^{n} \frac{(Z_{t,x}^{(i+1)})^{m_i^{(x)}}}{m_i^{(x)}!} = \frac{1}{(2\pi\sqrt{-1})^m} \oint \cdots \oint \prod_{1\leq i<j\leq m} \frac{v_j - v_i}{v_j - v_i - 1} \\ \times \prod_{k=1}^{\alpha} \prod_{r>\mathfrak{m}[1,k-1]}^{\mathfrak{m}[1,k]} \frac{(\kappa + v_r)^{t-c_k}}{\mathfrak{m}_k!} \cdot \mathfrak{p}_{\mu-1^m}(\mathrm{colour}(\mu); v_1, \dots, v_m) \prod_{i=1}^{m} dv_i, \tag{7-13}$$

where the integration contours are positively oriented and nested around 0, *with* v_j*-contour containing* (v_i*-contour*) $+ 1$ *for all* $i < j$.

Proof. We start with (7-9), set $\sigma - \rho$ to κ, change the integration variables via $u_i = -v_i - \sigma/2$, and take the limit. Since $\mathfrak{z}_{t,x}^{(i)} \sim \sigma^{x+i-t-2} Z_{t,x}^{(i)}$ and $\mathfrak{z}_{t,x}^{(i+1)} \sim \sigma^{x+i-t-1} Z_{t,x}^{(i)}$, the second term dominates, and employing Lemma 7.4, we obtain the convergence of the left-hand side of (7-9) to that of (7-13), with an additional power of σ that has exponent

$$\sum_{i,x\geq 1} m_i^{(x)}(x + i - t - 1) = |\mu| + \sum_{k=1}^{\alpha} c_k \mathfrak{m}_k - (t+1)m.$$

On the other hand, in the integrand we have factors of the form $\sigma/2 - u_* = \sigma + v_* \sim \sigma$ in the denominator, that make the terms with $j = 0$ dominate and produce the power of σ with the exponent

$$\sum_{k=1}^{\alpha} \mathfrak{m}_k (c_k - t) = \sum_{k=1}^{\alpha} c_k \mathfrak{m}_k - tm.$$

Finally, the limit relation (7-10) yields the power of σ with the exponent $|\mu| - m$, thus matching the powers of σ on both sides. All the signs cancel out, where we note that we used the second choice of the contours in Proposition 7.3 and changed the negative orientation to the positive one for (7-13). □

7D. ***O'Connell–Yor semidiscrete Brownian polymer.*** This model was first introduced in [O'Connell and Yor 2001]. It is defined using a family $\{B_n(t)\}_{n\geq 1, t\geq 0}$ of independent standard Brownian motions. For each $n \geq 1$ and $t \geq s \geq 0$ we define its point-to-point partition function (with one of the points situated on level 1) as

$$Z_{(1,s)\to(n,t)}^{OY} = \int_{s=\tau_0<\tau_2<\cdots<\tau_n=t} \exp\left(\sum_{i=1}^{n} (B_i(\tau_i) - B_i(\tau_{i-1}))\right) d\tau_1 \cdots d\tau_{n-1}. \tag{7-14}$$

The classical functional central limit theorem and the fact that a Gamma-distributed random variable with large parameter L divided by L is approximately equal to 1 plus a standard normal variable divided by $\sqrt{L}$, yield the convergence (see [Borodin et al. 2019, Section 7.3])

$$\lim_{L\to\infty} \frac{Z_{Lt,n}^{(Ls+1)}(\kappa = L)}{L^{L(t-s)}} = \exp\left(\tfrac{s-t}{2}\right) \cdot Z_{(1,s)\to(n,t)}^{OY}, \qquad t \geq s \geq 0, \quad n \geq 1, \tag{7-15}$$

where on the left we take the strict-weak polymer partition functions from the previous section with the parameter $\kappa = L$, and the convergence is in finite-dimensional distributions.

Proposition 7.7. *Let μ and the associate notation be as in Proposition 7.3, and fix $0 \le s_1 < \cdots < s_\alpha$. Then for any $t \ge s_\alpha$ we have*

$$\mathbb{E} \prod_{x \ge 1} \prod_{i \ge 1} \frac{\left(\exp\left(\frac{s_i - t}{2}\right) \cdot Z^{OY}_{(1,s_i) \to (x,t)}\right)^{m_i^{(x)}}}{m_i^{(x)}!} \\ = \frac{1}{(2\pi\sqrt{-1})^m} \oint \cdots \oint \prod_{1 \le i < j \le m} \frac{v_j - v_i}{v_j - v_i - 1} \\ \times \prod_{k=1}^{\alpha} \frac{\exp\left((t - s_k) \cdot \sum_{r > \mathfrak{m}[1,k-1]}^{\mathfrak{m}[1,k]} v_r\right)}{\mathfrak{m}_k!} \cdot \mathfrak{p}_{\mu - 1^m}(\mathrm{colour}(\mu); v_1, \dots, v_m) \prod_{i=1}^{m} dv_i, \tag{7-16}$$

where the integration contours are positively oriented and nested around 0, *with v_j-contour containing* $(v_i$*-contour*$) + 1$ *for all $i < j$.*

Proof. Let us take the limit $L \to \infty$ of (7-13) with $\kappa = L$ and the coloured composition μ replaced by $\mu^{(L)}$ that has exactly the same parts, but the colours of those parts are $[s_1 L], \dots, [s_\alpha L]$ instead of $c_1, \dots, c_\alpha$, respectively.

Let us look at the left-hand side first. The weak joint convergence of the random variables

$$\lim_{L \to \infty} \frac{Z^{([s_i L]+1)}_{Lt,x}(\kappa = L)}{L^{tL - [s_i L]}} = \exp\left(\frac{s_i - t}{2}\right) \cdot Z^{OY}_{(1,s_i) \to (n,t)}$$

for various i and x follows from the central limit theorem, as was mentioned above, and the convergence of moments of these random variables follows from the corresponding convergence of their integral representations; see [Corwin et al. 2015, Theorem 5.3] for moments of the left-hand side, and [Borodin and Corwin 2014, Proposition 5.2.8] for the moments of the right-hand side.[15] Lemma 7.4 then shows that the left-hand side of (7-13) divided by the power of L with exponent

$$\sum_{i,x \ge 1} m_i^{(x)} (tL - [s_i L]) = \sum_{k=1}^{\alpha} \mathfrak{m}_k (tL - [s_k L])$$

converges to that of (7-16).

On the other hand, for the right-hand side the convergence of the L-dependent factors in the integrand is elementary:

$$\lim_{L \to \infty} \frac{(L + v_r)^{tL - [s_k L]}}{L^{tL - [s_k L]}} = \exp((t - s_k) v_r),$$

uniformly for bounded v_r's, which leads to the right-hand side of (7-16). □

7E. *Continuum Brownian polymer.* One way to define partition functions of the continuum Brownian polymer in (1+1)-dimensions is through solving the stochastic heat equation with multiplicative

[15] That convergence is, in fact, a special case of the one we are about to observe for single-coloured μ.

two-dimensional white noise. More exactly, let $\widetilde{\mathcal{Z}}^{(y)}(t,x)$ be the unique solution of the following stochastic partial differential equation with the initial condition

$$\mathcal{Z}^{(y)}_t = \tfrac{1}{2}\mathcal{Z}^{(y)}_{xx} + \eta(t,x)\mathcal{Z}^{(y)}, \qquad t>0, \quad x\in\mathbb{R}; \qquad \mathcal{Z}^{(y)}(0,x)=\delta(x-y),$$

where $\eta=\eta(t,x)$ is the two-dimensional white noise. We refer to the survey [Quastel 2012] and references therein for an extensive literature on this equation and its close relation to continuum Brownian path integrals and the Kardar–Parisi–Zhang equation.

The solutions $\mathcal{Z}^{(y)}(t,x)$ arise naturally as limits of the partition functions of the semidiscrete Brownian polymer from the previous section:

$$\lim_{L\to\infty} \frac{\exp\bigl(\frac{y-t\sqrt{L}}{2}\bigr)\cdot Z^{OY}_{(1,y)\to(tL-x\sqrt{L},t\sqrt{L})}}{\exp(tL)\cdot L^{(x\sqrt{L}-tL)/2}} = \mathcal{Z}^{(y)}(t,x). \tag{7-17}$$

This was essentially verified on the level of convergence of integral representations for moments in [Borodin and Corwin 2014], and a complete proof for convergence of finite-dimensional distributions and moments with varying x was given in [Nica 2016] (in different scalings). It is very likely that the methods of [Nica 2016] are sufficient to achieve the same result for varying y as well; we will not address that here but rather focus on convergence of integral representations for joint moments instead. We need to start with an appropriate analogue of Lemmas 7.2 and 7.5.

Lemma 7.8. *Take a λ-coloured composition $\mu = 0^{\boldsymbol{m}^{(0)}}1^{\boldsymbol{m}^{(1)}}2^{\boldsymbol{m}^{(2)}}\cdots$ of length $m\ge 1$, where λ is a composition of weight $|\lambda|=m$, and assume that*

$$\mu_i = tL-\kappa_i\sqrt{L}+o(\sqrt{L}) \quad \text{as } L\to\infty, \quad 1\le i\le m,$$

for a fixed m-tuple of reals $\kappa=(\kappa_1,\dots,\kappa_m)$ and $t>0$. Then there exists a limit

$$\lim_{L\to\infty} \frac{e^{t\sqrt{L}(w_1+\cdots+w_m)}\cdot \mathfrak{p}_{\mu-1^m}(\sqrt{L}+w_1,\dots,\sqrt{L}+w_m)}{L^{-(\mu_1+\cdots+\mu_m)/2}} = e^{\frac{t}{2}(w_1^2+\cdots+w_m^2)}\cdot \mathfrak{e}_\kappa(\lambda;w_1,\dots,w_m), \tag{7-18}$$

uniformly for bounded w_i's, where the function $\mathfrak{e}_\kappa(w_1,\dots,w_m)$ can be characterized as follows.

(i) *For a rainbow μ, in the dominant sector $\kappa_1\ge\kappa_2\ge\ldots\ge\kappa_m$ one has (omitting λ from the notation)*

$$\mathfrak{e}_\mu(w_1,\dots,w_m)=\exp(\kappa_1w_1+\cdots+\kappa_mw_m), \tag{7-19}$$

and in the case $\kappa_i>\kappa_{i+1}$ for some $1\le i\le m-1$ one uses the exchange relations (7-7) with $(\mathfrak{e},w_*)$ instead of $(\mathfrak{f},u_*)$.

(ii) *For a general labelling composition λ, let $\theta:\{1,\dots,m\}\to\{1,\dots,n\}$ be a monotone map with $\theta_*((1,\dots,1))=\lambda$. Then*

$$\sum_{\text{rainbow }\kappa':\theta_*(\kappa')=\kappa} \mathfrak{e}_{\kappa'}(w_1,\dots,w_m)=\mathfrak{e}_\kappa(\lambda;w_1,\dots,w_m). \tag{7-20}$$

Proof. It suffices to check the convergence for rainbow antidominant μ's, as the other cases follow from that one by finite linear combinations (note that w_i's are shifted but not scaled in the left-hand side of (7-18)). Then we need to prove that

$$\lim_{L\to\infty} \frac{e^{t\sqrt{L}(w_1+\cdots+w_m)}(\sqrt{L}+w_1)^{-\mu_1}\cdots(\sqrt{L}+w_m)^{-\mu_m}}{L^{-(\mu_1+\cdots+\mu_m)/2}} = e^{\frac{t}{2}(w_1^2+\cdots+w_m^2)}\cdot e^{\kappa_1 w_1+\cdots+\kappa_m w_m}.$$

As the relation splits into a product over w_i's, it suffices to consider the case of a single variable. We have

$$\begin{aligned}-\mu_1\log(\sqrt{L}+w_1) &= -\mu_1\biggl(\log\sqrt{L}+\biggl(\frac{w_1}{\sqrt{L}}-\frac{w_1^2}{2L}+o(L^{-1})\biggr)\biggr)\\ &= -\frac{\mu_1\log(L)}{2}-(tL-\kappa_1\sqrt{L}+o(\sqrt{L}))\biggl(\frac{w_1}{\sqrt{L}}-\frac{w_1^2}{2L}+o(L^{-1})\biggr)\\ &= \log L^{-\mu_1/2}-t\sqrt{L}w_1+\kappa_1 w_1+\frac{tw_1^2}{2}+o(1),\end{aligned}$$

as required. □

We can now make a limiting statement for the moments of $\mathcal{Z}^{(y)}(t,x)$.

Proposition 7.9. *Let $\kappa=(\kappa_1,\dots,\kappa_m)$ be a coloured composition of length m with real coordinates, and let the colours $s_1<\cdots<s_\alpha$ of the parts of κ also take real values; denote*

$$m_i^{(x)} = \#\{j:\kappa_j = x \text{ and has colour } s_i\},\qquad \mathfrak{m}_i=\sum_x m_i^{(x)},\qquad 1\le i\le\alpha,\ x\in\mathbb{R}.$$

Then

$$\begin{aligned}&\mathbb{E}\prod_{(i,x)\,:\,m_i^{(x)}>0}\frac{(\mathcal{Z}^{(s_i)}(t,x))^{m_i^{(x)}}}{m_i^{(x)}!}\\ &\quad=\frac{1}{(2\pi\sqrt{-1})^m}\int\cdots\int\prod_{1\le i<j\le m}\frac{w_j-w_i}{w_j-w_i-1}\\ &\qquad\times\prod_{k=1}^{\alpha}\frac{\exp\bigl(-s_k\cdot\sum_{r>\mathfrak{m}[1,k-1]}^{\mathfrak{m}[1,k]}w_r\bigr)}{\mathfrak{m}_k!}\cdot\mathfrak{e}_\kappa(\mathrm{colour}(\kappa);w_1,\dots,w_m)\prod_{i=1}^m e^{tw_i^2/2}dw_i,\end{aligned}\tag{7-21}$$

where the integration is over upwardly oriented lines $w_i=a_i+\sqrt{-1}\cdot\mathbb{R}$ with $\Re a_j>\Re a_i+1$ for $j>i$.

Sketch of the proof. We obtain (7-21) as a limit of (7-16). The convergence of the left-hand sides was discussed below (7-17). The convergence of the right-hand sides is a standard steepest descent argument with the main contribution coming from a finite neighbourhood of the critical point $v=\sqrt{L}$; see the proof of [Borodin and Corwin 2014, Proposition 5.4.2] for a similar situation. The change of variables $v_i=\sqrt{L}+w_i$, $1\le i\le m$, together with Lemma 7.8, leads to the convergence of the integrands. □

Remark 7.10. Assume that the string $\kappa=(\kappa_1,\dots,\kappa_m)$ can be split into three sequential (possibly empty) substrings $\kappa=(\kappa',\kappa'',\kappa''')$, with all the coordinates of κ'' being (weakly) smaller than those of κ' and (weakly) larger than those of κ''', and with all the colours s_i of the coordinates of κ'' being (strictly) larger than those of κ' and (strictly) smaller than those of κ'''. A special case of this situation is the dominant sector $\kappa_1 \geq \ldots \geq \kappa_m$ with rainbow compositions served by (7-19).

It is not hard to show from the definition of the $\mathfrak{e}_\kappa$-functions in Lemma 7.8, that under a simultaneous shift of all the coordinates of κ'' by Δ that does not change the ordering conditions above, $\mathfrak{e}_\kappa(w_1, \ldots, w_m)$ is multiplied by $\exp(\Delta(w_a + \ldots, w_b))$, where $\kappa'' = (\kappa_a, \ldots, \kappa_b)$. Hence, if one simultaneously performs the shift of all the colours of the parts of κ'' by the same amount Δ, then the right-hand side of (7-21) is not going to change. Since this means that the moments in the left-hand side do not change either, it is natural to conjecture that the joint distribution of the participating $\mathcal{Z}$'s also does not change (the moments do not determine this distribution uniquely, though). When κ'' consists of one part, this conjecture was verified in [Borodin et al. 2019], along with its versions for higher models, up to coloured stochastic vertex models in general "down-right" domains. For general κ'' this conjecture was very recently proved in [Dauvergne 2020] and [Galashin 2020] by two different methods.

Acknowledgments

Borodin was partially supported by NSF grants DMS-1664619 and DMS-1853981. Wheeler was partially supported by ARC grant DP19010289.

References

[Aggarwal et al. 2019] A. Aggarwal, A. Borodin, and A. Bufetov, "Stochasticization of solutions to the Yang–Baxter equation", *Ann. Henri Poincaré* **20**:8 (2019), 2495–2554. MR

[Amir et al. 2011] G. Amir, I. Corwin, and J. Quastel, "Probability distribution of the free energy of the continuum directed random polymer in $1+1$ dimensions", *Comm. Pure Appl. Math.* **64**:4 (2011), 466–537. MR

[Barraquand and Corwin 2017] G. Barraquand and I. Corwin, "Random-walk in beta-distributed random environment", *Probab. Theory Related Fields* **167**:3-4 (2017), 1057–1116. MR

[Borodin 2017] A. Borodin, "On a family of symmetric rational functions", *Adv. Math.* **306** (2017), 973–1018. MR

[Borodin 2018] A. Borodin, "Stochastic higher spin six vertex model and Macdonald measures", *J. Math. Phys.* **59**:2 (2018), [art. id. 023301]. MR

[Borodin and Bufetov 2019] A. Borodin and A. Bufetov, "Color-position symmetry in interacting particle systems", 2019. arXiv

[Borodin and Corwin 2014] A. Borodin and I. Corwin, "Macdonald processes", *Probab. Theory Related Fields* **158**:1-2 (2014), 225–400. MR

[Borodin and Olshanski 2017] A. Borodin and G. Olshanski, "The ASEP and determinantal point processes", *Comm. Math. Phys.* **353**:2 (2017), 853–903. MR

[Borodin and Petrov 2017] A. Borodin and L. Petrov, "Integrable probability: stochastic vertex models and symmetric functions", pp. 26–131 in *Stochastic processes and random matrices*, edited by G. Schehr et al., Oxford Univ. Press, 2017. MR

[Borodin and Petrov 2018a] A. Borodin and L. Petrov, "Higher spin six vertex model and symmetric rational functions", *Selecta Math. (N.S.)* **24**:2 (2018), 751–874. MR

[Borodin and Petrov 2018b] A. Borodin and L. Petrov, "Inhomogeneous exponential jump model", *Probab. Theory Related Fields* **172**:1-2 (2018), 323–385. MR

[Borodin and Wheeler 2018] A. Borodin and M. Wheeler, "Coloured stochastic vertex models and their spectral theory", preprint, 2018. arXiv

[Borodin et al. 2014] A. Borodin, I. Corwin, and T. Sasamoto, "From duality to determinants for q-TASEP and ASEP", *Ann. Probab.* **42**:6 (2014), 2314–2382. MR

[Borodin et al. 2015a] A. Borodin, I. Corwin, L. Petrov, and T. Sasamoto, "Spectral theory for interacting particle systems solvable by coordinate Bethe ansatz", *Comm. Math. Phys.* **339**:3 (2015), 1167–1245. MR

[Borodin et al. 2015b] A. Borodin, I. Corwin, L. Petrov, and T. Sasamoto, "Spectral theory for the q-Boson particle system", *Compos. Math.* **151**:1 (2015), 1–67. MR

[Borodin et al. 2016a] A. Borodin, A. Bufetov, and M. Wheeler, "Between the stochastic six vertex model and Hall–Littlewood processes", preprint, 2016. To appear in *J. Comb. Theory, Series A*. arXiv

[Borodin et al. 2016b] A. Borodin, I. Corwin, and V. Gorin, "Stochastic six-vertex model", *Duke Math. J.* **165**:3 (2016), 563–624. MR

[Borodin et al. 2019] A. Borodin, V. Gorin, and M. Wheeler, "Shift-invariance for vertex models and polymers", preprint, 2019. arXiv

[Bosnjak and Mangazeev 2016] G. Bosnjak and V. V. Mangazeev, "Construction of R-matrices for symmetric tensor representations related to $U_q(\widehat{sl_n})$", *J. Phys. A* **49**:49 (2016), 495204, 19. MR

[Calabrese et al. 2010] P. Calabrese, P. L. Doussal, and A. Rosso, "Free-energy distribution of the directed polymer at high temperature", *EPL* (*Europhysics Letters*) **90**:2 (2010), 20002.

[Corwin and Petrov 2016] I. Corwin and L. Petrov, "Stochastic higher spin vertex models on the line", *Comm. Math. Phys.* **343**:2 (2016), 651–700. MR

[Corwin et al. 2015] I. Corwin, T. Seppäläinen, and H. Shen, "The strict-weak lattice polymer", *J. Stat. Phys.* **160**:4 (2015), 1027–1053. MR

[Dauvergne 2020] D. Dauvergne, "Hidden invariance of last passage percolation and directed polymers", preprint, 2020. arXiv

[Dotsenko 2010] V. Dotsenko, "Bethe ansatz derivation of the Tracy–Widom distribution for one-dimensional directed polymers", *EPL* (*Europhysics Letters*) **90**:2 (2010), [20003].

[Foda and Wheeler 2013] O. Foda and M. Wheeler, "Colour-independent partition functions in coloured vertex models", *Nuclear Phys. B* **871**:2 (2013), 330–361. MR

[Galashin 2020] P. Galashin, "Symmetries of stochastic colored vertex models", 2020. arXiv

[Garbali et al. 2017] A. Garbali, J. de Gier, and M. Wheeler, "A new generalisation of Macdonald polynomials", *Comm. Math. Phys.* **352**:2 (2017), 773–804. MR

[Gwa and Spohn 1992] L.-H. Gwa and H. Spohn, "Six-vertex model, roughened surfaces, and an asymmetric spin Hamiltonian", *Phys. Rev. Lett.* **68**:6 (1992), 725–728. MR

[Kardar 1987] M. Kardar, "Replica Bethe ansatz studies of two-dimensional interfaces with quenched random impurities", *Nuclear Phys. B* **290**:4 (1987), 582–602. MR

[Kardar et al. 1986] M. Kardar, G. Parisi, and Y.-C. Zhang, "Dynamic Scaling of Growing Interfaces", *Phys. Rev. Lett.* **56** (1986), 889–892.

[Kuan 2018] J. Kuan, "An algebraic construction of duality functions for the stochastic $\mathcal{U}_q(A_n^{(1)})$ vertex model and its degenerations", *Comm. Math. Phys.* **359**:1 (2018), 121–187. MR

[Kuniba et al. 2016] A. Kuniba, V. V. Mangazeev, S. Maruyama, and M. Okado, "Stochastic R matrix for $U_q(A_n^{(1)})$", *Nuclear Phys. B* **913** (2016), 248–277. MR

[Mimachi and Noumi 1998] K. Mimachi and M. Noumi, "A reproducing kernel for nonsymmetric Macdonald polynomials", *Duke Math. J.* **91**:3 (1998), 621–634. MR

[Nica 2016] M. Nica, "Intermediate disorder limits for multi-layer semi-discrete directed polymers", preprtint, 2016. arXiv

[O'Connell and Ortmann 2015] N. O'Connell and J. Ortmann, "Tracy–Widom asymptotics for a random polymer model with gamma-distributed weights", *Electron. J. Probab.* **20** (2015), [art. id.] 25. MR

[O'Connell and Yor 2001] N. O'Connell and M. Yor, "Brownian analogues of Burke's theorem", *Stochastic Process. Appl.* **96**:2 (2001), 285–304. MR

[Quastel 2012] J. Quastel, "Introduction to KPZ", pp. 125–194 in *Current developments in mathematics, 2011*, edited by D. Jerison et al., International Press, Somerville, MA, 2012. MR

[Sahi 1996] S. Sahi, "A new scalar product for nonsymmetric Jack polynomials", *Internat. Math. Res. Notices* 20 (1996), 997–1004. MR

[Sasamoto and Spohn 2010] T. Sasamoto and H. Spohn, "One-Dimensional Kardar–Parisi–Zhang Equation: An Exact Solution and its Universality", *Phys. Rev. Lett.* **104** (2010), 230602.

[Tracy and Widom 2008a] C. A. Tracy and H. Widom, "A Fredholm determinant representation in ASEP", *J. Stat. Phys.* **132**:2 (2008), 291–300. MR

[Tracy and Widom 2008b] C. A. Tracy and H. Widom, "Integral formulas for the asymmetric simple exclusion process", *Comm. Math. Phys.* **279**:3 (2008), 815–844. MR

[Tracy and Widom 2009] C. A. Tracy and H. Widom, "Asymptotics in ASEP with step initial condition", *Comm. Math. Phys.* **290**:1 (2009), 129–154. MR

Received 22 Mar 2020. Revised 6 Sep 2020. Accepted 21 Sep 2020.

ALEXEI BORODIN: borodin@math.mit.edu
Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, United States

and

Institute for Information Transmission Problems, Moscow, Russia

MICHAEL WHEELER: wheelerm@unimelb.edu.au
School of Mathematics and Statistics, University of Melbourne, Parkville VIC, Australia

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the submission page.

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 are usually in English or French, 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 refer to bibliography keys. 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 a Mathematics Subject Classification for the article, and, for each author, affiliation (if appropriate) and email address.

Format. Authors are encouraged to use LaTeX and the standard amsart class, but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should normally 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 as many details as you can about how your graphics were generated.

Bundle your figure files into a single archive (using zip, tar, rar or other format of your choice) and upload on the link you been provided at acceptance time. Each figure should be captioned and numbered so that it can float. Small figures occupying no more than three lines of vertical space can be kept in the text (“the curve looks like this:”). It is acceptable to submit a manuscript with all figures at the end, if their placement is specified in the text by means of comments such as “Place Figure 1 here”. The same considerations apply to tables.

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.

PROBABILITY and MATHEMATICAL PHYSICS

1:1 2020