Introduction to the Gopakumar–Vafa Large N Duality

DAVE AUCKLY
SERGIY KOSHKIN

Gopakumar–Vafa Large N Duality is a correspondence between Chern–Simons invariants of a link in a 3–manifold and relative Gromov–Witten invariants of a 6–dimensional symplectic manifold relative to a Lagrangian submanifold. We address the correspondence between the Chern–Simons free energy of S^3 with no link and the Gromov–Witten invariant of the resolved conifold in great detail. This case avoids mathematical difficulties in formulating a definition of relative Gromov–Witten invariants, but includes all of the important ideas.

There is a vast amount of background material related to this duality. We make a point of collecting all of the background material required to check this duality in the case of the 3–sphere, and we have tried to present the material in a way complementary to the existing literature. This paper contains a large section on Gromov–Witten theory and a large section on quantum invariants of 3–manifolds. It also includes some physical motivation, but for the most part it avoids physical terminology.

81T45; 81T30, 57M27, 17B37, 14N35

CONTENTS

Published: 21 September 2007 DOI: 10.2140/gtm.2006.8.195

Introduction

Large N duality is a conjectural correspondence between two very different types of mathematical objects: the large N limit of a gauge theory with structure group $U(N)$ and a string theory. Since gauge theories and string theories are both meant to describe the same universe it is natural to expect a correspondence between their predictions. There are several examples in physics and mathematics literature of this apparent duality. The original description of this correspondence goes back to the 1974 theoretical physics paper [148] by 't Hooft. At the time the main computational tool in both theories was perturbative expansion and 't Hooft noticed intriguing similarities that occur in those expansions. One of the first mathematical papers related to a Large N Duality is a 1992 paper where Kontsevich introduced a matrix model to resolve the conjectures of Witten about 2–dimensional gravity [87].

In oversimplified terms, gauge theory studies moduli spaces of connections on principal bundles while string theory studies spaces of maps from a particular class of domains into different targets. Both theories lead to invariants (of the base manifold in the gauge case and the target manifold in the string case). These invariants can be conveniently assembled into generating functions (called 'partition functions' by physicists). 't Hooft was considering expansions for gauge theories with SU(N) structure groups and noticed that as $N \to \infty$ they turn into partition functions one expects from a string theory.

In this generality 't Hooft's principle remains beyond the reach of mathematical theory for the foreseeable future. Ten years after Kontsevich's paper people began to understand the Large N Duality relating the Chern–Simons SU(N) gauge theory of 3–manifolds to the Gromov–Witten string theory of complex 3–folds.

The aspect of Large N Duality that we address in this survey is a duality between Chern–Simons theory and Gromov–Witten theory, the Gopakumar–Vafa Large N Duality duality. This duality states that the Chern–Simons (Reshetikhin–Turaev) invariants of a link in a 3–manifold are related to the relative Gromov–Witten invariants of a 6–dimensional symplectic manifold relative to a 3–dimensional Lagrangian submanifold. We address the correspondence between the Chern–Simons free energy of S^3 with no link and the Gromov–Witten invariant of the resolved conifold in great detail.

The key trait of both theories is that they are 'topological' in the sense that they do not depend on a background metric on the manifold in question. This greatly simplifies the setting and makes it possible to explain the Large N Duality in mathematical terms, something that remains impossible for other examples.

Mathematically the most that can be done is to compute the invariant on the Gromov–Witten side and compute the invariant on the Chern–Simons side and compare the two

answers. This is what we do. The first part of this paper (Part I) is an exposition of Gromov–Witten theory up to the point of the computation of the Gromov–Witten free energy of the resolved conifold (the full multiple cover formula). The definition of Gromov–Witten invariants is given in Section 3 with intuitive descriptions of some of the more technical elements. Formal definitions are given in the following subsections as they are motivated by and required for ever more complicated sample computations. We have included all of the relevant definitions. For example, the definition of a stack is given in Appendix A. Most expositions on Gromov–Witten theory avoid this definition.

One of the unique things that we do in this paper is a computation of the genus zero, degree two Gromov–Witten invariant of the resolved conifold directly from the definition; see Section 10. This case is addressed via localization in the book [43] by Cox and Katz.

The second part of this paper (Part II) is an exposition of Chern–Simons theory up to the point of the computation of the Chern–Simons free energy of the 3–sphere. The Chern–Simons invariants were motivated by a path integral expression. We outline the progression from this heuristic definition to formal definitions of perturbative invariants in Sections 14 and 15 and the first two parts of Section 16. The motivation for introducing the free energy is explained in Section 14.

The skein theory approach is described in Appendix C; this is the easiest way to describe the Chern–Simons invariants. It is however very difficult to compute from the resulting expressions, so we rely on the quantum group approach instead.

The main subsection in the second part is Section 16. It is here that we motivate and give a formal definition of the quantum group invariants. The second part of the paper ends with the computation of the Chern–Simons partition function.

Part III begins with Section 19 where we use special function techniques to derive the formal relation between the Gromov–Witten free energy and the Chern–Simons free energy. The remainder of this part is overview and history.

To get a fast introduction to the Gopakumar–Vafa Large N Duality one may just read the overview or read the first definition of the Gromov–Witten invariants from Section 3, the skein theory definition of the Chern–Simons invariants from Appendix C and the comparison of the two in Section 19. Some physical intuition may be obtained from the description of the perturbative expansion in Section 15.

Chern–Simons theory is defined for real 3–manifolds while the relevant Gromov–Witten theory is defined for Calabi–Yau complex 3–manifolds (in general one can define Gromov–Witten invariants for arbitrary symplectic manifolds, see Part I). Thus

the mathematically oriented reader can see Large N Duality as an interesting correspondence between 3–dimensional real and complex geometries and topologies. Physically the importance of the Calabi–Yau condition is that in string theory Calabi–Yau 3–folds (ie 6–real dimensional manifolds) provide complementary 'compactified' dimensions to the 4 observed ones of the classical space-time.

The existing literature on Large N Duality is vast and is growing exponentially so it would be impossible to survey it here. However, it appears to fall mostly into two categories: one written by physicists with extensive use of physical terminology (see the survey by Mariño [103] and references therein), another by or for algebraic geometers (see Cox and Katz [43] and the second half of Hori et al [76]). This reflects the fact that the physical insight and the complex side of the duality are at present the most developed parts.

One of the difficulties that impedes further progress in this subject is the amount of background material required to comprehend all the mathematical aspects of the conjectural duality (as one can judge, for example, from the size of the background chapters in Hori et al [76] and Turaev [152]).

We tried to provide a relatively self-contained introduction to the existing ideas and methods involved in Large N Duality that was complementary to the existing literature. We fill in details where we had trouble finding them and leave well-documented computations as exercises.

In order to keep the size of the the paper manageable we stuck to topics that could be formalized mathematically. In particular we cover the duality between the 3–sphere and the resolved conifold without including any knots. We also provide a number of computational and illustrative examples to make the matters clearer to a non-specialist. It is hoped that this paper will be accessible to advanced graduate students and will help to bring new blood into the field. Exercises are spread all over the text along with references to their solutions (or ideas for such). Although formally not necessary to get through the paper they are important for those who plan to acquire a working understanding of the subject matter.

The idea of writing this paper dates back to the workshop Interaction of Finite-type and Gromov–Witten Invariants at the Banff International Research Station in November 2003 co-organized by the first author and Jim Bryan. M Mariño gave a mini-series of lectures at this workshop on physics and mathematics of the Gopakumar–Vafa conjecture (as the Large N Duality was dubbed at the time). The plan was to write up lecture notes accessible to mathematics graduate students. However, M Mariño wrote an introduction that followed his lectures fairly closely [103] and it became clear that more details could not be given without writing a fair amount of background material.

The final version of the paper emerged out of friendly discussions between the two authors as we learned this material.

We would like to thank the directors and staff of the Banff International Research Station for providing an amazing place to share mathematics. We would like to thank Marcos Mariño for his illuminating lectures and Arthur Greenspoon for his careful editing of an earlier draft of this paper. The first author would like to thank Jim Bryan for first telling him about this duality. The second author would like to thank C-C M Liu and D Karp for fruitful discussions and suggestions related to the content of this work. Dave Auckly and Sergiy Koshkin were partially supported by NSF grant DMS-0204651.

1 Mathematical history of Large N Duality

It is instructive to trace the development of elements involved in the modern picture of Large N Duality. Hopefully this will also serve as a non-technical step-by-step introduction into the field. The reader should keep in mind that this is a mathematician's take on this task and a sporadic one at that. For instance, we almost completely ignore the physical undercurrent of the process except for a few landmark papers. A different perspective can be found in the introductory parts of Grassi–Rossi [67].

The history can be divided into four periods separated by physical breakthroughs into the mathematical realm: prehistory (1974–1989), formation of concepts and tools (1989–1998), the Gopakumar–Vafa conjecture (1998–2003), life after the vertex (2003–present). The dates in the text refer to arxiv submissions while references are given wherever possible to journal publications.

This historical overview describes the state of the subject in 2003 before the seminal paper 'The topological vertex' [3] re-shifted the perspective. Accordingly in this essay we only address the developments in 1974–2003. New results and directions that appeared after the November 2003 workshop are surveyed (or rather sketched) in the last section of this paper.

1.1 Prehistory (1974–1989)

When 't Hooft first formulated the Large N Duality principle neither Chern–Simons theory nor Gromov–Witten theory existed. However, this period saw the appearance of many ingredients that later fit into the picture.

In 1983 H Clemens in a paper called 'Double Solids' (ie complex threefolds) [41] studied extremal transitions between threefolds that include deformations of complex structure into a singular one with subsequent resolution of the singularity. The conifold transition

that connects cotangent bundles of 3–manifolds to their large N dual threefolds is an example of such a transition. The study of threefold singularities led to a conjecture known as Reid's fantasy (1987) that the moduli space of Calabi–Yau threefolds forms a single family related through such transitions (see Grassi and Rossi [67] and Reid [127]).

At about the same time, physicists realized that string theories remain well-defined on varieties with certain singularities (see Dixon, Harvey, Vafa and Witten [49]) and can therefore change smoothly as the underlying manifolds undergo a singular transition. This idea that the same theory can be described on topologically different bases is at the heart of Large N Duality.

On the other end, in 1985 V Jones introduced his polynomial invariant of knots [80]. However, his motivation came from operator algebras and no connection to Chern–Simons theory was known at the time. Already in 1985 several groups of authors generalized the Jones polynomial. Six authors published a joint paper [60] giving the new HOMFLY polynomial invariant (named with the initials of their last names). Two other authors [125] operated behind the iron curtain and their work remained unrecognized until somewhat later. To give full credit HOMFLY is now sometimes expanded to HOMFLYPT or THOMFLYP. It was later discovered that the Jones polynomial corresponds to the SU(2) and THOMFLYP to the SU(N) Chern–Simons theories respectively.

1.2 Formation of concepts and tools (1989–1998)

This period starts with the appearance of the first [159] of three seminal papers by E Witten. These papers revolutionized both the Chern–Simons and the Gromov–Witten theories and then tied them together. In 'Quantum field theory and the Jones polynomial' Witten introduced the (quantum) Chern–Simons theory as a gauge theory on 3–manifolds with Lagrangian density given by the Chern–Simons form (see Witten [159] and Baez–Muniain [20]). He then 'solved' it, that is, found explicit mathematical expressions for expectations of observables by reducing the computation to conformal field theory on Riemann surfaces (see Di Francesco–Mathieu–Sénéchal [45] or Appendix E). The observables turned out to be the so-called 'Wilson loops', that is, holonomies of connections over knots and links in the manifold. Witten's heuristic computation showed that expectation values of Wilson loops are the Jones and THOMFLYP polynomials with minor renormalizations. One can find a more mathematical account of Witten's ideas in M Atiyah's book [14].

Shortly after (1990–1991) Witten's results were put on a firm mathematical basis via quantum groups and conformal field theory. The quantum group approach was led by

N Reshetikhin and V Turaev [128; 129; 83; 152], who redefined Witten's quantum invariants using the machinery of quantum groups and modular tensor categories (see Section 16).

One first had to study perturbative expansions of the invariants rather than their exact expressions before the gauge/string duality would become apparent. Coefficients of such expansions represent another kind of knot invariants known as finite-type or Vassiliev invariants introduced in 1990 by V Vassiliev from completely different considerations. The connection between Chern–Simons theory and Vassiliev invariants along this line was established in the 1991 PhD thesis of D Bar-Natan (see also the subsequent papers [23; 24; 22]). The idea was to apply the the 'Feynman rules' from quantum field theory to the 'path integral' on the space of connections describing Chern–Simons theory (see Sections 14 and 15 for more details). Mathematically the universal finite type invariant now known as the Kontsevich integral was given in 1992 by M Kontsevich [90].

Similar developments also occurred in the field of 3–manifolds. In 1991 S Axelrod and I Singer formalized the Feynman integral construction for perturbative Chern–Simons 3–manifold invariants and in 1994 M Kontsevich demonstrated that his integral proved to be a universal finite-type invariant [88]. Finite-type invariants for 3–manifolds were introduced by T Ohtsuki in 1996 [117] and in in 1998 T Le, J Murakami and T Ohtsuki gave a detailed construction of a Kontsevich-type invariant for 3–manifolds and proved its universality among the Ohtsuki invariants for integral homology spheres [93]. Now known as the LMO invariant it is believed to capture the trivial connection contribution to Witten's quantum invariant (see Rozansky[132] and Bar-Natan–Garoufalidis–Rozansky–Thurston [26]).

Rapid progress on the complex side of the conjecture was initiated by Witten's 1991 paper '2D Gravity and Intersection Theory on Moduli Space' [160]. In this paper Witten defines what are now called tautological classes on the moduli space of stable algebraic curves and gives string, dilaton and divisor equations that are sufficient to compute all of the corresponding intersection numbers (see Section 4, Hori et al [76] or Vakil [154]). These intersection numbers could be included in the framework of invariants of symplectic manifolds introduced in M Gromov's 1985 paper [69] on pseudoholomorphic curves (Gromov was more interested in topological applications).

The next year M Kontsevich provided a proof of Witten's conjectured equations [87] using a presentation of the moduli space by ribbon graphs and reducing the generating function for the intersection numbers to a matrix integral of Gaussian type [88]. This can be considered the first mathematical instance of Large N Duality. The techniques of matrix models and graph combinatorics have become indispensable in computations related to Large N Duality.

In 1994 M Kontsevich generalized the definition of stable algebraic curves to stable maps [89] and paved the way to general definitions of (closed) Gromov–Witten invariants introduced by various authors in 1996 (see the discussion in McDuff–Salamon [108] or Cox–Katz [43]). Another important achievement of this paper is the discovery of the 'localization', technique for computing Gromov–Witten invariants (see Section 6). The basic idea dates back to the Atiyah–Bott paper [15] that shows how to compute an integral of functions equivariant under a torus action by localizing to an integral over the fixed points of the action. In 1997 this was generalized to 'virtual localization' by Graber and Pandharipande [66].

The third and final Witten paper that we ought to discuss 'Chern–Simons gauge theory as a string theory' was made available in 1992 although the first printed version appeared only in 1995 in the Floer Memorial Volume [161]. It contains an outline of the first step in the Large N Duality between the Chern–Simons gauge theory and a theory of open strings, so-called 'holomorphic instantons at infinity'. The second and more complicated step in 't Hooft's large N program that involves transition from open to closed strings had to wait until later.

As described above, quantum invariants of knots and links correspond to expectation values in Chern–Simons theory. Witten took the next step and along with a 3–manifold M considered its cotangent bundle T^*M with its natural symplectic structure. This allows one to define (pseudo-)holomorphic curves (stable maps) in T^*M and 'count' them with Gromov–Witten invariants. For this idea to work it is important that the dual threefold is a Calabi–Yau so that holomorphic curves (stable maps) are generically isolated and can be counted. Holomorphic curves ending on the zero section of T^*M are supposed to be the holomorphic instantons of Witten's theory. Witten argues moreover that the duality is exact, that is, it holds not only for large but for all N. The last property is due to the topological nature of the Chern–Simons theory. The partition function of this theory corresponds to the generating function of the Gromov–Witten invariants of holomorphic curves which is dual to the generating function of Chern–Simons invariants of knots and links.

There is one major problem with this picture. A cotangent bundle is a Calabi–Yau but a very degenerate one, in particular there are no non-constant holomorphic curves there either closed or ending on the zero section. This circumstance was well known to Witten and is explicitly pointed out in [161]. Recall that in 1992 there was no notion of Gromov–Witten invariants and even the modern 'translation' of holomorphic instantons as stable maps is incomplete. The physical notion is broader and includes objects like framed trivalent graphs, in particular framed knots and links in the zero section, a.k.a. instantons at infinity. One can think of them as made of infinitely thin ribbons thus representing degenerate Riemann surfaces with boundary. Even today we

do not have a grand theory that would incorporate such degeneracies. Some potential contenders have emerged recently (for example symplectic field theory, see more in the last Section 20). In a way, one can view the theory of Chern–Simons link invariants as 'the Gromov–Witten theory' of cotangent bundles. Building up on this idea R Gopakumar and C Vafa completed the 't Hooft's program by finding in 1998 a dual theory of closed strings that turned out to be a 'true' Gromov–Witten theory but not on T^*M.

1.3 The Gopakumar–Vafa conjecture (1998–2003)

Recall that string theories may change smoothly as the target space passes through some mild singular transitions. Gopakumar and Vafa conjectured that the dual theory of closed instantons lives on a threefold obtained from T^*M via such a transition. They succeeded in finding the dual threefold for $M = S^3$ [65]. The corresponding transition known only to few algebraic geometers from Clemens' paper [41] was very explicitly described in 1990 by P Candelas and X De La Ossa [39]. It involves the zero section for T^*S^3 shrinking to a nodal point and then getting resolved into an exceptional $\mathbb{C}\mathrm{P}^1$. Their terminology of deformed and resolved conifolds and the schematic picture of the conifold transition have since migrated from one paper to the next. They also showed that all three actors in the conifold transition admit Calabi–Yau metrics and computed them explicitly. The manifold on the resolved side of the transition is just the sum $\mathcal{O}(-1) \oplus \mathcal{O}(-1)$ of two tautological bundles over $\mathbb{C}\mathrm{P}^1$. It is now called the resolved conifold even by mathematicians.

Intuitively, as the zero section in T^*S^3 collapses into the nodal point the 'open curves' that end on it close up and stay closed after the resolution of singularity. Thus, Gopakumar and Vafa conjectured that Witten's open string theory on T^*S^3 turns into the usual Gromov–Witten theory of closed holomorphic curves on $\mathcal{O}(-1) \oplus \mathcal{O}(-1)$, in short:

$$\text{CS theory on } S^3 \xrightarrow{\text{large } N \text{ Duality}} \text{GW theory on } \mathcal{O}(-1) \oplus \mathcal{O}(-1)$$

This was the original meaning of the Gopakumar–Vafa conjecture on the level of partition functions.

Physicists make many claims like this; one reason this particular conjecture generated so much excitement in the mathematical community is that mathematical machinery was just mature enough to define both sides of the duality (if not the connection between them) in rigorous terms.

Another reason is that Gopakumar and Vafa did not stop at a general physical claim but made two important and completely mathematical predictions. First, based on Witten's

Chern–Simons computations they predicted an explicit form of the Gromov–Witten free energy function (and thus all closed Gromov–Witten invariants) for the resolved conifold. This can be compared to the prediction made by another physical duality, Mirror Symmetry, for the Gromov–Witten invariants of projective quintics (see Cox–Katz [43] and Hori et al [76]). Second, although Gromov–Witten invariants themselves are rational numbers (see Section 5) they can be represented as combinations of different numbers, later named the Gopakumar–Vafa invariants, that are integers. This integrality prediction later replaced the original meaning of 'the Gopakumar–Vafa conjecture' among algebraic geometers. The environment was so ripe that the prediction of the Gromov–Witten free energy was verified the same year by C Faber and R Pandharipande [55] by an explicit localization computation. The integrality conjecture for the resolved conifold was proved by A Okounkov and R Pandharipande in 2003 [118].

On the Chern–Simons side of the duality significant progress was made in computational techniques. In 1999 S Garoufalidis, D Bar-Natan, L Rozansky and D Thurston discovered an effective algorithm for computing the LMO invariant [26; 25]. It was called the Århus integral (in honor of the city in Denmark where they started their work in 1995) and uses the graphical calculus of Bar-Natan and formal Gaussian integration. The same year R Lawrence and L Rozansky gave a representation for the $\mathrm{SU}(2)$ LMO invariant in terms of integrals and residues distinguishing contributions from different flat connections. This type of representation led to the discovery by Mariño [104] of a relation between the LMO invariants and matrix integrals.

The next year D Bar-Natan and R Lawrence used surgery and the Århus integral to give an explicit formula for the LMO invariant of Seifert fibered homology spheres [27]. Contributions from nontrivial flat connections for Seifert fibered spaces were determined in 2002 by M Mariño using physical arguments [104]. Finally, in 2002 combinatorial expressions for the Reshetikhin–Turaev version of Witten's invariants associated to an arbitrary compact Lie group were derived for Seifert fibered spaces by S Hansen and T Takata [71].

The predictions in Gopakumar–Vafa [65] did not involve mathematical expectations of Chern–Simons observables, that is, quantum link invariants. This was rectified in a 2000 paper by H Ooguri and C Vafa [120], who conjectured that to each framed knot in S^3 there corresponds a Lagrangian submanifold in the resolved conifold and polynomial invariants of the knot give the generating function of open Gromov–Witten invariants. The latter 'count' open holomorphic curves with boundaries on the Lagrangian submanifold. For the case of the unknot Ooguri and Vafa gave an explicit construction of the corresponding Lagrangian submanifold in $\mathcal{O}(-1)\oplus\mathcal{O}(-1)$, and predicted the generating function of open invariants and their integral structure analogous to the Gopakumar–Vafa invariants in the closed case. A number of explicit

computations for nontrivial knots followed on the Chern–Simons side (see Labastida–Mariño–Vafa [92] and references therein).

Unfortunately, the situation on the Gromov–Witten side did not develop as successfully. For one thing, unlike in the closed case the definition of the open Gromov–Witten invariants was lacking. Also, it was not clear how to generalize the Ooguri–Vafa construction of the Lagrangian to nontrivial knots. Nevertheless, in 2001 two groups of researchers succeeded in verifying the unknot predictions of Ooguri and Vafa. To compute the as-of-yet undefined invariants S Katz and C-C M Liu used the virtual localization technique as applied by Faber and Pandharipande in the closed case. T Li and Y S Song on the other hand avoided the use of open invariants altogether [95] by replacing them with relative Gromov–Witten invariants, the theory of which was developed by J Li at the same time [96]. A year later M Liu gave a rigorous definition of open invariants for the case when the Lagrangian submanifold is invariant under a torus action [98].

On the other front J M F Labastida, M Mariño and C Vafa generalized the Ooguri–Vafa construction to all algebraic knots in 2000 [92] and in 2002 C H Taubes extended it even to all symmetric knots [149]. S Koshkin later gave a different construction for a Lagrangian associated to a knot that is valid for any knot [91]. By the time of the Banff workshop in 2003 the time seemed right for a general theory of open invariants to emerge and predictions of the Large N Duality to be verified. However, this did not happen.

The difficulties proved to be much more significant than in the closed case: Lagrangian submanifolds for nontrivial knots do not admit convenient torus actions so the standard computational techniques do not apply; moreover, the available definition of invariants does not work in those cases. At the same time (2003) M Aganagic, A Klemm, M Mariño and C Vafa discovered that all closed Gromov–Witten invariants of toric Calabi–Yau threefolds could be computed by 'slicing' them along Lagrangians corresponding to framed unknots. Their algorithm known as 'the topological vertex' [3] captured the attention of mathematicians and shifted the focus away from general knots. Shortly after, the topological vertex was restated in a rigorous form by J Li, C-C M Liu, K Liu and J Zhou [94] using relative invariants and thus eliminating a need for open ones altogether as far as toric varieties are concerned.

Thus 2003 closes the Gopakumar–Vafa period of research and we conclude our historical excursion. Progress made after 2003 went in a different direction and we shall give some indications about this in the last section of this paper.

2 Overview

In this section we introduce some minimal notation to explain in a nutshell what the rest of this fat paper is all about. The Gopakumar–Vafa Large N Duality [65] is a correspondence between two theories, one defined on S^3 and the other on $\mathcal{O}(-1)\oplus\mathcal{O}(-1)$. Here $\mathcal{O}(-1)$ is the tautological line bundle over $\mathbb{C}\mathrm{P}^1$ defined by

$$\mathcal{O}(-1) := \left\{([z_1 : z_2], w_1, w_2) \in \mathbb{C}\mathrm{P}^1 \times \mathbb{C}^2 \;\middle|\; \begin{vmatrix} w_1 & w_2 \\ z_1 & z_2 \end{vmatrix} = 0\right\}.$$

The projective line $\mathbb{C}\mathrm{P}^1$ is the collection of complex lines through the origin in $\mathbb{C}^2$ and the $\mathcal{O}(-1)$ bundle is simply the collection of pairs consisting of a line together with a point on that line (hence the name 'tautological line bundle'). The projection just maps each point to the corresponding line. The number -1 refers to the fact that the first Chern class of this bundle evaluates to -1 on the $\mathbb{C}\mathrm{P}^1$ cycle.

Intuitively, the large N correspondence can be traced to a geometric relation between the two spaces called the conifold transition $T^*S^3 \rightsquigarrow \mathcal{O}(-1)\oplus\mathcal{O}(-1)$. To understand this transition consider $T^*S^3 \simeq TS^3$ realized as $\mathrm{SL}_2\mathbb{C}$ in $\mathbb{C}^4 \simeq \mathrm{End}(\mathbb{C}^2)$:

$$T^*S^3 \simeq \left\{W = \begin{pmatrix} w_1 & w_2 \\ w_3 & w_4 \end{pmatrix} \in \mathrm{End}(\mathbb{C}^2) \;\middle|\; \det(W) = 1\right\} = \mathrm{SL}_2\mathbb{C}. \tag{1}$$

Exercise 2.1 Consider the standard embedding $S^3 \hookrightarrow \mathbb{R}^4$ as the unit sphere inducing the embedding $TS^3 \hookrightarrow \mathbb{R}^4 \times \mathbb{R}^4 \simeq \mathbb{C}^4$ and find an explicit automorphism $\mathbb{C}^4 \to \mathbb{C}^4$ that restricts to a diffeomorphism $TS^3 \to \mathrm{SL}_2\mathbb{C}$ and $S^3 \to \mathrm{SU}(2)$ (see Koshkin [91]).

Now set $\det(W) = \mu$ in equation (1). Taking $\mu \to 0$ produces a complex deformation of T^*S^3 into a singular variety that physicists and mathematicians call the conifold (it has an ordinary double point at the origin and is a higher dimensional analog of the usual double cone as in Figure 2.1):

$$\check{X}_{S^3} := \left\{\begin{pmatrix} w_1 & w_2 \\ w_3 & w_4 \end{pmatrix} \in \mathrm{End}(\mathbb{C}^2) \;\middle|\; \det(W) = 0\right\}.$$

The conifold admits a small resolution $X_{S^3} \xrightarrow{\rho} \check{X}_{S^3}$ of the singularity (see Harris [73]) small meaning that the exceptional locus $\rho^{-1}(0)$ is a curve rather than a surface as in the case of a blow-up. This resolution is the resolved conifold.

Definition 2.2 The resolved conifold is

$$X_{S^3} = \left\{([z_1 : z_2], w_1, w_2, w_3, w_4) \in \mathbb{C}\mathrm{P}^1 \times \mathbb{C}^4 \;\middle|\; \begin{vmatrix} w_1 & w_2 \\ z_1 & z_2 \end{vmatrix} = \begin{vmatrix} w_3 & w_4 \\ z_1 & z_2 \end{vmatrix} = 0\right\}.$$

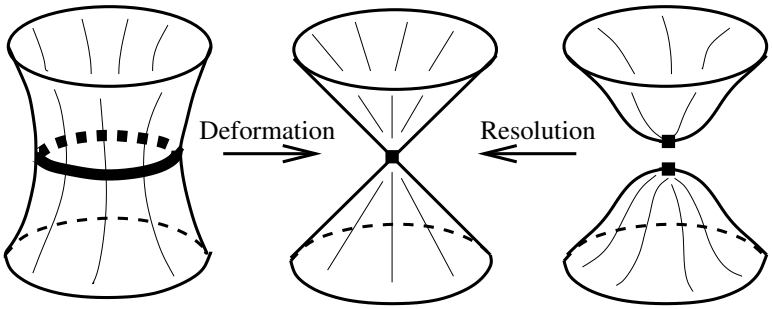


Figure 2.1: The 'conifold' transition two dimensions down $S^1\times\mathbb{R}^1 \rightsquigarrow S^0\times\mathbb{R}^2$

The resolved conifold X_{S^3} is easily seen to be biholomorphic to $\mathcal{O}(-1)\oplus\mathcal{O}(-1)$.

Geometrically the deformation shrinks the zero section (ie S^3) into the double point and the resolution replaces it with an exceptional $\mathbb{C}\mathrm{P}^1$ so topologically we have the surgery $S^3\times\mathbb{R}^3 \rightsquigarrow \mathbb{R}^4\times S^2$. A low-dimensional analog of the conifold transition is shown in Figure 2.1.

The resolved conifold X_{S^3} is a Kähler manifold and one can show that the expected dimension of the space of stable maps (holomorphic curves) from a Riemann surface into X_{S^3} is zero. Thus such maps are (formally) isolated and it makes sense to count them. The Gromov–Witten invariant $N_{g,d}(X_{S^3})$ is intuitively the number of maps of genus g surfaces representing the homology class $d[\mathbb{C}\mathrm{P}^1]$. The actual numbers can be fractional because one must assign fractional weights to curves with automorphisms. Invariants defined in this way can be conveniently collected into the full Gromov–Witten free energy (see Part I for details):

$$F^{\mathrm{GW}}_{X_{S^3}}(t,y) := \sum_{g=0}^{\infty}\sum_{d=0}^{\infty} N_{g,d}(X_{S^3})e^{-td}\,y^{2g-2}.$$

On the other side, the unnormalized Chern–Simons free energy is the logarithm of its partition function $F^{\mathrm{CS}}_{S^3} := \ln Z^{\mathrm{CS}}(S^3)$. Again intuitively, the latter is the value of the path integral:

$$Z^{\mathrm{CS}}(S^3) = \int_{\mathcal{A}} e^{iCS(A)}\mathcal{D}A,$$

taken over the space of connections on a trivial $\mathrm{SU}(N)$ bundle over S^3. In this formula

$$CS(A) := \frac{k}{4\pi}\int_M \mathrm{Tr}(A\wedge dA + \tfrac{2}{3}A\wedge A\wedge A)$$

is the $SU(N)$ Chern–Simons action, where k is an arbitrary integer called the level. Thus the unnormalized Chern–Simons free energy is a function of two parameters k, N

encoded as N and $x := \frac{2\pi}{k+N}$ just as the full Gromov–Witten free energy is a function of t, y.

Today there is a multitude of rigorous constructions that define invariants $Z^{\mathrm{CS}}(M)$ for any 3–manifold M that have the properties that one would conjecture based on heuristic path integral manipulations. Most of them use the fact due to Lickorish and Kirby that any 3–manifold M can be presented as a surgery on a framed link L_M and two different links present the same 3–manifold if and only if they are related by a sequence of the so-called Kirby moves (see Prasolov–Sossinsky [124] and Turaev [152]). Therefore, if one can come up with a framed link invariant that does not change under the Kirby moves one gets a 3–manifold invariant. N Reshetikhin and V Turaev were the first ones to come up with a systematic procedure for constructing Kirby-move invariant link invariants. Their invariants were based on the theory of quantum groups [128; 129]. We use a version of their construction in Sections 16.7 and 17. The corresponding invariant which is just the THOMFLYP polynomial at roots of unity can also be constructed in a number of other ways: skeins, TQFT, etc that all lead to the same quantity identified with the above partition function. We briefly touch on the skein and TQFT approaches in Appendices C and E respectively.

We now give an intuitive idea how the Gopakumar–Vafa duality arises from the physics of string theory (we are grateful to M Mariño for explaining this to us). As mentioned in the history section physicists work with a very broad notion of 'holomorphic instantons' described by a topological version of string theory known as the 'topological A–model'. Holomorphic instantons live in Calabi–Yau threefolds and can be closed or open, that is, have boundary. In the latter case their boundary lies on 'D-branes', located at ('wrapped around') special Lagrangian submanifolds of the threefold. Open and closed holomorphic curves and stable maps are examples of holomorphic instantons but there are more degenerate ones as well, for example 'instantons at infinity' (see Witten [161]) – trivalent ribbon graphs in Lagrangian submanifolds representing infinitely thin 'curves with boundary'. Physical quantities produced by the theory are called 'string amplitudes' and in good cases they can be identified with Gromov–Witten invariants of the threefolds. The Calabi–Yau condition is needed to make sure that holomorphic instantons are isolated and can be 'counted' with finite amplitudes.

Now consider two extreme cases of this picture. The first case is when there are no 'honest' holomorphic curves as in cotangent bundles T^*M to 3–manifolds since the symplectic form on them is exact. At the same time, the zero section is a special Lagrangian submanifold and knotted trivalent graphs in M (framed knots and links in the simplest case) can be seen as degenerate instantons with boundary. Witten discovered in [161] that the string amplitudes of a cotangent bundle with 'N D-branes

wrapped around the zero section' can be recovered from the quantum SU(N) Chern–Simons invariants of M computable via the surgery prescription from his earlier paper [159]. Moreover, one can also consider instantons ending on conormal bundles to links in M that are also special Lagrangian in T^*M. This time the string amplitudes coincide with the the quantum SU(N) link invariants. In other words, the topological A–model reduces to the quantum Chern–Simons theory in this case and can be viewed as 'the Gromov–Witten theory' of cotangent bundles. K Fukaya gives this idea a more precise meaning in terms of Floer homology in [61].

In the second case there are no D-branes in the picture and the only holomorphic instantons that remain are closed stable maps. This is the case of the resolved conifold and its usual Gromov–Witten theory. A striking feature of string theory discovered by Dixon, Harvey, Vafa and Witten [49] is that physically equivalent models can be set in different 'geometric backgrounds', that is, live on different threefolds. This occurs when the geometric backgrounds are related by special geometric transitions. Whereas physical quantities do not change, the underlying threefold may undergo a singular transition as some of the D-brains and/or holomorphic homology classes collapse or appear. The conifold transition is the simplest example of such a geometric transition.

Open instantons that end on the zero section close up as they shrink to the nodal point and transform into closed holomorphic curves in the resolved conifold. Unlike the zero section conormal bundles to links do not collapse and reappear as Lagrangian submanifolds in the resolved conifold. Instantons at infinity that ended on them therefore transform into open holomorphic curves. String amplitudes computed on both sides of the transition should be the same since the physics does not change. In a nutshell, this is the insight behind the Gopakumar, Ooguri and Vafa predictions of equality between the Chern–Simons 3–manifold and link invariants on one side and closed and open Gromov–Witten invariants on the other (see Gopakumar–Vafa [65] and Ooguri–Vafa [120]).

The name Large N Duality comes from the specific way string amplitudes for instantons at infinity are recovered from the Chern–Simons invariants. One needs to consider the latter not for a specific rank but for *all* ranks N. The resulting function turns out to be analytic in $\frac{1}{N}$ around 0 modulo some logarithmic terms and can be expanded into a Laurent series. The coefficients of this series are the string amplitudes in question. From the perspective of knot theory this means that string amplitudes are given not by the 'exact' invariants (such as Jones or THOMFLYP polynomials) but by the so-called Vassiliev (or finite-type or perturbative) invariants [22; 25]. One should not be misled by the name into believing that the duality holds at large N only. The Laurent coefficients match the Gromov–Witten invariants of the resolved conifold at all powers of $\frac{1}{N}$ and the duality is exact.

One of the most attractive traits of the Large N Duality is its computational power. It is usually much easier to compute gauge theory quantities (partition functions, correlators, etc) than the corresponding string amplitudes. This is due to the apparatus of informal but effective path integral manipulations successfully applied by physicists for quite some time. String theory techniques (such as equivariant localization) are more recent and much more cumbersome. It turns out that contributions from instantons at infinity can be reduced to path integrals and even those contributions of honest holomorphic curves can be represented by path integrals with extra insertions. Thus the computational machinery of the gauge theory becomes available for string theories as well. One remarkable achievement of this approach is the topological vertex algorithm originally derived from the Chern–Simons path integral. This algorithm computes Gromov–Witten invariants of all toric Calabi–Yau threefolds (see Aganagic–Klemm–Mariño–Vafa [3] and Mariño [103]).

At present we are very far removed from a mathematical definition of the topological A–model in anywhere near the generality used by physicists. However, the prediction of the equality of the Gromov–Witten and Chern–Simons free energies or partition functions on X_{S^3} and S^3 is a well-defined mathematical statement. This equality should not be taken too literally, one has to renormalize and change variables to make it work. But it is true that one function can be recovered from the other as Gopakumar and Vafa convincingly demonstrated by correctly predicting the values of the Gromov–Witten invariants of X_{S^3} in [65].

The computation that verifies the Gopakumar–Vafa predictions was originally done by C Faber and R Pandharipande in [55] but it does not cover the Chern–Simons side relying on the formulas obtained by path integral methods. Later papers [67; 103] that compute and compare both free energies skip many of the details to stay within a limited length. In this paper we provide the background material on both theories necessary to understand the structures behind these computations and reproduce the computations themselves (Section 19). The result of comparison can be packaged in the following form:

Theorem 2.3 *The full Gromov–Witten free energy and the unnormalized Chern–Simons free energy are related by*

$$Re\big(F^{GW}_{X_{S^3}}(iNx,x) - F^{CS}_{S^3}(N,x)\big) = \tfrac{5}{12}\ln x + \zeta(3)x^{-2} - \tfrac{1}{2}\ln(2\pi) - \zeta'(-1).$$

Some comments are in order about the form of this formula. First of all, even though the free energies are complex-valued the relevant coefficients in their expansions, that is, the invariants themselves are real so it suffices to consider only the real parts. Secondly,

the free energies fail to be holomorphic in x at zero where the expansions are taken, but they do so in a very minor way. The terms on the right appear as a result of regularization and do not indicate any meaningful discrepancy.

To appreciate how powerful this theorem is note that the full Gromov–Witten free energy encapsulates the Gromov–Witten invariants of the resolved conifold in *all degrees* and *all genera*. In its turn, the Chern–Simons partition function for the Hopf link contains the $\mathfrak{sl}_n\mathbb{C}$ Vassiliev invariants of *all knots and links* in the three-sphere. It should come as no surprise that the duality for 'just' this one example led to computation of the Gromov–Witten invariants for all local Calabi–Yau threefolds [3; 103]. Despite its somewhat unappealing form this formula is a very strong confirmation of the Gopakumar–Vafa conjecture. We finish this introduction by stating a far-reaching generalization of Theorem 2.3 suggested by M Mariño in his Banff lectures.

Conjecture 2.4 *For every rational homology 3–sphere M there exists a large N dual Calabi–Yau threefold X_M such that the Chern–Simons theory on M is equivalent to the Gromov–Witten theory on X_M. In particular, the corresponding invariants can be recovered from each other.*

For a reader who may think that the task of learning so much algebraic geometry and quantum algebra is overwhelming we promise that learning these complex but remarkable structures is well worth the effort. While navigating the deep waters of abstraction the reader should always keep in mind that we are merely computing two complex-valued functions – the free energies of X_{S^3} and S^3.

Part I Gromov–Witten invariants

The theory of Gromov–Witten invariants is the mathematical theory closest to string theory in physics. These invariants arise as generalizations of enumerative invariants. In this part, we will outline the definition of Gromov–Witten invariants and give some sample computations.

The first ingredient in understanding these invariants is the cohomological interpretation of intersection theory. As a simple example consider counting the number of zeros of a degree d polynomial in $\mathbb{C}[x]$, say $p(x)$. The answer is easier when we are using complex coefficients. In more general counting problems the answers will be more uniform if we work in complex projective spaces. Given a degree d polynomial p we can define a function, $f_p\colon \mathbb{C}\mathrm{P}^1 \to \mathbb{C}\mathrm{P}^1$ given by $f_p([z:w]) = [p(z/w)w^d : w^d]$. This induces a map on the second cohomology, $f_p^*\colon H^2(\mathbb{C}\mathrm{P}^1;\mathbb{Z}) \to H^2(\mathbb{C}\mathrm{P}^1;\mathbb{Z})$. Since $H^2(\mathbb{C}\mathrm{P}^1;\mathbb{Z}) \cong \mathbb{Z}$, this map is just multiplication by some integer. This integer is known as the degree of the map and it coincides with the degree of the original polynomial. We can write this as

$$\#(p^{-1}(0)) = \#(f_p^{-1}([0:1])) = \int_{[\mathbb{C}\mathrm{P}^1]} f_p^* \omega_{\mathbb{C}\mathrm{P}^1}.$$

Here # represents a signed count of a set of points in general position. We will later describe methods to address non-generic situations. The integral represents the cap product pairing of homology and cohomology, so $[\mathbb{C}\mathrm{P}^1]$ is the fundamental homology cycle ($\mathbb{C}\mathrm{P}^1$ has a natural orientation coming from the complex structure) and $\omega_{\mathbb{C}\mathrm{P}^1}$ is the orientation class. Using the de Rham model for cohomology, the integral will become an honest integral.

Table 2.1 provides a correspondence between geometric intersections and cohomological operations. We will describe various lines in this table as we use them. General topological folk wisdom suggests thinking via intersections and proving via cohomology.

We now turn to a more serious motivating question: how many lines pass through two generic points in a plane? More generally, how many degree d parameterized curves pass through the 'right' number of points in a plane modulo reparameterization of the domain? We may describe the space of lines with two marked points in the plane as

$$M_{0,2}(\mathbb{C}\mathrm{P}^2, 1[\mathbb{C}\mathrm{P}^1]) = \{u\colon \mathbb{C}\mathrm{P}^1 \to \mathbb{C}\mathrm{P}^2,\, p_1,\, p_2 \mid u_*[\mathbb{C}\mathrm{P}^1] = 1[\mathbb{C}\mathrm{P}^1],\, \overline{\partial} u = 0,\, p_1 \neq p_2 \in \mathbb{C}\mathrm{P}^1\}/\sim.$$

Maps in this set are explicitly given by $u([z:w]) = [az+bw : cz+dw : ez+fw]$. Now count the dimension of this space. There are 6 complex parameters in the definition of

Intersections	Cohomology
A codimension k homology submanifold, A	A cohomology class $\alpha \in H^k(M;\mathbb{Z})$
$\#(A \cap F)$	$\alpha([F]) = \alpha \cap [F] = \int_{[F]} \alpha$
$A \cap B$	$\alpha \cup \beta = \alpha \wedge \beta$
$f^{-1}(A)$	$f^*\alpha$
$\sigma_1^{-1}(\sigma_0(X))$	$c_1(L)$ or $e(E) \in H^r(X)$
$\sigma_0(X)$	Thom class $\Phi \in H^r(E, E-\sigma_0(X))$

Table 2.1: Geometric intersections vs cohomology

our degree one parameterized curve. However the points in the projective plane are only defined up to a scale, so we subtract one parameter. We wish to count two maps as equivalent if they are related by a reparameterization of the domain (this is what the $\sim$ represents in the equation for $M_{0,2}(\mathbb{C}\mathrm{P}^2, 1[\mathbb{C}\mathrm{P}^1])$). The holomorphic isomorphisms (reparameterizations) of $\mathbb{C}\mathrm{P}^1$ are just the linear fractional transformations. More explicitly, we define $(u, p_1, p_2) \sim (v, q_1, q_2)$ to hold if and only if there is a linear fractional transformation, say φ, such that $u = v \circ \varphi$ and $\varphi(p_k) = q_k$. A similar count allows one to conclude that the space of linear fractional transformations has complex dimension 3. It follows that the space of complex projective lines in the complex projective plane has complex dimension two. Adding two points in the domain adds two more complex parameters, so the complex dimension of $M_{0,2}(\mathbb{C}\mathrm{P}^2, 1[\mathbb{C}\mathrm{P}^1])$ is four.

Now we have two natural evaluation maps taking $M_{0,2}(\mathbb{C}\mathrm{P}^2, 1[\mathbb{C}P^1])$ to $\mathbb{C}\mathrm{P}^2$, given by $\mathrm{ev}_k([u, p_1, p_2]) := u(p_k)$. Since $\mathbb{C}\mathrm{P}^2$ is two complex-dimensional the following integral makes sense and represents the number of lines passing through two points:

$$\int_{[\overline{M}_{0,2}(\mathbb{C}\mathrm{P}^2, 1[\mathbb{C}\mathrm{P}^1])]} \mathrm{ev}_1^*\omega_{\mathbb{C}P^2} \wedge \mathrm{ev}_2^*\omega_{\mathbb{C}\mathrm{P}^2}$$

There is an infinite number of lines passing through one fixed point, and there are no lines passing through three generic points in the plane, thus two is the 'right' number of points to mark when counting lines.

Exercise 2.5 Count the dimension of the space of degree d genus zero parameterized curves into $\mathbb{C}\mathrm{P}^2$ modulo reparameterization of the domain. Write an expression similar to the integral above representing the number of such curves through the 'right' number of points. We will outline two different ways to compute these numbers later in Part I.

In the next section we will define the Gromov–Witten invariant $\langle \gamma_1, \dots, \gamma_n \rangle_{g,\beta}^X$ The intuitive interpretation of $\langle \gamma_1, \dots, \gamma_n \rangle_{g,\beta}^X$ is the number of genus g curves in the class

β that intersect the cycles $\Gamma_1, \dots, \Gamma_n$ Poincaré dual to $\gamma_1, \dots, \gamma_n$. In general, this interpretation fails. In fact, $\langle \gamma_1, \dots, \gamma_n \rangle^X_{g,\beta}$ are only rational numbers, not integers. This is because curves must be counted with fracitonal weights to get a correct definition.

3 The coarse moduli space

Gromov–Witten invariants extend the ideas described in the above problem. These invariants may be defined for symplectic manifolds or for projective algebraic varieties. We mostly use the symplectic definition in this section, but give some idea of the algebraic definition at the end.

3.1 The symplectic construction

Recall that a symplectic manifold is a (real) $2n$–dimensional manifold with a 2–form ω such that $d\omega = 0$ and $(n!)^{-1}\omega^{\wedge n}$ is a volume form on X. Any symplectic manifold admits a compatible almost complex structure. An almost complex structure $J \in \Gamma(\mathrm{End}(TX))$ is an endomorphism of the tangent bundle which squares to negative one $J^2 = -I$. Such is compatible with a symplectic form, say ω, if the tensor defined by $g(X, Y) = \omega(X, JY)$ is a Riemannian metric. Any two of g, ω or J uniquely determine the third via the compatibility condition. The standard symplectic structure on $\mathbb{C}$ is given by $\omega = \frac{i}{2} dz \wedge d\bar{z}$ and the standard symplectic structure on $\mathbb{C}\mathrm{P}^1$ is given by

$$\omega = \frac{i\,(dz \wedge d\bar{z} + dw \wedge d\bar{w})}{2(|z|^2 + |w|^2)^2}.$$

Together with the standard complex structure this produces a round metric of radius of $\frac{1}{2}$ on the Riemann sphere $\mathbb{C}\mathrm{P}^1$. There are similar symplectic structures on all complex projective spaces. The standard complex structure on $\mathbb{C}$ considered as a real vector space is just multiplication by i. In the basis $\{1, i\}$ it is given by the matrix

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

In tensor notation it is given by

$$J = \partial_y \otimes dx - \partial_x \otimes dy = i\,\partial_z \otimes dz - i\,\partial_{\bar{z}} \otimes d\bar{z}.$$

In a local chart the complex structure on any complex manifold takes this form (generalized in the obvious way to $\mathbb{C}^n$). The product of two symplectic manifolds, (X_k, ω_k), $k = 1, 2$, is the symplectic manifold $(X_1 \times X_2, p_1^*\omega_1 + p_2^*\omega_2)$ where p_k are the

natural projections. Thus $\mathbb{C}\mathrm{P}^1 \times \mathbb{C}^4$ inherits a natural symplectic structure as a product. This leads to a symplectic structure on our main example, the *resolved conifold*

$$X_{S^3} := \left\{ [z_1, z_2], (w_1, w_2, w_3, w_4) \in \mathbb{C}\mathrm{P}^1 \times \mathbb{C}^4 \;\middle|\; \begin{vmatrix} w_1 & w_2 \\ z_1 & z_2 \end{vmatrix} = \begin{vmatrix} w_3 & w_4 \\ z_1 & z_2 \end{vmatrix} = 0 \right\}.$$

This is our main example because it will turn out to be the large N dual of S^3. We will describe the corresponding moduli space in Section 9, compute the Gromov–Witten invariants in Section 11 and see that this is the large N dual of S^3 in Section 19.

The notion of a holomorphic curve can be generalized from algebraic varieties to symplectic manifolds. A symplectic manifold with a compatible almost complex structure provides exactly the data needed for the target of a pseudoholomorphic curve. A Riemann surface Σ has an associated almost complex structure which is automatically a complex structure denoted by j. The Cauchy–Riemann operator is defined on a map $u\colon \Sigma \to X$ by

$$\overline{\partial} u = \tfrac{1}{2}\left(du + J_u \circ du \circ j\right).$$

By definition, a map u is pseudoholomorphic when $\overline{\partial} u \equiv 0$.

Exercise 3.1 Write out the Cauchy–Riemann operator on maps $f\colon \mathbb{C} \to \mathbb{C}$ using x–y coordinates on the first factor and u–v coordinates on the second assuming the natural symplectic and almost complex structures.

Definition 3.2 A smooth, genus g, n–marked, pseudoholomorphic curve in X is a tuple $(\Sigma, j, u, p_1, \ldots, p_n)$ where Σ is an oriented genus g surface, j is an almost complex structure on Σ, u is a pseudoholomorphic map $u\colon \Sigma \to X$ and $p_k \in \Sigma$ are distinct. A morphism between $(\Sigma, j, u, p_1, \ldots, p_n)$ and $(\Sigma', j', u', p_1', \ldots, p_n')$ is a holomorphic map, $\varphi\colon \Sigma \to \Sigma'$, such that $u' \circ \varphi = u$ and $\varphi(p_k) = p_k'$. A genus g, n–marked, pseudoholomorphic curve in X will be called stable if it has a finite automorphism group.

We can now define the (coarse) moduli space of genus g, n–marked, stable pseudoholomorphic curves in a homology class $\beta \in H_2(X;\mathbb{Z})$ to be the set of equivalence classes of such,

$$M_{g,n}(X, \beta) = \{[\Sigma, j, u, p_1, \ldots, p_n] | u_*[\Sigma] = \beta\}/\sim.$$

There are natural evaluation maps, $\mathrm{ev}_k\colon M_{g,n}(X, \beta) \to X$ given by

$$\mathrm{ev}_k([\Sigma, j, u, p_1, \ldots, p_n]) = u(p_k).$$

Using these, we have our first definition of the Gromov–Witten invariants. Let $\gamma_1, \ldots, \gamma_n \in H^*(X;\mathbb{Q})$, $\beta \in H_2(X;\mathbb{Z})$ and define the Gromov–Witten invariants by

$$\langle \gamma_1, \ldots, \gamma_n \rangle^X_{g,\beta} := \int_{[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\mathrm{vir}}} \mathrm{ev}_1^*\gamma_1 \wedge \ldots \wedge \mathrm{ev}_n^*\gamma_n.$$

You will notice that there are many notations in this definition that we have not defined yet. We will slowly compute the Gromov–Witten invariants of $\mathbb{C}\mathrm{P}^2$ (hence how many cubic parameterized curves pass through 8 points etc) and the Gromov–Witten invariants of X_{S^3}. Along the way, we will define the extra notations used in the above formula. For now when you see $[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\mathrm{vir}}$ you should just think $M_{g,n}(X,\beta)$. We will see that the overline refers to a compactification of this space later in this article. The calligraphic font refers to the stack structure on the moduli space. Intuitively the stack structure 'adds' a group to each point of the space in order to have a proper count of points taking symmetry into account. The necessity of stacks is motivated in the first article in Section 5, and the definition of the moduli stack $\overline{\mathcal{M}}_{g,n}(X,\beta)$ is given in the second article of this subsection. The general definition of a stack is given in Appendix A, and the best example is contained in Section 10. The square brackets with vir superscript indicate the virtual fundamental class. This is motivated and defined in Section 9. The short explanation is that when intersections are counted one generally assumes that objects are in general position. However, one can still get sensible answers when the objects are not in general position provided one uses the correct virtual fundamental classes.

To make sense of the integral in the definition we need to have a fundamental cycle to integrate over. This is easiest to establish when the domain of integration is compact. As defined above the coarse moduli spaces would not be compact because we insist that the marked points be distinct. Even without considering marked points these spaces will fail to be compact. To see the problem, consider the family of degree two parameterized curves $u_n\colon \mathbb{C}\mathrm{P}^1 \to \mathbb{C}\mathrm{P}^2$ given by $u_n([s:t]) := [n^{-1}(s^2+t^2) : 2n^{-1}st : s^2-t^2]$. The limit appears to be given by $u_\infty([s:t]) = [0:0:s^2-t^2]$, but this is not a well defined map (consider points where $s = \pm t$). The image of the map u_n in an affine chart is just the hyperbola, $x^2 - y^2 = n^{-2}$. The geometric limit of this sequence is just a pair of lines. This is a clue that motivates the correct definition of a compactification of the moduli space. The correct limit is a map from the one point union of two copies of the complex projective line. The domains and (real part of the) images of u_n and the limit u_∞ are displayed in Figure 3.1. To be precise the limit is defined by

$$u_\infty\colon \mathbb{C}\mathrm{P}^1 \times \{\pm 1\} \to \mathbb{C}\mathrm{P}^2; \quad u_\infty([s:t],\epsilon) := [s : \epsilon s : t]; \quad \epsilon = \pm 1.$$

The splitting off of an extra component in the limit is called bubbling.

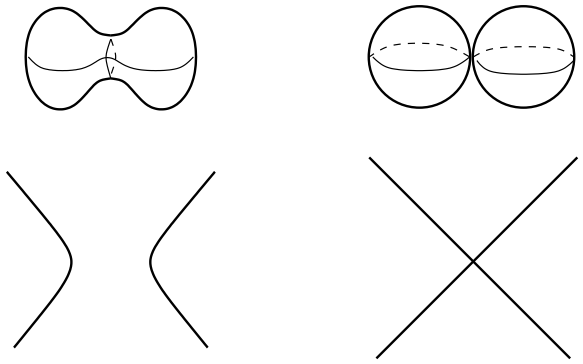


Figure 3.1: Bubbling

To describe domains of stable maps we take a disjoint union of Riemann surfaces (called the normalization) and glue them together along special points called nodes. The resulting 'surface' needs to be connected. These objects are called prestable curves or nodal Riemann surfaces. The original surface without any identifications is called the normalization. The nodes are locally modeled on $\{(x, y)|xy = 0\}$. There is also a notion of a smoothing of a prestable curve obtained by replacing each node by $\{(x, y)|xy = \epsilon\}$. An example of a marked prestable curve together with its normalization and smoothing is shown in Figure 3.2. Precise definitions are given in the paper by Siebert on Gromov–

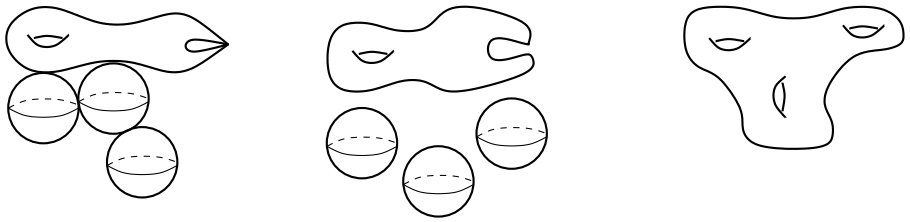


Figure 3.2: Prestable curve, normalization and smoothing

Witten invariants for general symplectic manifolds [141]. In fact, this is a superb reference for the definition of Gromov–Witten invariants in the symplectic category for readers who are more familiar with differential topology than algebraic geometry. While we are on the subject of references, we should mention the book by Hori et al [76], the book by Cox and Katz [43] and the little-known book edited by Aluffi [8]. These are the main references from which this section on Gromov–Witten invariants was derived. For more information see these sources and the references contained therein.

What follows is a more formal definition of a marked prestable curve.

Definition 3.3 (Siebert [141]) A marked prestable curve is a pair, (Σ, p) where Σ is a reduced, compact, connected, one-dimensional, complex, projective variety with no worse than ordinary double point singularities, and p is an n–tuple of pairwise distinct regular points.

An ordinary double point is a point modeled on $\{(x, y)|xy = 0\}$ (see Griffiths–Harris [68] and Hartshorne [75]). Reduced means that there is no locally defined non-zero holomorphic function with a power equal to zero. To see how such a function could arise consider the variety defined by $x^2y = 0$. This looks like a one point union of a pair of lines, but it is not a reduced variety because xy is a non-zero holomorphic function with square zero. The one point union of a pair of lines is represented by the reduced variety defined by $xy = 0$.

A prestable map is a holomorphic map from a prestable curve to a symplectic manifold or projective variety. The definition of a morphism of marked pseudoholomorphic curves, and resulting notions of equivalence, stability and automorphism group extend naturally to prestable curves. This allows one to define a compactification of the moduli space as follows.

Definition 3.4 The (compactified) coarse moduli space of genus g, n–marked curves is

$$\overline{M}_{g,n}(X, \beta) = \Big\{[\Sigma, p, u] \mid u_*[\Sigma] = \beta \text{ and } (\Sigma, p) \text{ is a genus } g, n\text{–marked prestable curve such that } [\Sigma, p, u] \text{ is stable}\Big\}/\sim .$$

It may seem weird that this space is compact when it is required that the marked points be disjoint from each other and the nodes. The limit of a sequence where two or more marked points or nodes collide will be described in the discussion of the boundary divisors and ψ classes in Section 4 on the cohomology of the moduli space. Basically, the idea is that one or more new components bubble off at the point of collision and the marked points move into these bubbles while staying distinct. Before we talk about a compactification, we should introduce a topology on the moduli space. We will conclude this subsection with a description of the topology used in the symplectic category.

In the symplectic case a topology on the space of stable maps can be described via Gromov convergence . The main difficulty in defining it is that the limit map may have a different domain (as we saw in the example depicted in Figure 3.1). To circumvent this difficulty define a resolution $\kappa\colon \widetilde{\Sigma} \to \Sigma$ between two prestable curves to be a map satisfying:

(1) If $p \in \widetilde{\Sigma}$ is a node, then $\kappa(p)$ is a node;

(2) If $q \in \Sigma$ is a node, then $\kappa^{-1}(q)$ is either a node or a circle disjoint from nodes;

(3) If V is any neighborhood of all nodes in Σ, then $\kappa\,|_{\kappa^{-1}(V)}$ is a diffeomorphism onto its image.

One says that a sequence of stable maps (Σ_k, u_k) Gromov converges to a map (Σ, u) if there is a sequence of resolutions $\kappa_k\colon \Sigma_k \to \Sigma$ such that for any neighborhood V of all nodes in Σ:

(1) $u_k \circ \kappa_k^{-1} \xrightarrow[k\to\infty]{} u$ in $C^\infty(\Sigma\backslash V)$;

(2) $d\kappa_k \circ j_k \circ d\kappa_k^{-1} \xrightarrow[k\to\infty]{} j$ in $C^\infty(\Sigma\backslash V)$, where j_k, j are complex structures on Σ_k, and Σ respectively;

(3) $\mathrm{Area}(u_k(\Sigma_k)) \xrightarrow[k\to\infty]{} \mathrm{Area}(u(\Sigma))$.

Equipped with the Gromov topology the moduli space is Hausdorff and compact. One can check that the image homology class $\beta \in H_2(X, \mathbb{Z})$ and the (arithmetic) genus are preserved in this limit. With this topology the coarse moduli space has the structure of a generalization of a manifold called an orbifold. See Siebert [141] and the references contained therein for more details.

Remark 3.5 For genus zero invariants of convex spaces such as $\mathbb{C}\mathrm{P}^n$ the moduli space is in fact a manifold. We will discuss the moduli space as if it were a manifold until Section 5 where we explain where the orbifold singularities arise.

3.2 The algebraic construction

The construction of the coarse moduli space for projective varieties works a bit differently. One can start working from the ground up by studying some examples. The easiest example has X equal to a point (so $\beta = 0$) and the genus equal to zero. This amounts to studying configurations of n distinct points in $\mathbb{C}\mathrm{P}^1$ modulo equivalence by degree one rational maps. Any degree one holomorphic map from $\mathbb{C}\mathrm{P}^1 \to \mathbb{C}\mathrm{P}^1$ takes the form

$$\varphi([z : w]) = [az + bw : cz + dw]$$

where $ad - bc \neq 0$. In affine coordinates on the domain and codomain this takes the form $\varphi(z) = \frac{az+b}{cz+d}$. It is easy to see that the following function for fixed z_0, z_1 and z_2 is a linear fractional transformation taking $z_3 = z_0$ to 0, $z_3 = z_1$ to 1 and $z_3 = z_2$ to ∞.

Definition 3.6 The cross ratio is the function defined by

$$\gamma(z_0, z_1, z_2, z_3) := \frac{(z_1 - z_2)(z_3 - z_0)}{(z_0 - z_1)(z_2 - z_3)}.$$

It is easy to say what the cross ratio means – it is nothing more than the image of the fourth point under the unique linear fractional transformation taking the first three points to 0, 1, and ∞ respectively. We now follow the exposition of D Salamon which utilizes cross ratios to realize $\overline{M}_{0,n}(\mathrm{pt}, 0)$ as projective algebraic varieties (see Salamon [133]). For each tuple (i, j, k, ℓ) of distinct positive integers less than or equal to n define a function, $\gamma_{i,j,k,\ell}\colon M_{0,n}(\mathrm{pt}, 0) \to \mathbb{C}\mathrm{P}^1$ by $\gamma_{i,j,k,\ell}([p]) := \gamma(p_i, p_j, p_k, p_\ell)$. One can easily check that these functions satisfy the relations

$$\begin{aligned}
\gamma_{j,i,k,\ell} &= \gamma_{i,j,\ell,k} = 1 - \gamma_{i,j,k,\ell},\\
\gamma_{i,j,k,\ell}\gamma_{i,k,j,\ell} - \gamma_{i,k,j,\ell} &= \gamma_{i,j,k,\ell},\\
\gamma_{j,k,\ell,m}\gamma_{i,j,k,m} - \gamma_{j,k,\ell,m}\gamma_{i,j,k,\ell} &= \gamma_{i,j,k,m} - 1.
\end{aligned}$$

As a particular case, we see that the cross ratio maps $M_{0,4}$ isomorphically to $\mathbb{C}\mathrm{P}^1 - \{0, 1, \infty\}$. It is natural to guess that this extends to a bijection between $\overline{M}_{0,4}$ and $\mathbb{C}\mathrm{P}^1$, which in fact it does. The stable maps corresponding to 1 and ∞ are displayed on the left of Figure 4.4 in the next subsection. More generally, the cross ratios of the marked points may be used to identify $\overline{M}_{0,n}$ with the projective subvariety of $(\mathbb{C}\mathrm{P}^1)^{\times N}$ specified by the solutions to the above displayed equations in the γ. Here $N = n(n-1)(n-2)(n-3)$ is just the number of possible distinct 4–tuples marked points. In the following aside we continue with a very brief description of the algebraic construction of the coarse moduli spaces of higher genus curves as projective varieties.

Aside 3.7 We can now step things up a bit and consider moduli of higher genus curves. Here our exposition follows that of D Mumford from [114]. Let Σ be a genus $g > 1$ curve and K_Σ be the canonical bundle (top exterior power of the cotangent bundle). Using the Riemann–Roch theorem one can compute that the dimension of $H^0(\Sigma, K_\Sigma^3)$ is $5g-5$. The space $H^0(\Sigma, K_\Sigma^3)$ is just the space of globally defined holomorphic forms of the form $f(z)dz^{\otimes 3}$. Given a basis for $H^0(\Sigma, K_\Sigma^3)$, say $\{\omega_k\}$, define a map to $\mathbb{C}\mathrm{P}^{5g-6}$ by $\phi(z) = [\omega_k(z)]$. Here we use any trivialization of K_Σ^3 around z to identify the $\omega_k(z)$ with complex numbers. Changing the trivialization clearly does not change the projective equivalence class. The Weierstrass points of Σ are defined to be those points in Σ for which the tangent plane to Σ in $\mathbb{C}\mathrm{P}^{5g-6}$ matches to order $5g-5$ or more. There are $g(5g-5)^2$ such points counted with multiplicity, label them by z_j. Now take a large N and consider the following set of functions from $5g-5$ element subsets of $E = \{1, \dots, g(5g-5)^g\}$ to the non-negative integers

$$R = \big\{r : \{I \subset E \,||I| = 5g-5\} \to \mathbb{Z} | r(J) \geq 0 \text{ for all } J \text{ and } \textstyle\sum_{k\in I} r(I) = N\big\}.$$

Define an embedding $M_{g,0} \to \mathbb{C}\mathrm{P}^{|R|-1}$ by

$$[\Sigma] \mapsto \left[\sum_{\sigma \in \mathrm{perm}(E)} \prod_{I \subset E, |I|=5g-5} \left(\det_{j \in I}(\omega_k(z_j))\right)^{r(\sigma(I))} \right].$$

The determinants in the above expression lie in K_Σ^{3N}; they can be interpreted as complex numbers by evaluation in any trivialization.

Exercise 3.8 Assuming that this construction works, jazz it up to define an embedding of $M_{g,n}$ into a sufficiently large projective space. This is difficult, but luckily it is not needed.

There are other approaches to proving that $M_{g,n}$ and $\overline{M}_{g,n}$ admit the structure of quasiprojective and projective varieties respectively, but no way is easy. See Mumford [112] for the standard exposition.

The next step is to describe the structure of $\overline{M}_{g,n}(\mathbb{C}\mathrm{P}^r, d)$. The final step is to define $\overline{M}_{g,n}(X, \beta)$ for general projective varieties X. These last two steps are not so bad. Our exposition comes from the lectures by Aluffi [8]. Given $[\Sigma, p] \in \overline{M}_{g,n+d(r+1)}$ such that the divisors $(p_{n+kd+1} + \cdots + p_{n+kd+d})$ and $(p_{n+\ell d+1} + \cdots + p_{n+\ell d+d})$ are linearly equivalent for $k, \ell = 0, \ldots, r$ and non-zero sections s_k of the line bundle associated to these equivalent divisors such that $s_k(p_{n+kd+1}) = s_k(p_{n+kd+d}) = 0$ one can associate a stable curve $[u, \Sigma, q] \in \overline{M}_{g,n}(\mathbb{C}\mathrm{P}^r, d)$ by $q_j = p_j$ for $j = 1, \ldots, n$ and $u(z) = [s_k(z)]$. It is not hard to see that two pairs $([\Sigma, p], s)$ and $([\Sigma', p'], s')$ produce the same stable map if and only if s and s' agree up to a constant factor and $[\Sigma, p]$ and $[\Sigma', p']$ agree after a permutation in the marked points fixing p_j for $j = 1, \ldots, n$ and each divisor $(p_{n+kd+1} + \cdots p_{n+kd+d})$. The subset of $\overline{M}_{g,n+d(r+1)}$ satisfying the equivalent divisor condition is a subvariety, and the set of data that we have described here forms a $(\mathbb{C}^\times)^{r+1}$ bundle over this subvariety. The quotient of this bundle by the group generated by the change of scale and permutations produces a quasiprojective variety that embeds into $\overline{M}_{g,n}(\mathbb{C}\mathrm{P}^r, d)$ as an open set. Of course we can embed it in a different way by composing each map with a fixed holomorphic isomorphism of $\mathbb{C}\mathrm{P}^r$. The fact is that by choosing a finite number of such isomorphisms we can completely cover $\overline{M}_{g,n}(\mathbb{C}\mathrm{P}^r, d)$. To see this, pick a basis, $\{t_k\}$, for $H^0(\mathbb{C}\mathrm{P}^r, \mathcal{O}(-1))$. Then to any generic $[u, \Sigma, q] \in \overline{M}_{g,n}(\mathbb{C}\mathrm{P}^r, d)$ we associate $([\Sigma, p], t_k \circ u)$, where p is obtained by adjoining the zeros of the $t_k \circ u$ to q.

4 Cohomology of the moduli space

We see that our first definition of Gromov–Witten invariants is just the evaluation of natural cohomology classes on the moduli space. This leads one to ask if there are any other cohomology classes on the moduli space. Indeed there are other interesting

classes. In this subsection we will define some of them and derive several important recurrence relations between the Gromov–Witten invariants. The definition of a new set of cohomology classes appears in the first article and the recurrence relations are described in subsequent ones. The paper by R Vakil [154] is also a good reference for this material.

Here we describe the new cohomology classes as Chern classes of natural vector bundles over the moduli space. In general, as explained in Section 5 one needs a generalization of vector bundles called orbibundles that allow finite quotient singularities. However, in a number of examples singularities do not appear (Remark 3.5). For now we assume that everything is smooth and introduce more general examples in Section 5.

4.1 Gromov–Witten invariants and descendants

We recall the intersection theory definition of the first Chern class of a line bundle. There are many other possible definitions, see Milnor–Stasheff [110], Griffiths–Harris [68] and Bott–Tu [30]. Given a line bundle $L \to X$ and two generic (transverse) sections $\sigma_0, \sigma_1 \colon X \to L$, the first Chern class of the line bundle may be defined to be the cohomology class Poincaré dual to $\sigma_1^{-1}(\sigma_0(X))$. As an example, consider the tangent bundle to the 2–sphere. A section of the tangent bundle is nothing other than a vector field. We can (and generally will) take σ_0 to be the zero section. We can take σ_1 to be a vector field that flows up from the south pole to the north pole. As a set $\sigma_1^{-1}(\sigma_0(S^2))$ consists of exactly two points. One can see that the intersections are transverse, and conclude that $c_1(TS^2)[S^2]$ must be -2, 0 or 2. It is in fact 2.

Exercise 4.1 Write out consistent orientation conventions and verify that the signed count of zeros implies that $c_1(TS^2)[S^2] = 2$.

Exercise 4.2 Prove the following properties of line bundles and their Chern classes.

(1) $c_1(\underline{\mathbb{C}}) = 0$ (Here and elsewhere $\underline{V}$ will denote the trivial bundle with fiber V.)

(2) $c_1(L_1 \otimes L_2) = c_1(L_1) + c_1(L_2)$.

(3) $c_1(f^*L) = f^*c_1(L)$.

(4) $L \otimes L^* \cong \underline{\mathbb{C}}$.

There are n distinguished line bundles defined over $\overline{M}_{g,n}(X, \beta)$, denoted by $\mathcal{L}_k$ for $k = 1, \dots, n$. Intuitively, these bundles are specified by an identification of the fiber over each point as

$$\mathcal{L}_k|_{[u,\Sigma,p]} = T^*\Sigma_{p_k}.$$

This allows one to define new cohomology classes on $\overline{M}_{g,n}(X, \beta)$ and extend the definition of the Gromov–Witten invariants.

Definition 4.3 The ψ–classes are defined by $\psi_k := c_1(\mathcal{L}_k) \in H^2(\overline{M}_{g,n}(X,\beta))$. The descendant Gromov–Witten invariants are defined by

$$\langle \tau_{a_1}(\gamma_1), \ldots, \tau_{a_n}(\gamma_n) \rangle_{g,\beta}^X := \int_{[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\text{vir}}} \psi_1^{a_1} \wedge \text{ev}_1^* \gamma_1 \wedge \ldots \wedge \psi_n^{a_n} \wedge \text{ev}_n^* \gamma_n \,.$$

Here $\gamma_1, \ldots, \gamma_n \in H^*(X;\mathbb{Q})$, $\beta \in H_2(X;\mathbb{Z})$ and the integral of a form over a space is defined to be zero if the degree of the form does not match the dimension of the space.

It is often possible to reduce the computation of Gromov–Witten invariants on one space to a computation of descendant invariants on a smaller space. In addition there are recursion relations relating various descendant invariants.

We should now make the definition of the bundle $\mathcal{L}_k$ more precise. To do so, we need to study the relationship between moduli spaces with different numbers of marked points. There is a natural projection from $\overline{M}_{g,n+1}(X,\beta)$ to $\overline{M}_{g,n}(X,\beta)$ given by ignoring the final point. One subtle point is that after deleting a marked point a stable map with marked points may no longer be stable. This can be fixed by stabilization.

The stabilization of a prestable map, $\text{st}([u,\Sigma,p])$ is defined by identifying to a point any component of the normalization of Σ on which an infinite subgroup of the automorphism group acts effectively (that is, only the identity element fixes everything). This gives,

$$\pi\colon \overline{M}_{g,n+1}(X,\beta) \to \overline{M}_{g,n}(X,\beta);\ \pi([u,\Sigma,p_1,\ldots,p_{n+1}]) := \text{st}([u,\Sigma,p_1,\ldots,p_n])\,.$$

See Figure 4.1 to see the result of projection with nontrivial stabilization.

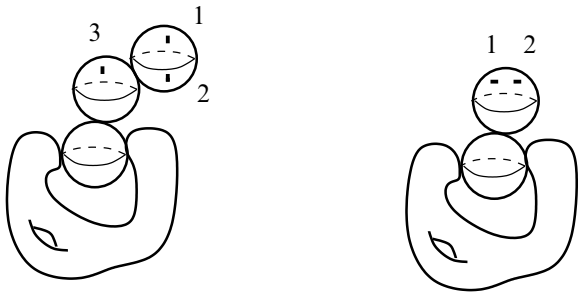


Figure 4.1: Projection from $\overline{M}_{2,3}$ to $\overline{M}_{2,2}$

There are also inclusion maps going in the other direction defined as follows:

$$\rho_k\colon \overline{M}_{g,n}(X,\beta) \to \overline{M}_{g,n+1}(X,\beta);$$
$$\rho_k([u,\Sigma,p_1,\dots,p_n]) := \left[\overline{u}, \Sigma \cup_{p_k=[0:1]} \mathbb{C}\mathrm{P}^1, p_1,\dots,p_{k-1},[1:1],p_{k+1},\dots,p_n,[1:0]\right],$$

where $\overline{u}$ is defined by $\overline{u}|_\Sigma = u$ and $\overline{u}|_{\mathbb{C}\mathrm{P}^1} = u(p_k)$. See Figure 4.2 to see the result of inclusion at a marked point.

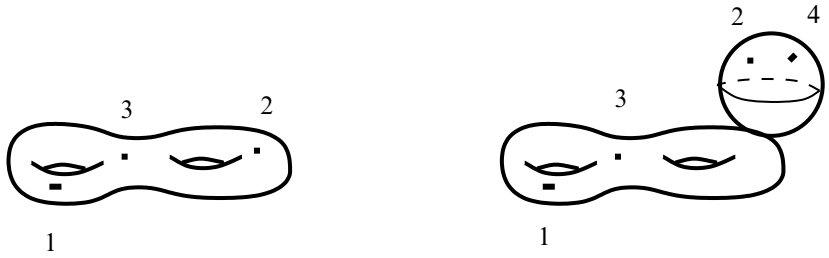


Figure 4.2: Inclusion of $\overline{M}_{2,3}$ in $\overline{M}_{2,4}$ at p_2

As an aside we can use these last two figures to explain the limit of a sequence in which one marked point approaches a node or approaches a different marked point. If a third marked point is added to one of the genus zero components of the curve on the right side of Figure 4.1, the limit as this point approaches the node between the genus zero components will be the curve on the left. If a fourth marked point is added to the curve on the left of Figure 4.2, the limit as this point approaches p_2 is the curve on the right.

Setting $\mathcal{U} = \overline{M}_{g,n+1}(X,\beta)$, the collection of maps,

$$(\pi\colon \mathcal{U} \to \overline{M}_{g,n}(X,\beta), \rho_k\colon \overline{M}_{g,n}(X,\beta) \to \mathcal{U}, \mathrm{ev}_{n+1}\colon \mathcal{U} \to X)$$

is an example of a family of stable maps. The key property is that for every $s_0 \in \overline{M}_{g,n}(X,\beta)$, we have that $[\mathrm{ev}|_{\pi^{-1}(s_0)}, \pi^{-1}(s_0), \rho_1(s_0),\dots,\rho_n(s_0)] \in \overline{M}_{g,n}(X,\beta)$.

Exercise 4.4 Given that s_0 has trivial automorphism group prove that s_0 is isomorphic to the following stable map.

$$[\mathrm{ev}|_{\pi^{-1}(s_0)}, \pi^{-1}(s_0), \rho_1(s_0),\dots,\rho_n(s_0)]\,.$$

This almost implies that $\mathcal{U}$ is a 'universal' family of stable maps. What happens if s_0 has nontrivial automorphisms? See Section 5.2.

We are now ready to give a precise definition of the bundles $\mathcal{L}_k$. This definition will only work as intended when stable maps have no nontrivial automorphisms. To define V in general one needs to use the language of stacks. For starters, define the subbundle V of vertical vectors on $\mathcal{U}$ to be those vectors that are tangent to the fibers of the projection π. Next, recall the definition of the pull-back of a vector bundle. Given a vector bundle $\pi\colon E \to Y$ and a map $f\colon X \to Y$ the pull back is defined by

$$f^*E := \{(x, e) \in X \times E \,|\, f(x) = \pi(e)\}.$$

Now we have,

Definition 4.5 Let

$$V := \ker(\pi_*\colon T\mathcal{U} \to T\overline{M}_{g,n}(X, \beta)).$$

be the vertical subbundle. Then the tautological bundles over the moduli space are $\mathcal{L}_k := \rho_k^* V^*$, where V^* is the dual of V.

We have seen that in the absence of automorphisms the inverse image of a stable map $s_0 \in \overline{M}_{g,n}(X, \beta)$ under the projection from $\overline{M}_{g,n+1}(X, \beta)$ to $\overline{M}_{g,n}(X, \beta)$ is isomorphic to the original stable map s_0. The image $\rho_k(s_0)$ corresponds to the marked point p_k in the isomorphic copy of s_0. It follows that the fiber of $\rho_k^* V$ over $s_0 = [u, \Sigma, p]$ is isomorphic to $T_{p_k}\Sigma$. Thus the fiber of $\mathcal{L}_k$ over a stable map is isomorphic to the cotangent space of the underlying curve at the associated marked point. The analogous construction in the category of stacks will work when there are automorphisms.

4.2 Boundary divisors

We need a way to describe the cohomology of the moduli spaces. By intersection theory we can identify a k–dimensional cohomology class with a real codimension k cycle. One sees that the set of stable maps with one node has complex codimension 1, so may be used to define real dimension 2 cohomology classes. The classes defined in this way are called boundary divisors.

Definition 4.6 Given two disjoint sets A and B such that $A \cup B = \{1, \dots, n+m\}$ with order preserving bijections $j_a\colon \{1, \dots, n\} \to A$ and $j_b\colon \{1, \dots, m\} \to B$, non-negative integers g and h and second cohomology classes α and β in X, we define the boundary divisor $D(g, A, \alpha|h, B, \beta)$ to be the push-forward of the fundamental cycle by the map,

$$\begin{aligned}\overline{M}_{g,n+1}(X, \alpha) \times \overline{M}_{h,m+1}(X, \beta) &\to \overline{M}_{g+h,n+m}(X, \alpha + \beta);\\ ([u, \Sigma, p], [u', \Sigma', p']) &\mapsto [u \cup u', \Sigma \cup_{p_{n+1}=q_{m+1}} \Sigma', r],\end{aligned}$$

where $r_{j_a(k)} := p_k$ and $r_{j_b(k)} := q_k$.

The boundary divisor $D(g, A, \alpha | h, B, \beta)$ corresponds to the closure of the subset of moduli space of the set of stable maps with one node joining two irreducible components having data g, A, α and h, B, β respectively. See Figure 4.3. There is an additional

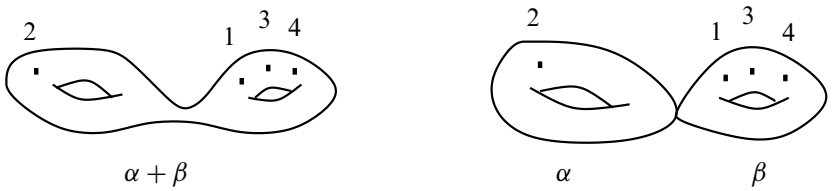


Figure 4.3: Degeneration to the boundary divisor $D(1, \{2\}, \alpha | 1, \{1, 3, 4\}, \beta)$

type of boundary divisor corresponding to the degeneration of a non-separating simple closed curve in the domain. This divisor is typically denoted by Δ_0.

Exercise 4.7 Give a formal definition of Δ_0.

When factors in the definition of a boundary divisor are clear from the context we will leave them out of the notation. For example $D(\{1, 3\}|\{2, 4\})$ in $\overline{M}_{0,4}$ is short-hand notation for $D(0, \{1, 3\}, 0|0, \{2, 4\}, 0)$, as there is no possible way to have nontrivial genus or homology in this case. There is an additional boundary divisor obtained by identifying two points on one curve, or equivalently degenerating a non-separating simple closed curve. This divisor is typically denoted by Δ_0. The map combining two curves into one nodal curve is very similar to constructions gluing two surfaces with boundary along a boundary component. This latter operation is common in the analysis of topological quantum field theories. As an example consider the divisors $D(\{1, 2\}|\{3, 4\})$ and $D(\{1, 3\}|\{2, 4\})$ in $\overline{M}_{0,4}$. We have identified $\overline{M}_{0,4}$ with $\mathbb{C}\mathrm{P}^1$ via the cross ratio. Under this identification we are mapping the first point to zero, the second to one and the third to infinity and then looking at the coordinates of the last point. Thus $D(\{1, 2\}|\{3, 4\})$ is identified with infinity and $D(\{1, 3\}|\{2, 4\})$ is identified with one, so these divisors are distinct. They are however linearly equivalent, which implies that they represent the same cohomology class. The equality of these cohomology classes is the starting point for the Witten–Dijkgraaf–Verlinde–Verlinde equations (WDVV equations). The WDVV equations are similar to a consequence of the fusion rule for monoidal functors, see Figure 4.4.

This similarity is one good reason to believe that there may be some relationship between Gromov–Witten invariants and Chern–Simons invariants.

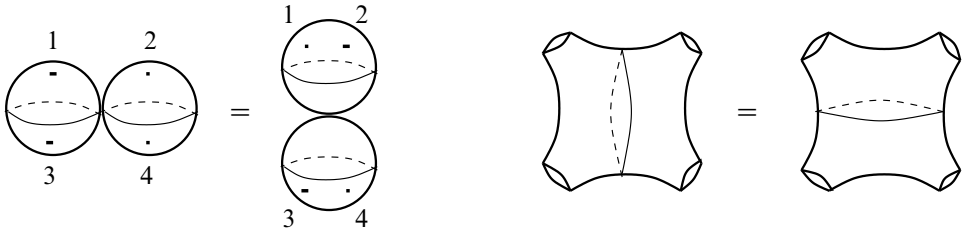


Figure 4.4: The WDVV equation and the fusion rule

The linear equivalence of these two simple divisors is more powerful than it may first appear. The projection map from $\overline{M}_{0,n}(X,\beta)$ to $\overline{M}_{0,4}$ defined via stabilization whenever $n \geq 4$ translates this equivalence into many more moduli spaces. To demonstrate the power of this equivalence, the next exercise outlines a proof of Kontsevich's recursion for the genus zero Gromov–Witten invariants of $\mathbb{C}\mathrm{P}^2$. The answer to this exercise is explained nicely in the book by Hori et al [76].

Exercise 4.8 Let N_d be the Gromov–Witten invariant

$$\langle \mathrm{pt}, \dots, \mathrm{pt} \rangle_{0,d[\mathbb{C}\mathrm{P}^1]}^{\mathbb{C}\mathrm{P}^2}.$$

This is the number of degree d parameterized curves that pass through $3d-1$ generic points in the plane. Let $\pi\colon \overline{M}_{0,3d}(\mathbb{C}\mathrm{P}^2, d[\mathbb{C}\mathrm{P}^1]) \to \overline{M}_{0,4}$ be the standard projection. Finally, let Y be a cycle in $\overline{M}_{0,3d}(\mathbb{C}\mathrm{P}^2, d[\mathbb{C}\mathrm{P}^1])$ representing

$$\mathrm{ev}_1^*[\mathbb{C}\mathrm{P}^1]\mathrm{ev}_2^*[\mathbb{C}\mathrm{P}^1]\mathrm{ev}_3^*[\mathrm{pt}]\dots\mathrm{ev}_{3d}^*[\mathrm{pt}].$$

We are denoting cohomology classes by their Poincaré duals. Of course one should take a generic collection of two hyperplanes and $3d-2$ points when writing a representative for Y.

(1) Express $\pi^* D(\{1,2\}|\{3,4\})$ and $\pi^* D(\{1,3\}|\{2,4\})$ in terms of boundary divisors on $\overline{M}_{0,3d}(\mathbb{C}\mathrm{P}^2, d[\mathbb{C}\mathrm{P}^1])$.

(2) Show that $\#(Y \cap D(\{1,2\}, 0|\{3,\dots,3d\}, d[\mathbb{C}\mathrm{P}^1])) = N_d$.

(3) Argue that $\#(Y \cap D(A, e|B, f))$ must be zero unless B has the 'right' number of points. When B does have the right number of points construct a covering projection

$$Y \cap D(A,e|B,f) \to \big(\overline{M}_{0,3e-1}(\mathbb{C}\mathrm{P}^2, e[\mathbb{C}\mathrm{P}^1]) \cap \mathrm{ev}_1(q_1) \cap \dots \mathrm{ev}_{3e-1}(q_{3e-1})\big) \times \big(\overline{M}_{0,3f-1}(\mathbb{C}\mathrm{P}^2, f[\mathbb{C}\mathrm{P}^1]) \cap \mathrm{ev}_1(r_1) \cap \dots \mathrm{ev}_{3f-1}(r_{3f-1})\big).$$

(4) Express $\#(Y \cap \pi^* D(\{1,2\}|\{3,4\}))$ in terms of the numbers N_e. Hint: When computing $\#(Y \cap D(A, e|B, f))$ recall that a degree e curve intersects a degree f curve in ef points when you are enumerating the locations of the node and the two marked points that lie on the lines.

(5) Express $\#(Y \cap \pi^* D(\{1,3\}|\{2,4\}))$ in a similar way.

(6) Use the linear equivalence to deduce a recurrence relation between the various N_d, and test your recurrence by computing a few values. You should get $N_2 = 1$, $N_3 = 12$, and $N_4 = 620$.

If these Gromov–Witten invariants are put into a suitable generating function the recurrence relation will translate into a differential equation. One important example of this allows one to conclude that the descendant Gromov–Witten invariants of a point are encoded into the solutions of the KdV equation. This may be considered as one of the first tests of Large N Duality for the case of zero-dimensional space. We however, will not emphasize this aspect of Large N Duality here. See the paper by Witten on 2D gravity [160] for a discussion of this when it was still a conjecture and the paper by Kontsevich for a proof [87].

4.3 The string and dilaton equations

We now turn to a description of the recursion relations between the ψ–classes. For clarity, we will denote the first ψ class on $\overline{M}_{g,n}(X,\beta)$ by $\psi_1^{(n)}$ when we are comparing ψ–classes on $\overline{M}_{g,n}(X,\beta)$ and $\overline{M}_{g,n+1}(X,\beta)$. We will use a similar notation for the tautological bundles. We have by definition,

$$\psi_1^{(n+1)} - \pi^* \psi_1^{(n)} = c_1(\mathcal{L}_1^{(n+1)} \otimes (\pi^* \mathcal{L}_1^{(n)})^*)\,.$$

To compute this Chern class, we just need to construct a section of the bundle. A section of this bundle may be viewed as a vector bundle morphism, $s\colon \rho_1^* V^{(n+1)} \to \pi^* \rho_1^* V^{(n)}$. We can view tangent vectors to the universal bundle as paths of stable maps. A natural morphism is given by

$$s([u,\Sigma,p_1,\dots,p_{n+1}],[u,\Sigma,p_1,\dots,p_{n+1},q(t)]) := \\ ([u,\Sigma,p_1,\dots,p_{n+1}],\mathrm{st}[u,\Sigma,p_1,\dots,p_n],\underline{\mathrm{st}}[u,\Sigma,p_1,\dots,p_n,q(t)]).$$

Here $q(t)$ is a smooth path in Σ with $\lim_{t\to 0} q(t) = p_1$, and the above formula is correct when $t \neq 0$. The underline on the second stabilization refers to the fact that it is the result of stabilizing the $[u,\Sigma,p]$ and then placing the $q(t)$ at the corresponding point of the stabilization. The above formula appears complicated because of the pull-backs in the definitions of the two bundles. This bundle map is identically zero

over the boundary divisor $D(0,\{1,n+1\},0|g,\{2,\dots,n\},\beta)$ because the entire bubble containing p_1 and p_{n+1} collapses to a point when p_{n+1} is forgotten. The resulting map is stabilized and thus the path $q(t)$ becomes constant. In fact, this is the only way that this section can be identically zero over some point. One concludes

$$\psi_1^{(n+1)} - \pi^*\psi_1^{(n)} = D(0,\{1,n+1\},0|g,\{2,\dots,n\},\beta).$$

The recursion between cohomology classes that we just derived implies three important recursion relations between descendant Gromov–Witten invariants. We will prove two of these recursions; the string equation and the dilaton equation. This will require use of the Thom isomorphism and the Euler class of a vector bundle.

Recall that relative cohomology classes may be represented as the Poincaré duals of closed cycles. Given a real rank r vector bundle, $\pi\colon E\to X$ over a closed base, there is a natural relative cohomology class, $\Phi\in H^r(E,E-\sigma_0(X))$ (the Thom class) given as the Poincaré dual of $\sigma_0(X)$. The Thom isomorphism is the map

$$\begin{aligned} H^k(X) &\to H^{r+k}(E,E-\sigma_0(X)) \\ \alpha &\mapsto \Phi\cup\pi^*\alpha. \end{aligned}$$

The Euler class of the vector bundle is the pull-back of the Thom class by a section, $e(E):=\sigma_1^*\Phi\in H^r(X)$. Notice that the Euler class of the real bundle underlying a complex line bundle is the first Chern class of the line bundle. It is easier to put all of this on a firm theoretical foundation by using cohomology; however the conceptual picture is harder to follow. See for example Milnor–Stasheff [110] Spanier [147].

Using this technology we will establish two important recursion relations for Gromov–Witten invariants. The first is the string equation,

$$\langle\tau_{a_1}(\gamma_1)\dots\tau_{a_n}(\gamma_n)\tau_0(X)\rangle_{g,\beta}^X = \sum_{k=1}^n\langle\tau_{a_1}(\gamma_1)\dots\tau_{a_k-1}(\gamma_k)\dots\tau_{a_n}(\gamma_n)\rangle_{g,\beta}^X,$$

and the second is the dilaton equation,

$$\langle\tau_{a_1}(\gamma_1)\dots\tau_{a_n}(\gamma_n)\tau_1(X)\rangle_{g,\beta}^X = (2g-2+n)\langle\tau_{a_1}(\gamma_1)\dots\tau_{a_n}(\gamma_n)\rangle_{g,\beta}^X.$$

We can now prove the string equation (we denote cycles and their Poincaré duals by the same symbols). To start the computation, we use the definition of the descendant Gromov–Witten invariants and the relation between the ψ–classes on the moduli of $(n+1)$–marked curves and the moduli of n–marked curves. After expanding the powers, we recognize that the first integral is trivial because the moduli space of n–marked curves is lower dimensional than the moduli space of $(n+1)$–marked curves. Any cycle containing a factor of D_iD_j for $i\neq j$ is empty because the set of all stable maps

with p_i and p_{n+1} isolated in a single bubble is disjoint from the set of all stable maps with just p_j and p_{n+1} in an isolated bubble, so the last integral below is trivial. Using $\mathrm{ev}_k \circ \pi = \mathrm{ev}_k$ on the remaining integral gives

$$\begin{aligned}
\langle \tau_{a_1}(\gamma_1) \dots \tau_{a_n}(\gamma_n) \tau_0(X) \rangle_{g,\beta}^X &= \int_{[\overline{\mathcal{M}}_{g,n+1}(X,\beta)]^{\mathrm{vir}}} \psi_1^{a_1} \dots \psi_n^{a_n} \mathrm{ev}_1^* \gamma_1 \dots \mathrm{ev}_n^* \gamma_n \\
&= \int_{[\overline{\mathcal{M}}_{g,n+1}(X,\beta)]^{\mathrm{vir}}} (\pi^* \psi_1 + D_1)^{a_1} \dots (\pi^* \psi_n + D_n)^{a_n} \mathrm{ev}_1^* \gamma_1 \dots \mathrm{ev}_n^* \gamma_n \\
&= \int_{[\overline{\mathcal{M}}_{g,n+1}(X,\beta)]^{\mathrm{vir}}} \pi^* (\psi_1^{a_1} \dots \psi_n^{a_n} \mathrm{ev}_1^* \gamma_1 \dots \mathrm{ev}_n^* \gamma_n) \\
&\quad + \sum_{k=1}^{n} \sum_{p=1}^{a_k} \binom{a_k}{p} \int_{[\overline{\mathcal{M}}_{g,n+1}(X,\beta)]^{\mathrm{vir}}} D_k^p \pi^* (\psi_1^{a_1} \dots \psi_n^{a_n} \psi_k^{-p} \mathrm{ev}_1^* \gamma_1 \dots \mathrm{ev}_n^* \gamma_n) \\
&\quad + \sum \int_{[\overline{\mathcal{M}}_{g,n+1}(X,\beta)]^{\mathrm{vir}}} (D_i D_j \dots).
\end{aligned}$$

We next change the integrand on the remaining integral to $N(D_k)$ because D_k is supported in this tubular neighborhood. The equation we use in the second step below, $\int_E \Phi\beta = \int_X \sigma^* \beta$, is easy to understand from the view of intersection theory. Any variation of a stable map in the boundary divisor

$$D_k := D(0, \{k, n+1\}, 0 | g, \{1, \dots, \widehat{k}, \dots, n\}, \beta)$$

may be decomposed into variations that move p_k and p_{n+1} out of a common bubble and variations that leave these two points in a common bubble. The normal bundle of D_k consists of those variations that move the points out of a common bubble, that is, the vertical bundle restricted to D_k. It follows that the Euler class of the normal bundle is given by

$$e(N(D_k)) := c_1(\mathcal{L}_k^*) = -\psi_k.$$

Together with a combinatorial identity from the binomial theorem, this allows us to complete the derivation of the string equation.

$$\begin{aligned}
&= \sum_{k=1}^{n} \sum_{p=1}^{a_k} \binom{a_k}{p} \int_{N(D_k)} D_k^p \pi^* (\psi_1^{a_1} \dots \psi_n^{a_n} \psi_k^{-p} \mathrm{ev}_1^* \gamma_1 \dots \mathrm{ev}_n^* \gamma_n) \\
&= \sum_{k=1}^{n} \sum_{p=1}^{a_k} \binom{a_k}{p} \int_{[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\mathrm{vir}}} \rho_k^* \big(D_k^{p-1} \pi^* (\psi_1^{a_1} \dots \psi_n^{a_n} \psi_k^{-p} \mathrm{ev}_1^* \gamma_1 \dots \mathrm{ev}_n^* \gamma_n) \big) \\
&= \sum_{k=1}^{n} \sum_{p=1}^{a_k} \binom{a_k}{p} \int_{[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\mathrm{vir}}} e(N(D_k))^{p-1} \psi_1^{a_1} \dots \psi_n^{a_n} \psi_k^{-p} \mathrm{ev}_1^* \gamma_1 \dots \mathrm{ev}_n^* \gamma_n
\end{aligned}$$

$$= \sum_{k=1}^{n} \sum_{p=1}^{a_k} \binom{a_k}{p} (-1)^{p-1} \int_{[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\mathrm{vir}}} \psi_1^{a_1} \dots \psi_n^{a_n} \psi_k^{-1} \mathrm{ev}_1^* \gamma_1 \dots \mathrm{ev}_n^* \gamma_n$$
$$= \sum_{k=1}^{n} \int_{[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\mathrm{vir}}} \psi_1^{a_1} \dots \psi_n^{a_n} \psi_k^{-1} \mathrm{ev}_1^* \gamma_1 \dots \mathrm{ev}_n^* \gamma_n$$
$$= \sum_{k=1}^{n} \langle \tau_{a_1}(\gamma_1) \dots \tau_{a_k - 1}(\gamma_k) \dots \tau_{a_n}(\gamma_n) \rangle_{g,\beta}^{X}$$

Example 4.9 We will use the string equation to compute a couple of descendant invariants. First consider the integral

$$\int_{[\overline{\mathcal{M}}_{0,6}(\mathrm{pt},0)]^{\mathrm{vir}}} \psi_1^2 \psi_2 \,.$$

This integral is denoted by $\langle \tau_2(\mathrm{pt}) \tau_1(\mathrm{pt}) (\tau_0(\mathrm{pt}))^4 \rangle_{0,0}^{\mathrm{pt}}$. Using the string equation we obtain

$$\begin{aligned} \langle \tau_2(\mathrm{pt}) \tau_1(\mathrm{pt}) (\tau_0(\mathrm{pt}))^4 \rangle_{0,0}^{\mathrm{pt}} &= \langle \tau_1(\mathrm{pt}) \tau_1(\mathrm{pt}) (\tau_0(\mathrm{pt}))^3 \rangle_{0,0}^{\mathrm{pt}} + \langle \tau_2(\mathrm{pt}) (\tau_0(\mathrm{pt}))^4 \rangle_{0,0}^{\mathrm{pt}} \\ &= 3 \langle \tau_1(\mathrm{pt}) (\tau_0(\mathrm{pt}))^3 \rangle_{0,0}^{\mathrm{pt}} = 3 \langle (\tau_0(\mathrm{pt}))^3 \rangle_{0,0}^{\mathrm{pt}} = 3 \,. \end{aligned}$$

More generally consider the integral

$$\int_{[\overline{\mathcal{M}}_{0,n}(\mathrm{pt},0)]^{\mathrm{vir}}} \psi_1^{a_1} \dots \psi_k^{a_k},$$

where $a_1 + \dots + a_k = n - 3$. Using the string equation this can be reduced to a sum of terms with the power of one of the ψ decremented. Repeating this, a total of $n-3$ subtractions without combining like terms will lead to a sum of ones. Each term of this sum will correspond to selecting a_1 of the subtractions, then a_2 of the subtractions, etc. This shows that the value of the integral is $\binom{n-3}{a_1 a_2 \dots a_k}$. The symmetries in this formula arising from rearranging the a_j correspond to the geometric operations on the moduli space obtained by rearranging the marked points.

The derivation of the dilaton equation is similar.

$$\begin{aligned} \langle \tau_{a_1}(\gamma_1) \dots \tau_{a_n}(\gamma_n) \tau_1(X) \rangle_{g,\beta}^{X} &= \int_{[\overline{\mathcal{M}}_{g,n+1}(X,\beta)]^{\mathrm{vir}}} \psi_1^{a_1} \dots \psi_n^{a_n} \psi_{n+1} \mathrm{ev}_1^* \gamma_1 \dots \mathrm{ev}_n^* \gamma_n \\ &= \int_{[\overline{\mathcal{M}}_{g,n+1}(X,\beta)]^{\mathrm{vir}}} (\pi^* \psi_1 + D_1)^{a_1} \dots (\pi^* \psi_n + D_n)^{a_n} \psi_{n+1} \mathrm{ev}_1^* \gamma_1 \dots \mathrm{ev}_n^* \gamma_n \end{aligned}$$

$$= \int_{[\overline{\mathcal{M}}_{g,n+1}(X,\beta)]^{\mathrm{vir}}} \psi_{n+1}\pi^*(\psi_1^{a_1}\dots\psi_n^{a_n}\mathrm{ev}_1^*\gamma_1\dots\mathrm{ev}_n^*\gamma_n)$$
$$+\sum_{k=1}^{n}\sum_{p=1}^{a_k}\binom{a_k}{p}\int_{[\overline{\mathcal{M}}_{g,n+1}(X,\beta)]^{\mathrm{vir}}} \psi_{n+1}D_k^p\pi^*(\psi_1^{a_1}\dots\psi_n^{a_n}\psi_k^{-p}\mathrm{ev}_1^*\gamma_1\dots\mathrm{ev}_n^*\gamma_n)$$
$$+\sum\int_{[\overline{\mathcal{M}}_{g,n+1}(X,\beta)]^{\mathrm{vir}}}(D_iD_j\dots)$$

Two new observations need to be used. The first is that the evaluation of ψ_{n+1} on the fiber of the projection is the evaluation of the first Chern class of the cotangent bundle of a surface (on the surface), which is $2g-2$. The second observation is that the pull-back of ψ_{n+1} under the natural map that interchanges p_k and p_{n+1} is ψ_k. The restriction of this map to the divisor D_k is trivial, so these two bundles agree over this divisor.

$$= (2g-2)\int_{[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\mathrm{vir}}} \pi^*(\psi_1^{a_1}\dots\psi_n^{a_n}\mathrm{ev}_1^*\gamma_1\dots\mathrm{ev}_n^*\gamma_n)$$
$$+\sum_{k=1}^{n}\sum_{p=1}^{a_k}\binom{a_k}{p}\int_{[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\mathrm{vir}}} \rho_k^*\big(\psi_{n+1}D_k^{p-1}\pi^*(\psi_1^{a_1}\dots\psi_n^{a_n}\psi_k^{-p}\mathrm{ev}_1^*\gamma_1\dots\mathrm{ev}_n^*\gamma_n)\big)$$
$$= (2g-2)\int_{[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\mathrm{vir}}} \pi^*(\psi_1^{a_1}\dots\psi_n^{a_n}\mathrm{ev}_1^*\gamma_1\dots\mathrm{ev}_n^*\gamma_n)$$
$$+\sum_{k=1}^{n}\sum_{p=1}^{a_k}\binom{a_k}{p}\int_{[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\mathrm{vir}}} \psi_k e(N(D_k))^{p-1}\psi_1^{a_1}\dots\psi_n^{a_n}\psi_k^{-p}\mathrm{ev}_1^*\gamma_1\dots\mathrm{ev}_n^*\gamma_n$$
$$= (2g-2)\int_{[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\mathrm{vir}}} \pi^*(\psi_1^{a_1}\dots\psi_n^{a_n}\mathrm{ev}_1^*\gamma_1\dots\mathrm{ev}_n^*\gamma_n)$$
$$+\sum_{k=1}^{n}\sum_{p=1}^{a_k}\binom{a_k}{p}(-1)^{p-1}\int_{[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\mathrm{vir}}} \psi_1^{a_1}\dots\psi_n^{a_n}\mathrm{ev}_1^*\gamma_1\dots\mathrm{ev}_n^*\gamma_n$$
$$= (2g-2+n)\langle\tau_{a_1}(\gamma_1)\dots\tau_{a_n}(\gamma_n)\rangle_{g,\beta}^X$$

To conclude this subsection we should remark that there is a different natural family of cohomology classes on these moduli spaces that arise as the Chern classes of the Hodge bundle. This bundle will appear after we discuss the deformation-obstruction sequence in Section 5.6. In addition there is a third important recurrence relation known as the divisor equation (see Hori et al [76]).

5 Local structure of moduli spaces

We begin this section by pointing out what seems to be a paradox in this theory. When the Gromov–Witten invariants of certain integral classes are computed one can get

non-integral answers. A simple example can explain the origin of this paradox. This is how we begin the following subsection.

5.1 Orbifolds and $\overline{M}_{1,1}$

The group of orientation preserving isometries of an octahedron acts on S^2 and on TS^2 in a natural way. The quotient of S^2 by this group action is topologically a new copy of S^2 and the degree of the quotient projection is 24. Note that the projection is not a covering projection, rather it is a branched cover, branched over the points labeled with finite cyclic groups in Figure 5.1. The cyclic group labels are just the stabilizer groups of the points in the preimage of each branch point. By keeping track of the local symmetry groups, one can develop a theory of branched covers that is very close to the theory of covering spaces. This gives rise to the notions of orbifolds, orbibundles and their characteristic classes. Orbibundles are not locally trivial over labeled points, but rather look (locally) like quotients of such bundles by the label groups.

Thus an orbifold is locally a quotient of Euclidean space by a finite group where one keeps track of the local symmetry groups. Note that orbifolds are not manifolds with singularities. Topologically, the points labeled with the finite cyclic groups in Figure 5.1 are locally homeomorphic to $\mathbb{C}$ (although this need not be the case in general), but the stabilizer groups make these points special.

The formal definition is a Deligne–Mumford stack over the category DIFF. This is formalized in Appendix A, but we wanted to give an intuitive description first. We work out further examples in Section 10 below.

Continuing with the quotient of S^2 by the octahedral group, note that by the third part of exercise in Exercise 4.2, we know that the Chern class of a pull-back bundle is the pull-back of the Chern class of the bundle. Since the first Chern class of TS^2 evaluates to 2 on the fundamental class of S^2 we would have to conclude that the first Chern class of the quotient bundle of TS^2 by the orientation preserving octahedral group would have to be $\frac{1}{12}$. Of course there is a map $TS^2/\sim \to S^2/\sim$; however, this map does not give $TS^2/\sim$ the structure of a vector bundle. It gives it the structure of an orbibundle. Chern classes of orbibundles are defined; however, they are not always integral classes.

To see that this situation arises in Gromov–Witten theory consider the structure of the moduli space $\overline{M}_{1,1} := \overline{M}_{1,1}(\text{point}, 0)$. It is a fact that every smooth genus 1 Riemann surface is equivalent to the quotient of $\mathbb{C}$ by a lattice (see Griffiths–Harris [68]). All of these admit a natural group structure, so without loss of generality we may assume that the marked point is the equivalence class of zero. Next, notice that any biholomorphism

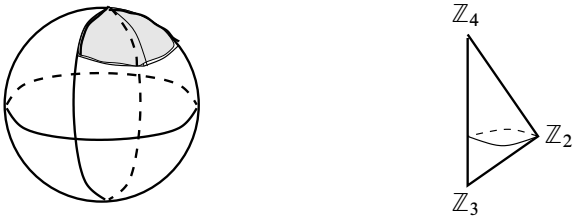


Figure 5.1: The octahedral orbifold

between a pair of tori must be induced from a linear endomorphism of $\mathbb{C}$. To see this, compose a given biholomorphism, $\varphi\colon T \to T'$, with the projection $\mathbb{C} \to T$ and notice that the resulting map lifts to a biholomorphism $\overline{\varphi}\colon \mathbb{C} \to \mathbb{C}$. Now $(\overline{\varphi}(z^{-1}))^{-1}$ has an isolated singularity at zero that must be removable as this function is bounded near zero. It follows that the map $\overline{\varphi}$ extends to a biholomorphism of $\mathbb{C}\mathrm{P}^1$ taking infinity to infinity. Such must be a linear map when restricted to $\mathbb{C}$. The map $z \mapsto -z$ is an automorphism of any torus, thus every point in $\overline{M}_{1,1}$ has stabilizer containing $\mathbb{Z}_2$.

Exercise 5.1 What is the stabilizer of the unique nodal curve in $\overline{M}_{1,1}$?

We may assume that one generator of any lattice used to construct a torus is 1 and that the other generator is in the upper half plane by applying a suitable linear transformation. Equivalently, the second generator is chosen so that the generators of the lattice form a positively oriented basis. It is standard to parameterize a once-marked torus by the

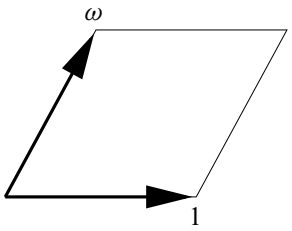


Figure 5.2: Lattice generators

location of the second lattice point. The choice of second lattice point is however not unique. Clearly, adding one to the second generator $z \mapsto z + 1$ does not change the torus. Similarly, interchanging the two generators and changing the sign of one and then renormalizing does not change the torus, $z \mapsto -z^{-1}$. These two operations generate an action of $SL_2\mathbb{Z}$ on the upper half plane. The quotient of the upper half

plane by this action may be identified with $M_{1,1}$; adding a single point at the cusp representing the nodal marked torus gives $\overline{M}_{1,1}$. A fundamental domain for this action and the quotient is displayed in Figure 5.3.

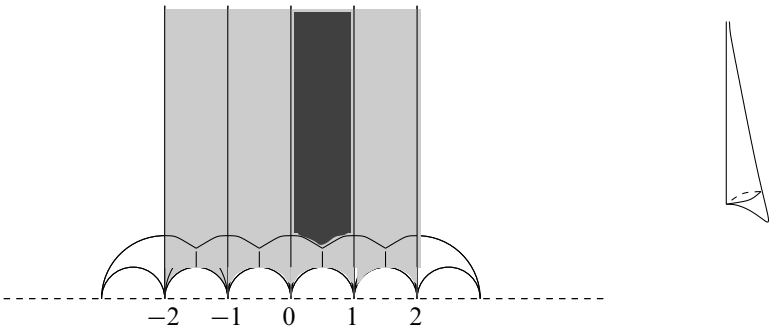


Figure 5.3: The orbifold $M_{1,1}$

We can now analyze the group of automorphisms of one of these marked tori. Consider the torus labeled by $\omega := e^{i\frac{\pi}{3}}$. We have seen that any automorphism must be of the form $[z] \mapsto [\alpha z]$. Since 1 is equivalent to 0 modulo the lattice it must be mapped to a new lattice point, say $\alpha = a + b\omega$. This implies that ω gets mapped to a lattice point, $(a + b\omega)\omega = -b + (a + b)\omega$. This allows us to represent the automorphism as a real linear transformation by a 2×2 matrix. Since the map is invertible and takes the lattice to itself the determinant must be ± 1. Orientation preserving implies that the determinant must be 1, so

$$1 = \det \begin{bmatrix} a & -b \\ b & a+b \end{bmatrix} = a^2 + ab + b^2 = \left(a + \tfrac{1}{2}b\right)^2 + \left(\tfrac{\sqrt{3}}{2}b\right)^2.$$

It follows that (a, b) must be one of the six pairs, $\pm(1, 0)$, $\pm(0, 1)$ or $\pm(1, -1)$. These correspond exactly to the rotations generated by multiplication by powers of ω. Notice that the computation of the stabilizer of ω under the action of $SL_2\mathbb{Z}$ on the upper half plane is exactly the same, thus the upper half plane quotient accurately models the orbifold structure on $M_{1,1}$. In particular notice that $SL_2\mathbb{Z}$ does not act effectively on the upper half plane as every point is fixed by $-I$. This corresponds to the fact that a generic elliptic curve has $\mathbb{Z}_2$ automorphism group. Thus, $PSL_2\mathbb{Z}$ is definitely not the right group to study from the point of view of algebraic geometry.

Exercise 5.2 Compute the automorphism group of the marked torus labeled by i.

We will now take a detour via exercises to prove that $SL_2\mathbb{Z}$ is generated by two matrices and sketch a proof that $\langle\tau_1(\mathrm{pt})\rangle^{\mathrm{pt}}_{1,0} = \frac{1}{24}$. The matrices are

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

After the detour, we will give the correct definition of the moduli stack. This encodes the orbifold structure of the moduli space.

Let $PSL_2\mathbb{Z}$ be the quotient of $SL_2\mathbb{Z}$ by ± 1. This acts on the upper half plane by linear fractional transformations. One can define an orbifold fundamental group. This group should satisfy a van Kampen theorem and the orbifold fundamental group of an X/Γ should be Γ when X is a simply connected orbifold and the action is nice (see Ratcliffe [126]).

Exercise 5.3 Conclude that $\pi_1^{orb}(\mathbb{H}/PSL_2\mathbb{Z}) = PSL_2\mathbb{Z}$. Now use the orbifold van Kampen theorem and Figure 5.3 to prove that this is $\langle [S], [W] | [S]^2 = 1, [W]^3 = 1\rangle$, where $W = ST$ and $[A]$ is the image of $A \in SL_2\mathbb{Z}$ under the natural projection.

Exercise 5.4 Check that the map from the group $\langle s, t | s^2 = (st)^3, s^4 = 1\rangle$ to $SL_2\mathbb{Z}$ taking s to S and t to T is a well-defined group homomorphism. Show that the kernel of the composition of this map with projection to $PSL_2\mathbb{Z}$ is $\mathbb{Z}_2$. Conclude that $SL_2\mathbb{Z} \cong \langle s, t | s^2 = (st)^3, s^4 = 1\rangle$.

Exercise 5.5 Give a definition of a line bundle in the orbifold category. Also define the first Chern class of an orbifold bundle. You should see that the Chern class may be computed by counting zeros with sign, but one must divide each term by the order of the stabilizer group. Also prove that this Chern class satisfies the usual multiplication by degree for pull-backs.

Now $\langle\tau_1(\mathrm{pt})\rangle^{\mathrm{pt}}_{1,0}$ is just the evaluation of the first Chern class of the tautological bundle over $\overline{M}_{1,1}$. To do this we would like to have a section of $\mathcal{L}_1$. Such a section would associate a holomorphic form to each marked torus. It is natural to try a section that would associate $f(\tau)dz$ to $\mathbb{C}/\langle 1, \tau\rangle$. Transformations by elements of $SL_2\mathbb{Z}$ fix the torus and so should fix the holomorphic form. Multiplication by $c\tau + d$ is an isomorphism $\mathbb{C}/\langle 1, \frac{a\tau+b}{c\tau+d}\rangle \to \mathbb{C}/\langle c\tau + d, a\tau + b\rangle = \mathbb{C}/\langle 1, \tau\rangle$. It follows that we need $f\left(\frac{a\tau+b}{c\tau+d}\right) = (c\tau + d)\, f(\tau)$ in order for this section to be well defined.

Definition 5.6 The Dedekind eta function is the analytic function on the upper half plane given by

$$\eta(\tau) := e^{\frac{\pi i \tau}{12}} \prod_{n=1}^{\infty} (1 - e^{2\pi i n \tau}).$$

The following functional equation is proved in Apostol [12].

$$\eta^2\left(\frac{a\tau+b}{c\tau+d}\right) = -i\epsilon(a,b,c,d)^2(c\tau+d)\eta^2(\tau)\,,$$

where

$$\epsilon(a,b,c,d) := \exp\left\{\pi i\left(\frac{a+d}{12c} + s(-d,c)\right)\right\}\,,$$

and

$$s(h,k) := \sum_{r=1}^{k-1}\frac{r}{k}\left(\frac{hr}{k} - \left[\frac{hr}{k}\right] - \frac{1}{2}\right)\,.$$

Thus $\tau \mapsto \eta^2(\tau)dz$ is almost a holomorphic section of the tautological bundle $\mathcal{L}_1$. In fact, $(-i\epsilon(a,b,c,d)^2)^{12} = 1$ so $\tau \mapsto \eta^2(\tau)dz$ is a well-defined section on the pull-back of $\mathcal{L}_1$ to a 12–fold branched cover of $\overline{M}_{1,1}$.

Exercise 5.7 Label translates of the fundamental domain in Figure 5.3 by the group elements in $PSL_2\mathbb{Z}$ used to translate them. Now compute $-i\epsilon([A])^2$ for each group element and use this as a label. Pick one translate with each of the 12 different labels. The collection of these translates is a fundamental domain for the 12–fold branched cover of $M_{1,1}$. Also label the adjoining translates to figure out the identifications on the larger fundamental domain.

Exercise 5.8 Show that tubular neighborhoods of the cusps in each of these 12 labeled translates glue together to a once-punctured disk, and the function $\eta^2(\tau)$ extends across this disk with a simple zero at the center. (Use the group label to translate back to the small fundamental domain without losing the $-i\epsilon([A])^2$ factor and introduce $w = e^{2\pi i\tau}$ as a coordinate.)

Adding the extra point described in Exercise 5.8 gives the 12 fold cover together with a section of the pull-back of $\mathcal{L}_1$ to this cover. The negative of the identity matrix acts trivially on every point of this cover. This means that the stabilizer group of every point in the cover is $\mathbb{Z}_2$ so the evaluation of the first Chern class of the pull-back bundle is $\frac{1}{2}$ of the number of zeros which is just $\frac{1}{2}$. Thus since Chern classes are multiplicative under covers $\langle\tau_1(\mathrm{pt})\rangle^{\mathrm{pt}}_{1,0} = \frac{1}{24}$ as claimed concluding our detour.

5.2 Moduli stacks

The proper structure to encode the orbifold idea in algebraic geometry is a stack. One does not have to understand the definition of a stack to get a feel for Gromov–Witten invariants so we do not include the definition here. However most sources do not even

define a stack, so we have given the definition together with some motivating examples in Appendix A. The main points to keep in mind are that a stack adds extra structure that remembers the stabilizer groups, and that the 'universal' family is a stack.

A family of stable curves $\{\Sigma_y\}_{y\in Y}$ is encoded as a map say f from one space X to a second space Y such that for every point $y \in Y$, $f^{-1}(y)$ can be thought of as a stable curve. One thinks of this as the set of stable maps $f^{-1}(y)$ together with some topology linking everything together.

To define families of stable maps, we need a technical definition.

Definition 5.9 An R–module M is *flat* if $0 \to M \otimes_R A \to M \otimes_R B \to M \otimes_R C \to 0$ is exact whenever $0 \to A \to B \to C \to 0$ is exact. Given a morphism between analytic spaces, $f\colon X \to Y$ the ring of germs of analytic functions over a point $x_0 \in X$ (denoted by $\mathcal{O}_X(x_0)$) is a module over the corresponding ring of germs on Y, $\mathcal{O}_Y(f(x_0))$. The map f is called *flat* if $\mathcal{O}_X(x_0)$ is a flat $\mathcal{O}_Y(f(x_0))$–module for every $x_0 \in X$.

Intuitively a map f is flat if it has a nice fiber structure. See Hartshorne [75] for a discussion and examples.

Definition 5.10 A family of stable maps is a flat morphism $\pi\colon \mathcal{V} \to S$ together with maps, $u\colon \mathcal{V} \to X$, and $\rho_k\colon S \to \mathcal{V}$ such that $f \circ \rho_k = \mathrm{id}_S$ and

$$[u|_{\pi^{-1}(s_0)}, \pi^{-1}(s_0), \rho_1(s_0), \ldots, \rho_n(s_0)] \in \overline{M}_{g,n}(X, \beta)$$

for every $s_0 \in S$. A morphism of families of stable maps is a pair of maps, $G\colon \mathcal{V} \to \mathcal{V}'$ and $g\colon S \to S'$ that intertwine the structure maps, u, u', ρ_k and ρ'_k.

A universal family of stable curves is one such that there is a unique morphism of families from any given family into the universal one. If one sticks to families over schemes or analytic spaces, there is no universal family; see the discussion in Mumford [113], Harris–Morrison [74] or Section 10. The moduli stack is just a formal construction of a universal family. The resulting construction generalizes the category of schemes or analytic spaces. Given a space T one constructs the contravariant functor $\underline{T}$ that takes a scheme S to the set of all morphisms from S to T. See Appendix A or Section 10 for more information.

Definition 5.11 The moduli stack, $\overline{\mathcal{M}}_{g,n}(X, \beta)$ is the functor from the category of schemes (or analytic spaces) to the category of sets taking a scheme S to the set of equivalence classes of families of genus g, n–marked stable curves representing β in $H_2(X)$. (Recall the definition of a family of stable maps from Definition 5.10).

The definition of the moduli stack is very elegant and nicely avoids questions about the topology of moduli space. However, one still has to work to extend intersection theory to stacks; see Vistoli [155].

5.3 Deformation complexes

Finally, we come to the main point of this section – the local structure of the moduli space. It is a general principle that nondegenerate spaces or maps have the same local structure as their linear approximations. The implicit function theorem is one version of this principle. We are interested in a generalization to spaces with group actions; see Bredon [33], Audin [17] and Atiyah–Bott [15]. It will be convenient to consider right G–spaces, ie assume that the group G acts on the right in contrast to the standard convention.

As a warm-up we shall study the following situation. Let $F\colon X \to Y$ be a smooth equivariant map of right G–spaces and let $y_0 := F(x_0)$ be a G fixed point of Y. To get a local model of $F^{-1}(y_0)/G$ in a neighborhood of $[x_0]$ we linearize the following sequence of maps,

$$G \xrightarrow{L_{x_0}} X \xrightarrow{F} Y\,,$$

where $L_{x_0}(g) := x_0 g$. The linearization is

$$T_1 G \xrightarrow{TL_{x_0}} T_{x_0} X \xrightarrow{TF} T_{y_0} Y\,.$$

Exercise 5.12 Prove that the above sequence is a complex, that is, $TF \circ TL_{x_0} = 0$ given that y_0 is a fixed point. It is called the deformation complex.

We will call the stabilizer group of x_0 the automorphism group of x_0, $\mathrm{Aut}(x_0) := \{g \in G | x_0 g = x_0\}$. The kernel of TL_{x_0} is the zeroth cohomology of the deformation complex and it is isomorphic to the Lie algebra of the automorphism group, $\mathfrak{aut}(x_0)$. Provided TF is surjective, $F^{-1}(y_0)$ will be a manifold. The cokernel of TF measures the failure of TF to be surjective. This cokernel is the second cohomology of the deformation complex and is called the obstruction space, $\mathfrak{ob}(x_0)$. Assuming that the obstruction space vanishes, $F^{-1}(y_0)$ is a manifold locally homeomorphic to the kernel of TF. The quotient of this manifold by the G action can be locally identified with the first cohomology of the deformation complex provided that the automorphism group is trivial. This cohomology group is called the deformation space of x_0 and is denoted by $\mathfrak{def}(x_0)$. The point is that the exponential map applied to the orthogonal complement of the image of TL_{x_0} in the kernel of TF is a local slice for the action of G on $F^{-1}(y_0)$. In other words, it intersects each nearby G orbit in exactly one

point. The obstruction, deformation and automorphism spaces glue together to form the obstruction, deformation and automorphism bundles over the space $F^{-1}(y_0)/G$. If the obstruction space is trivial at a point, but the automorphism group is nontrivial, then there is a natural action of the automorphism group on the deformation space. Furthermore, $F^{-1}(y_0)/G$ is locally homeomorphic to the quotient of the deformation space by the action of the automorphism group. If the obstruction space is nontrivial, there is a map from the deformation space to the obstruction space and the quotient $F^{-1}(y_0)/G$ is locally homeomorphic to the quotient of the inverse image of zero under this map by the automorphism group. This special map is called a Kuranishi map. The Kuranishi map is described in the following exercise for the case of trivial automorphism group.

Exercise 5.13 Let $F\colon H^1 \oplus V \to H^2 \oplus W$ be a smooth (non-linear) map between linear spaces satisfying $F(0) = 0$, $H^1 = \ker(T_0 F)$, and $H^2 = \operatorname{coker}(T_0 F)$. Define a map $\Phi\colon H^1 \oplus H^2 \oplus V \to H^1 \oplus H^2 \oplus W$ given by

$$\Phi(x, y, z) = (x, y + F_1(x, z), F_2(x, z)) .$$

Use the inverse function theorem to prove that there is a locally defined inverse and smooth maps $\psi\colon H^1 \to H^2$ and $\xi\colon H^1 \to V$ defined on open neighborhoods of zero in H^1 such that $\Phi(x, \psi(x), \xi(x)) = 0$. Conclude that $\psi^{-1}(0)$ is locally homeomorphic to $F^{-1}(0)$. The map ψ is the Kuranishi map.

Exercise 5.14 Apply these ideas to various orbit types in the quotient of S^3 (viewed as the unit sphere in $\mathbb{C}^2$) by the natural action of $S^1 \times S^1$.

5.4 Deformations of stable maps

We will now apply these ideas to the moduli space of stable maps. A stable map is specified by a complex structure on a surface, a collection of marked points and a J–holomorphic map into a symplectic manifold. Recall the relevant definitions from Section 3. We will separate the deformations of a stable curve into deformations of the marked points and complex structure and deformations of the map with a fixed set of marked points and fixed complex structure.

Begin by considering deformations of the map with the marked points and complex structure fixed. By definition, a map $u\colon \Sigma \to X$ is J–holomorphic if $\overline{\partial} u = 0$. The expression $\overline{\partial} u$ may be applied to a tangent vector in $T_{x_0}\Sigma$ to produce a tangent vector in $T_{u(x_0)}X$. This may be reinterpreted to say that $\overline{\partial} u \in C^\infty(\wedge^1 \Sigma \otimes u^* TX)$, ie $\overline{\partial} u$ is a 1–form with values in the pullback of the tangent bundle. The expression $\overline{\partial} u$ extends to a map from the complexified tangent space of Σ to the complexified tangent space

of X. It is completely determined by an induced map from $\wedge^{0,1}\Sigma$ to u^*TX. Thus we usually view $\overline{\partial}u \in C^\infty(\wedge^{0,1}\Sigma \otimes u^*TX)$.

Exercise 5.15 Let V be a real vector space with almost complex structure J, and let $V^{\mathbb{C}} := V \otimes_{\mathbb{R}} \mathbb{C}$ be the complexification. The (anti)holomorphic subspace is the $(-)i$–eigenspace of $J^{\mathbb{C}} := J \otimes 1$ acting on $V^{\mathbb{C}}$. These are denoted by $(V^{0,1})$ or $V^{1,0}$.

(1) Show that the projection $P^{1,0}\colon V^{\mathbb{C}} \to V^{1,0}$ given by

$$P(X) = \tfrac{1}{2}(X - J \otimes iX)$$

restricts to an isomorphism of complex vector spaces from V to $V^{1,0}$ when V is viewed as a complex vector space via J and is viewed as a subspace of $V^{\mathbb{C}}$ via $V \otimes 1$. (For this reason we often identify the holomorphic subspace of $V^{\mathbb{C}}$ with V by $P^{1,0}$.)

(2) Let $\overline{\partial}u$ act on the complexified tangent space of Σ in the natural way. Show that it takes $\wedge^{1,0}\Sigma := T\Sigma^{1,0}$ to $u^*TX^{0,1}$ (it takes holomorphic vectors to antiholomorphic vectors.)

(3) Conclude that $\overline{\partial}u$ is uniquely determined by the restriction of $P^{1,0}\overline{\partial}u$ to $\wedge^{0,1}\Sigma$.

Now consider a one-parameter family of maps passing through u, say v_t. The derivative of this family (at $t = 0$) associates a tangent vector in $T_{u(x_0)}X$ to a point $x_0 \in \Sigma$, so we may view $\dot{v}_0 \in C^\infty(u^*TX)$. The linearization of $\overline{\partial}$ maps $\dot{v}_0$ to the derivative of $\overline{\partial}v_t$ at $t = 0$. We will abuse notation and denote this by $\overline{\partial}\dot{v}_0$. This may be viewed as an element of $C^\infty(\wedge^{0,1}\Sigma \otimes u^*TX)$ as explained in the previous exercise. An expression for the linearization of $\overline{\partial}$ may be found in Salamon [133], McDuff–Salamon [108] and Audin–Lafontaine [18]. We represent it by

$$\overline{\partial}\colon C^\infty(u^*TX) \to C^\infty(\wedge^{0,1}\Sigma \otimes u^*TX). \tag{2}$$

The kernel of this map is the deformation space of maps with fixed complex structure and marked points. It is denoted by $\mathfrak{def}(u)$. The space of stable maps with underlying marked curve $[\Sigma, p]$ is locally diffeomorphic to $\mathfrak{def}(u)$ at $[u, \Sigma, p]$ provided that the obstruction space (cokernel of the above map) $\mathfrak{ob}(u)$ vanishes. There are no automorphisms that act on maps with fixed marked domain.

We now consider automorphisms and deformations of the underlying marked curve. An automorphism is just a holomorphic map $\varphi\colon \Sigma \to \Sigma$ fixing the marked points, so $\overline{\partial}\varphi = 0$ and $\varphi(p_k) = p_k$. Given a one-parameter family of maps φ_t, we can differentiate the defining conditions of the automorphism group to obtain the conditions specifying infinitesimal automorphisms: $\overline{\partial}\dot{\varphi}_0 = 0$ and $\dot{\varphi}_0|_{p_k} = 0$. The nice way to

encode these holomorphic vector fields that vanish at the marked points is to use the kernel of the following operator,

$$\overline{\partial}\colon C^\infty(T\Sigma\otimes[-p]) \to C^\infty(\wedge^{0,1}\Sigma\otimes T\Sigma\otimes[-p])\,. \tag{3}$$

Here $-p=-p_1-\cdots-p_n$ is negative the divisor associated to the marked points and $[-p]$ is the complex line bundle associated with this divisor; see Griffiths–Harris [68]. One may pick a (unique up to scale) non-zero meromorphic section of this line bundle with a simple pole at each marked point, say s^-. The map $\dot\varphi_0\mapsto\dot\varphi_0\otimes s^-$ identifies the infinitesimal automorphisms $\mathfrak{aut}([\Sigma,p])$ with the kernel of the above map.

Since a complex structure satisfies $J^2=-I$, the derivative of a one-parameter family of complex structures satisfies $\dot J_0 J + J\dot J_0 = 0$. For a holomorphic vector X we have

$$J(J\dot J_0 X) = -J(\dot J_0 J X) = -iJ\dot J_0 X\,.$$

Thus $\frac12 J\dot J_0$ takes holomorphic vectors to antiholomorphic vectors and may be interpreted as an element of $\wedge^{0,1}\Sigma\otimes T\Sigma$ similar to the way that $\overline{\partial}\dot v_0$ may be interpreted as an element of $\wedge^{0,1}\Sigma\otimes u^*TX$. A one-parameter family of deformations J_t is trivial if there is a one-parameter family of reparametrizations φ_t so that $d\varphi_t\circ J = J_t\circ d\varphi_t$.

Exercise 5.16 Show that that $d\varphi_t\circ J = J_t\circ d\varphi_t$ implies that $\frac12 J\dot J_0 = \overline{\partial}\dot\varphi_0$.

One also has to consider deformations of the marked points. Surprisingly, deformations of the marked points may be modeled by deformations of the complex structure that are fixed at the marked points. This is easiest to understand on the Riemann sphere $\mathbb{C}\mathrm{P}^1$ because $\mathbb{C}\mathrm{P}^1$ has exactly one complex structure up to reparametrizations.

Consider a collection of four or more points on $\mathbb{C}\mathrm{P}^1$, say $p=(p_1,\dots,p_n)$. Since we know that there is a unique complex structure on $\mathbb{C}\mathrm{P}^1$ up to reparametrizations on $\mathbb{C}\mathrm{P}^1$, given any family of complex structures J_t we may find a family of reparametrizations φ_t so that $d\varphi_t\circ J = J_t\circ d\varphi_t$. The expression $\varphi_t(p)$ describes the associated one-parameter family of marked points. This generalizes to deformations of the complex structure and marked points on any marked curve. To summarize, each deformation of the equivalence class of a marked curve $[\Sigma,p]$ is uniquely specified by an element $(\frac12 J\dot J_0)$ of the cokernel of the map in equation (3). The obstruction space vanishes for dimensional reasons.

One may think that it is impossible to have nontrivial automorphisms and nontrivial deformations of the same marked Riemann surface. This is true for smooth surfaces, but it is not true for nodal curves. The marked curve in Figure 5.4 has a four-complex-dimensional family of automorphisms corresponding to linear reparametrizations of

each side bubble. In addition it has a seven-complex-dimensional family of deformations, three for the marked points in the center bubble, two more for the locations of the nodes and two more for the resolutions of the side bubbles.

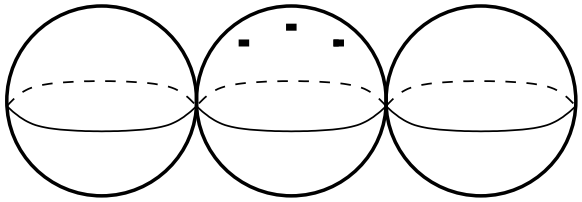


Figure 5.4: A nodal curve with automorphisms and deformations

We can now assemble the deformation complexes of a map (equation (2)) and marked curve (equation (3)) to obtain the deformation complex of a stable map $[u, \Sigma, p]$. We first form the double complex

$$\begin{array}{ccc} C^\infty(u^*TX) & \xrightarrow{\ \overline{\partial}\ } & C^\infty(\wedge^{0,1}\Sigma \otimes u^*TX) \\ {\scriptstyle du(-\otimes s^+)}\uparrow & & \uparrow{\scriptstyle du(-\otimes s^+)} \\ C^\infty(T\Sigma \otimes [-p]) & \xrightarrow[\ \overline{\partial}\ ]{} & C^\infty(\wedge^{0,1}\Sigma \otimes T\Sigma \otimes [-p]) \end{array} \tag{4}$$

Here s_+ is a non-zero section of $[p]$ with a simple zero at each marked point. A bit of thought allows one to conclude that the associated total complex,

$$C^\infty(T\Sigma\otimes[-p]) \to C^\infty(\wedge^{0,1}\Sigma\otimes T\Sigma\otimes[-p]) \oplus C^\infty(u^*TX) \to C^\infty(\wedge^{0,1}\Sigma\otimes u^*TX)\,,$$

is the deformation complex of the stable map $[u, \Sigma, p]$. The kernel of the first map $((\overline{\partial}, du(-\otimes s^+)))$ encodes the conditions that an infinitesimal automorphism corresponds to a one-parameter family of holomorphic maps φ_t fixing the marked points and satisfying $u \circ \varphi_0 = u$. The kernel of the second map $(\overline{\partial} - du(-\otimes s^+))$ encodes the fact that each map in a deformation of a stable map must be J–holomorphic with respect to the corresponding structure ($J \circ du_t = du_t \circ J_t$).

5.5 Homological description of deformations

Introducing a bit of homological algebra will allow us to formulate the deformation complex for nodal curves, where the notion of C^∞ sections may be less clear, and will provide additional computational tools for the study of the local structure of the moduli space. The homological definition of the deformation complex will also have

the added benefit of defining bundles of infinitesimal automorphisms, deformations and obstructions over the moduli space as opposed to just defining vector spaces attached at each point. The book by Weibel [156] is a good reference for the homological algebra that we use.

Definition 5.17 An R–module I is called injective if the functor $\mathrm{Hom}_R(-, I)$ is exact (that is, takes exact sequences to exact sequences). An *injective resolution* of an R–module M is an exact sequence,

$$0 \to M \to I_0 \to I_1 \to \cdots,$$

where the I_k are injective. Given a complex of R–modules,

$$A_* := \cdots \to A_k \to A_{k+1} \to \cdots$$

and an R–module B, one may take an injective resolution of B say $0 \to B \to I_*$, and form the double complex $\mathrm{Hom}_R(A_i, I_j)$. The *hyperext* group $\mathbb{E}xt^k_R(A_*, B)$ is by definition the kth cohomology of the associated double complex.

Notice that we have isomorphisms

$$\begin{aligned}
C^\infty(u^*TX) &\cong \mathrm{Hom}_{\mathcal{O}_\Sigma}(u^*\Omega_X, C^\infty(\wedge^0\Sigma)) \\
C^\infty(\wedge^{0,1}\Sigma \otimes u^*TX) &\cong \mathrm{Hom}_{\mathcal{O}_\Sigma}(u^*\Omega_X, C^\infty(\wedge^{0,1}\Sigma)) \\
C^\infty(T\Sigma \otimes [-p]) &\cong \mathrm{Hom}_{\mathcal{O}_\Sigma}(\Omega_\Sigma([p]), C^\infty(\wedge^0\Sigma)) \\
C^\infty(\wedge^{0,1}\Sigma \otimes T\Sigma \otimes [-p]) &\cong \mathrm{Hom}_{\mathcal{O}_\Sigma}(\Omega_\Sigma([p]), C^\infty(\wedge^{0,1}\Sigma)).
\end{aligned}$$

Here $\mathcal{O}_\Sigma$ is the sheaf of holomorphic functions on Σ, $\Omega_\Sigma(L)$ is the sheaf of holomorphic differentials on Σ taking values in a line bundle L and Ω_X is the sheaf of holomorphic differentials on X. Any sheaf $\mathcal{E}$ may be realized as the sections of an associated sheaf space $E \to X$. By definition, the pull-back of a sheaf under a map $f\colon Y \to X$ is $f^*\mathcal{E} := \Gamma(f^*E)$.

Exercise 5.18 The third isomorphism listed above is defined by

$$F(X \otimes s)(\alpha \otimes t) := \alpha(X)st\,.$$

Find the other three isomorphisms.

We can now see how to represent the infinitesimal automorphisms, deformations and obstructions as hyperext groups. Notice that a partition of unity argument may be used to show that a sheaf of C^∞ sections is fine. This in turn implies that it is $\mathrm{Hom}_{\mathcal{O}_\Sigma}(du, -)$–acyclic thus the hyperext groups can be computed with these sheaves;

see Hartshorne [75]. Also recall that the definition of an exact sequence of sheaves is defined by the requirement that the sequence be exact at the level of germs. This means that the usual de Rham complex is an exact sequence of sheaves (even though the sequence of global sections fails to be exact as measured by the de Rham cohomology). The point is that a closed form becomes exact when it is restricted to a small enough open set. It follows that the following is an acyclic resolution of $\mathcal{O}_\Sigma$:

$$0 \to \mathcal{O}_\Sigma \to C^\infty(\wedge^0\Sigma) \to C^\infty(\wedge^{0,1}\Sigma) \to 0\,.$$

We will use this resolution and the definition of the hyperext groups to construct the deformation-obstruction complex.

5.6 The deformation-obstruction sequence

Using the four isomorphisms and the definition of the hyperext groups we arrive at the following definitions which generalize our proceeding discussion for smooth stable maps.

Definition 5.19 The spaces in the deformation-obstruction sequence of a stable map are given by:

$$\begin{aligned}
\mathfrak{def}(u) &:= \mathbb{E}\mathrm{xt}^0_{\mathcal{O}_\Sigma}(u^*\Omega_X, \mathcal{O}_\Sigma)\\
\mathfrak{ob}(u) &:= \mathbb{E}\mathrm{xt}^1_{\mathcal{O}_\Sigma}(u^*\Omega_X, \mathcal{O}_\Sigma)\\
\mathfrak{aut}([\Sigma, p]) &:= \mathbb{E}\mathrm{xt}^0_{\mathcal{O}_\Sigma}(\Omega_\Sigma([p]), \mathcal{O}_\Sigma)\\
\mathfrak{def}([\Sigma, p]) &:= \mathbb{E}\mathrm{xt}^1_{\mathcal{O}_\Sigma}(\Omega_\Sigma([p]), \mathcal{O}_\Sigma)\\
\mathfrak{aut}([u, \Sigma, p]) &:= \mathbb{E}\mathrm{xt}^0_{\mathcal{O}_\Sigma}(u^*\Omega_X \to \Omega_\Sigma([p]), \mathcal{O}_\Sigma)\\
\mathfrak{def}([u, \Sigma, p]) &:= \mathbb{E}\mathrm{xt}^1_{\mathcal{O}_\Sigma}(u^*\Omega_X \to \Omega_\Sigma([p]), \mathcal{O}_\Sigma)\\
\mathfrak{ob}([u, \Sigma, p]) &:= \mathbb{E}xt^2_{\mathcal{O}_\Sigma}(u^*\Omega_X \to \Omega_\Sigma([p]), \mathcal{O}_\Sigma)\,.
\end{aligned}$$

Homological algebra will lead to a long exact sequence relating these groups. The following aside sketches the derivation of the exact sequence.

Aside 5.20 Since the hyperext groups are defined as the cohomology of the total complex of a double complex and every double complex has an associated spectral sequence, there is a spectral sequence related to the hyperext groups. As the double complex only has two terms in any direction, the spectral sequence will collapse at the E_2 term and produce a long exact sequence. See Bott and Tu [30], Brown [34] and Spanier [147] for a detailed discussion of spectral sequences. Briefly, to any double complex $C^{p,q}$ one associates a total complex

$$TC^n := \oplus_{k+\ell=n} C^{k,\ell}$$

filtered by

$$F^p TC^n := \oplus_{k+\ell=n, k\geq p} C^{k,\ell}.$$

The zeroth page of the spectral sequence is defined by

$$E_0^{p,q} := \frac{F^p TC^{p+q}}{F^{p+1} TC^{p+q}}.$$

Similarly the last page of the spectral sequence is defined by

$$E_\infty^{p,q} := \frac{F^p H^{p+q}(TC^*)}{F^{p+1} H^{p+q}(TC^*)}.$$

Exercise 5.21 Prove that for the spectral sequence of a double complex one has $E_0^{p,q} = C^{p,q}$.

The differentials of the double complex may be used to define differentials on the pages of the spectral sequence, $d_n\colon E_n^{p,q} \to E_n^{p+n,q+1-n}$. Here the pages may be inductively defined by $E_{n+1} := H(E_n, d_n)$. Applying this to the double complex of equation (4) we see that the E_1–page of the associated spectral sequence is given by taking the cohomology in the vertical direction. Combined with the definition of the deformation spaces this gives:

$$E_1 = \begin{matrix} \mathfrak{def}([\Sigma, p]) & \mathfrak{ob}(u) \\ \mathfrak{aut}([\Sigma, p]) & \mathfrak{def}(u). \end{matrix}$$

Taking cohomology in the horizontal direction gives:

$$E_2 = \begin{matrix} \ker(\mathfrak{def}([\Sigma, p]) \to \mathfrak{ob}(u)) & \operatorname{coker}(\mathfrak{def}([\Sigma, p]) \to \mathfrak{ob}(u)) \\ \ker(\mathfrak{aut}([\Sigma, p]) \to \mathfrak{def}(u)) & \operatorname{coker}(\mathfrak{aut}([\Sigma, p]) \to \mathfrak{def}(u)). \end{matrix}$$

Since all groups outside of this square are zero all other differentials are zero and we conclude that $E_\infty = E_2$. Using the definition of the E_∞ page and the definition of the deformation groups we conclude that

$$E_\infty^{0,1} = \frac{F^0 H^1(TC^*)}{F^1 H^0(TC^*)}, \quad E_\infty^{1,0} = F^0 H^1(TC^*) \quad \text{and} \quad F^0 H^1(TC^*) = \mathfrak{def}([u, \Sigma, p]).$$

Combining this with the above computation of the E_∞ page gives the following exact sequence:

$$0 \to \operatorname{coker}(\mathfrak{aut}([\Sigma, p]) \to \mathfrak{def}(u)) \to \ker(\mathfrak{def}([\Sigma, p]) \to \mathfrak{ob}(u)) \to 0.$$

Exercise 5.22 Continue in this way to prove that the following deformation-obstruction sequence is exact.

The deformation-obstruction sequence is

$$\begin{aligned}(5)\quad 0 \to \mathfrak{aut}([u,\Sigma,p]) &\to \mathfrak{aut}([\Sigma,p]) \\ &\to \mathfrak{def}(u) \to \mathfrak{def}([u,\Sigma,p]) \to \mathfrak{def}([\Sigma,p]) \\ &\to \mathfrak{ob}(u) \to \mathfrak{ob}([u,\Sigma,p]) \to 0\,.\end{aligned}$$

It is worth pointing out what the maps in the deformation-obstruction sequence are. Any automorphism of $[u,\Sigma,p]$ is an automorphism of $[\Sigma,p]$, so the first map is just the inclusion. The second map is given by $\dot\varphi_0 \mapsto \frac{d}{dt}u\circ\varphi|_{t=0}$. The third map is $\dot v \mapsto (\dot v, 0)$; the fourth is $(\dot v, B)\mapsto B$; the fifth is $B\mapsto du(B\otimes s^+)$. The sixth map is just the projection because both $\mathfrak{ob}(u)$ and $\mathfrak{ob}([u,\Sigma,p])$ are quotients of the same group, but the latter is a quotient by a larger equivalence. A few more comments will clarify these notions.

First consider the infinitesimal automorphisms and deformations of a smooth marked curve. We have the following acyclic resolution of $\mathcal{O}_\Sigma(T\Sigma\otimes[-p])$:

$$0\to\mathcal{O}_\Sigma(T\Sigma\otimes[-p])\to C^\infty(T\Sigma\otimes[-p])\to C^\infty(\wedge^{0,1}\Sigma\otimes T\Sigma\otimes[-p])\to 0\,.$$

It follows that we may make the identifications

$$\begin{aligned}\mathfrak{aut}([\Sigma,p]) &\cong H^0(\Sigma,\mathcal{O}_\Sigma(T\Sigma\otimes[-p])) \\ \mathfrak{def}([\Sigma,p]) &\cong H^1(\Sigma,\mathcal{O}_\Sigma(T\Sigma\otimes[-p]))\,.\end{aligned}$$

The simplest version of the Riemann–Roch theorem (see Forster [58]) states that

$$\dim_{\mathbb{C}} H^0(\Sigma;L)-\dim_{\mathbb{C}} H^1(\Sigma;L)=c_1(L)[\Sigma]+1-g\,.$$

This gives

$$\begin{aligned}&\dim_{\mathbb{C}}\mathfrak{aut}([\Sigma,p])-\dim_{\mathbb{C}}\mathfrak{def}([\Sigma,p]) \\ &\quad=\dim_{\mathbb{C}} H^0(\Sigma;\mathcal{O}_\Sigma(T\Sigma\otimes[-p]))-\dim_{\mathbb{C}} H^1(\Sigma;\mathcal{O}_\Sigma(T\Sigma\otimes[-p])) \\ &\quad=\deg(T\Sigma\otimes[-p])+1-g=3-3g-n.\end{aligned}$$

This is in fact true for the deformations and infinitesimal automorphisms of any marked curve. For the special case of $\Sigma=\mathbb{C}\mathrm{P}^1$ the bundle $T\mathbb{C}\mathrm{P}^1\otimes[-p]$ is just the degree $2-n$ bundle over the sphere. The *Kodaira vanishing theorem* (see Griffiths–Harris [68]) states that $H^q(X;\Omega^p(L))=0$ for $p+q>n$ when L has positive degree. With 3 or fewer marked points the Kodaira vanishing theorem implies that $H^1(\Sigma,\mathcal{O}_\Sigma(T\Sigma\otimes[-p]))=0$ and we conclude that $\dim_{\mathbb{C}}\mathfrak{aut}([\Sigma,p])=3-n$ and $\dim_{\mathbb{C}}\mathfrak{def}([\Sigma,p])=0$. One may also directly compute H^0 as homogenous degree $2-n$ polynomials in this case. One

sees that this matches perfectly with linear fractional transformations fixing 3 or fewer points.

With three or fewer marked points H^1 is trivial and H^0 can be nontrivial. We will see that the situation with 4 or more marked points is just the opposite– H^0 is trivial and H^1 can be nontrivial. To be more precise, recall that for smooth varieties there is a non-degenerate pairing,

$$H^{0,k}(X, \mathcal{O}(E)) \times H^{0,n-k}(X, \mathcal{O}(E^* \otimes \wedge^{n,0} X)) \to \mathbb{C}\,,$$

given by $(\alpha, \beta) \mapsto \int_X \alpha \wedge \beta$. Here the E and E^* components pair to give a number and we have k components $d\overline{z}_i$ from the first term, $n-k$ from the second term and n components dz_i from the coefficients in the second term. This pairing gives the duality,

$$H^{0,k}(X, \mathcal{A}) \cong (H^{0,n-k}(X, \Omega_X \otimes \mathcal{A}))^*\,,$$

known as *Kodaira–Serre duality* [68]. Here Ω_X is the sheaf of top-dimensional holomorphic forms. (The bundle $\wedge^{n,0} X$ is called the canonical bundle and is denoted by K_X. Applying this duality to the deformations gives,

$$\mathfrak{def}([\Sigma, p]) \cong H^0(\Sigma, \Omega_\Sigma \otimes \Omega_\Sigma([p]))^*\,.$$

The elements of this last group take the form $f(z)\,dz \otimes dz$ in local coordinates and are called quadratic differentials.

Kodaira–Serre duality generalizes to more general projective varieties where there is a dualizing sheaf, denoted by ω_X, so that

$$H^{0,k}(X, \mathcal{A}) \cong (H^{0,n-k}(X, \omega_X \otimes \mathcal{A}))^*\,.$$

See Hartshorne [75]. The dualizing sheaf leads to the last family of cohomology classes that we will need. One can construct a bundle over the moduli space, such that the fiber over any point $[u, \Sigma, p]$ is just the space of sections of the dualizing sheaf $H^0(\Sigma, \omega_\Sigma)$. This bundle is called the Hodge bundle and it is denoted by $\mathbb{E}$. The Chern classes of the Hodge bundle are called Hodge classes. The formal definition of the dualizing sheaf is as follows (see Hori et al [76]):

Definition 5.23 The *dualizing sheaf* over a nodal curve Σ denoted by ω_Σ is the sheaf of meromorphic differentials that:

(1) are holomorphic away from the nodes

(2) have at worst a pole of order one at each node branch

(3) have residues that sum to zero on each pair of node branches.

Exercise 5.24 Give a formal definition of the Hodge bundle analogous to the definition of the tautological bundles $\mathcal{L}_k$.

According to the previous application of the Riemann–Roch theorem, for every extra marked point the difference between the dimensions of the automorphisms and deformations decreases by one. Returning to the example in Figure 5.4, we see that adding one point to the left bubble would reduce the dimension of the automorphism group by one because the restriction of the automorphism to the left bubble would have to fix two points, not just one as before. However, this would not change the space of deformations. One might think that a possible position of the new marked point needs to be taken into account. But any motion of the new point is a trivial deformation. Indeed, an automorphism changing its position can can be applied to the space before adding the point. A similar thing occurs with the addition of a second marked point to the left bubble. When one adds a third marked point to the left bubble, the dimension of the group of automorphisms will not change from the dimension with two marked points because there are no automorphisms acting nontrivially on the left bubble (such an automorphism would have to fix at least three points). There would be an extra deformation corresponding to changing the location of the third point.

We now have a fairly good local description of the moduli stack. Since the moduli space models stable curves there are no infinitesimal automorphisms. Assuming that the obstruction space vanishes, we see that the moduli space is locally isomorphic to the quotient of $\mathfrak{def}([u, \Sigma, p])$ by a finite group. The deformation-obstruction sequence can be combined with the Riemann–Roch theorem to compute the dimension of $\mathfrak{def}([u, \Sigma, p])$. This is called the virtual dimension of the moduli space.

Exercise 5.25 Recall that the alternating sum of the dimensions of an exact sequence of spaces is zero. Use the Riemann–Roch theorem

$$\sum_{k=0}^{n}(-1)^k \dim_{\mathbb{C}} H^k(X, E) = \mathrm{Td}(TX)\mathrm{ch}(E)[X]$$

to compute $\dim_{\mathbb{C}} \mathfrak{def}(u) - \dim_{\mathbb{C}} \mathfrak{ob}(u)$. Combine this with our earlier computation of

$$\dim_{\mathbb{C}} \mathfrak{aut}([\Sigma, p]) - \dim_{\mathbb{C}} \mathfrak{def}([\Sigma, p])$$

and the deformation-obstruction sequence to conclude

$$\mathrm{virdim}_{\mathbb{C}}\ \overline{M}_{g,n}(X, \beta) = \int_{\beta} c_1(TX) + (\dim_{\mathbb{C}} X - 3)(1 - g) + n\,.$$

6 Localization

We have come a long way in our review of Gromov–Witten invariants. We described all the technical elements in their definition with the exception of the virtual fundamental class (addressed in a later subsection). We also performed a number of non-trivial computations via recursion or direct reasoning. In this section we will describe a new computational tool called localization used to compute the Gromov–Witten invariants of the resolved conifold X_{S^3}. We start with a general description of localization and an outline of the virtual localization formula for Gromov–Witten invariants. Some readers may prefer to skip down to the sample computations that we give for $\mathbb{C}\mathrm{P}^2$ after the general discussion.

6.1 The Umkehrung

Localization is a technique reducing a computation of an integral over a higher-dimensional space to an integral over a lower-dimensional space. Of course this is impossible in general, but it is instructive to try.

Given an inclusion (or any map) $\iota_F\colon F \hookrightarrow M$, we have the well-known pull-back $\iota_F^*\colon H^k(M) \to H^k(F)$, given on the level of Poincaré duals by the inverse image. There is a less well-known push-forward $\iota_{F!}\colon H^k(F) \to H^{k+m-f}(M)$ defined for any map between oriented manifolds. This map (called the Umkehrung) is defined by the following diagram.

$$\begin{array}{ccc} H^k(F) & \xrightarrow{\iota_{F!}} & H^{k+m-f}(M) \\ \Big\downarrow{\scriptstyle PD} & & \Big\uparrow{\scriptstyle PD^{-1}} \\ H_{f-k}(F) & \xrightarrow[\iota_{F*}]{} & H_{f-k}(M) \end{array}$$

Here m is the dimension of M, f is the dimension of F and PD is Poincaré duality. We will apply these ideas to orbifolds. For orbifolds everything goes through as in the manifold case provided one uses rational coefficients.

There are nice descriptions of the Umkehrung for fibrations and embeddings. As an example, if $\pi_M\colon M \to \mathrm{pt}$, the map $\pi_{M!}\colon H^m(M) \to H^0(\mathrm{pt})$ is just given by integration $f_!\alpha = \int_M \alpha$. When $\pi\colon E \to M$ is a fiber bundle, the Umkehrung is just integration over the fiber (see Bott and Tu [30]).

In the case of an embedding $\iota_F\colon F \hookrightarrow M$, we can compute $\iota_{F!}1$ and get an interesting answer. The cycle dual to 1 in F is just F and so the cycle dual to $\iota_{F!}1$ is just the

image of F in M which is the zero section of the normal bundle to F in M. Recall that the image of the zero section of a bundle is dual to the *Thom class* . The *Euler class* of the bundle is the pull-back under any section of the Thom class [30]. It follows that

$$\iota_F^* \iota_{F!} 1 = e(N(F)) .$$

Remark 6.1 The Umkehrung for an arbitrary map can be decomposed into one for an embedding and one for a projection. Starting with a cycle in F dual to the given cohomology class one obtains a cycle in $F \times M$ by the natural inclusion $F \hookrightarrow F \times M$ taking x to $(x, \iota_F(x))$. Given this cycle in $F \times M$ one obtains a cycle in M by composition with the projection. The class dual to this final cycle is the value of the Umkehrung.

If one could invert the Euler class, one might hope that the maps $\frac{\iota_F^*}{e(N(F))}\colon H^*(M) \to H^*(F)$ and $\iota_{F!}\colon H^*(F) \to H^*(M)$ would be inverses, thus reducing integrals over M to integrals over F. Of course this is too much to expect in general. Unfortunately, cohomology classes of positive degree on a finite-dimensional manifold are all nilpotent (since high enough powers would be forms of degree larger than the dimension of the manifold) and therefore not invertible.

6.2 Equivariant cohomology

It is much more reasonable to expect such a reduction to work if everything is invariant under a group action because as we will see, equivariant cohomology has more invertible elements. The insight of Atiyah and Bott [15] was that this could be made to work when F is the fixed point locus of a torus action on M. The geometric intuition behind the reduction of an integral over a larger set to an integral over a smaller set for equivariant functions is that the symmetry allows one to sample their values at a smaller collection of points.

We now need a brief review of equivariant cohomology. If G is any Lie group and EG is a contractible, free, right, proper G–space and M is any left G–space one can form the twisted product $EG \times_G M := EG \times M/\sim$ where $(eg^{-1}, gm) \sim (e, m)$. The equivariant cohomology of M is defined to be

$$H_G^*(M) := H^*(EG \times_G M) .$$

As an example take $G = T^2$, then $EG = S^\infty \times S^\infty$ where S^∞ is viewed as the unit sphere is an infinite-dimensional complex space and the action is given by

$(x_0, x_1)\cdot(\lambda_0, \lambda_1) := (x_0\lambda_0, x_1\lambda_1)$. The equivariant cohomology of a point is computed as follows [15],

$$H^*_{T^2}(\mathrm{pt};\mathbb{Q}) = H^*(\mathbb{C}\mathrm{P}^\infty \times \mathbb{C}\mathrm{P}^\infty;\mathbb{Q}) \cong \mathbb{Q}[\alpha_0, \alpha_1]\,.$$

Notice that $\mathbb{C}\mathrm{P}^\infty$ is infinite dimensional and α_k have degree 2. Given a representation $\mu\colon G \to GL_n\mathbb{C}$ we get a vector bundle $EG \times_G \mathbb{C}^n \to EG \times_G \mathrm{pt}$, and associated equivariant classes given by the Chern classes of this bundle $c_k(EG \times_G \mathbb{C}^n)$. If $\mu_k\colon T^n \to GL_1\mathbb{C}$ is the projection to the kth factor and μ_k^* is the dual representation then the first Chern class of the line bundle associated to μ_k^* is the standard generator α_k of $H^*_{T^2}(\mathrm{pt};\mathbb{Q})$. This corresponds to the fact that the first Chern class of the tautological line bundle over projective space evaluates to -1 on a standardly oriented generator of the second homology. In fact homogeneous polynomials of degree d may be regarded as sections of the degree d (as measured by the first Chern class) line bundle over projective space and this is where all of the sign conventions come from.

A second example of equivariant cohomology is the equivariant cohomology of the group with the natural left action. The result is

$$H^*_G(G) := H^*(EG \times_G G) = H^*(EG) = \mathbb{Q}\,.$$

Thus when the action was free the equivariant cohomology was trivial and when the action was trivial the equivariant cohomology was interesting. In some sense equivariant cohomology is generated by the fixed point set of the action.

From here on forward we will work with torus actions only and T will always denote a torus. In order to make the observation that the equivariant cohomology is generated by the fixed point set more precise and follow the reduction outline from the beginning of this section we need to invert elements of the equivariant cohomology and ultimately invert the Euler class of the normal bundle to the fixed point locus. Notice that $H^*_{T^{n+1}}(\mathrm{pt}) \cong \mathbb{Q}[\alpha_0, \ldots, \alpha_n]$ is an integral domain. This is in stark contrast to the cohomology of finite dimensional manifolds. We let

$$F^*_{T^{n+1}} \cong \mathbb{Q}(\alpha_0, \ldots, \alpha_n)$$

be the associated fraction field. The obvious map $EG \times_G M \to EG \times_G \mathrm{pt}$ induces a map $H^*_G(\mathrm{pt}) \to H^*_G(M)$ giving $H^*_G(M)$ the structure of an $H^*_G(\mathrm{pt})$–module. This leads us to the first version of the theorem of Atiyah and Bott.

Theorem 6.2 *Let Fix be the fixed point locus of a torus T action on M. The map $ET \times_T Fix \to ET \times_T M$ induces an isomorphism*

$$\iota^*_{Fix}\colon H^*_T(M) \otimes_{H^*_T(pt)} F^*_T \to H^*_T(Fix) \otimes_{H^*_T(pt)} F^*_T\,.$$

Atiyah and Bott actually state a more refined theorem that specifies which elements need to be inverted in order to obtain an isomorphism. The proof of this theorem is to use the formula $\iota_F^*\iota_{F!} = e(N(F))$ to see that the maps $\iota_{\mathrm{Fix}!}$ and $Q := \sum_F \frac{\iota_F^*}{e(N(F))}$ are inverses of each other. Here F represents a component of Fix and the sum is taken over all such components.

In order to apply these ideas to the integration of ordinary cohomology classes on M, notice that the map $j\colon M \to ET \times_T M$ taking x to the equivalence class of (e_0, x) induces a map $H_T^*(M) \to H^*(M)$ via pull-back. A class in the image of this map is called an equivariant class. One standard way to construct equivariant classes is to start with an equivariant vector bundle $E \to M$ and take characteristic classes. An equivariant bundle is just a vector bundle together with a torus action compatible with the bundle structure. To any such bundle one can associate the pull-back of the induced bundle $ET \times_T E \to ET \times_T M$. Any characteristic class of $ET \times_T E$ is then an equivariant class. Given any equivariant class $\phi \in H^m(M)$ with equivariant lift $\widehat{\phi} \in H_T^m(M)$ we have

$$\begin{aligned}(6)\quad \int_M \phi = \pi_{M!}\phi = \pi_{M!}j^*\widehat{\phi} = \iota_{\mathrm{pt}}^*\pi_{M!}\widehat{\phi} = \iota_{\mathrm{pt}}^*\pi_{M!}\iota_{\mathrm{Fix}!}\sum_F \frac{\iota_F^*\widehat{\phi}}{e(N(F))}\\ = \iota_{\mathrm{pt}}^*\pi_{\mathrm{Fix}!}\sum_F \frac{\iota_F^*\widehat{\phi}}{e(N(F))} = \sum_F \int_F \frac{\iota_F^*\widehat{\phi}}{e(N(F))}.\end{aligned}$$

This is the standard way of thinking about the Atiyah–Bott localization formula.

6.3 Equivariant cohomology of $\mathbb{C}\mathrm{P}^n$

To use the localization formula (6) to integrate a cohomology class ϕ one must find an equivariant lift $\widehat{\phi}$ that maps to ϕ. The standard way to do this is to express ϕ as a product of characteristic classes of vector bundles and then extend the group action over the vector bundles. Such an extension is called a *linearization of the action*. For example, we can represent the hyperplane bundle over $\mathbb{C}\mathrm{P}^n$ by $(\mathbb{C}^{n+1} - \{0\}) \times \mathbb{C}/\sim$ with $(z_0, \dots, z_n, \xi) \sim (wz_0, \dots, wz_n, w\xi)$. Define a family of linearizations of this bundle by $\lambda\cdot(z_0, \dots, z_n, \xi) := (\lambda_0 z_0, \dots, \lambda_n z_n, \lambda_0^{q_0} \dots \lambda_n^{q_n}\xi)$. Denoting the hyperplane bundle with this linearization by $L_{q_0,\dots,q_n}$, we have

$$c_1(L_{q_0,\dots,q_n}) = h - q_0\alpha_0 - \dots - q_n\alpha_n\,.$$

This serves to define equivariant cohomology classes of $\mathbb{C}\mathrm{P}^n$. The class h is defined to be the first Chern class of the bundle $L_{0,\dots,0}$, and α_k is defined as a difference of

Chern classes. Let us outline a proof that

$$H_T^*(\mathbb{C}\mathrm{P}^n) = \mathbb{Q}[h, \alpha_0, \dots, \alpha_n] / \left(\prod_{k=0}^{n} (h - \alpha_k) \right) .$$

First consider a finite-dimensional model $X_{p,n,N}$ for $ET \times_T \mathbb{C}\mathrm{P}^p$. If $T = T^{n+1}$ acts on $\mathbb{C}\mathrm{P}^p$ in the usual way then

$$X_{p,n,N} := \left\{ ([x_0], \dots, [x_n]; [y_0 : \dots : y_p]) \in (\mathbb{C}\mathrm{P}^N)^{n+1} \times \mathbb{C}\mathrm{P}^{(p+1)(N+1)-1} \;\middle|\; \wedge^2 \begin{bmatrix} x_k \\ y_k \end{bmatrix} = 0 \right\}.$$

A specific case helps explain what this means. Take $n = 2$, $p = 2$, and $N = 4$. We will write $[x_0] = [x_0 : x_1 : x_2 : x_3 : x_4]$, $[x_1] = [y_0 : \dots : y_4]$, $[x_2] = [z_0 : \dots : z_4]$ and

$$[y_0 : \dots : y_2] = [p_0 : p_1 : p_2 : p_3 : p_4 : q_0 : q_1 : q_2 : q_3 : q_4 : r_0 : r_1 : r_2 : r_3 : r_4]$$

The condition on the second wedge states that all 2×2 determinants pairing x's and p's etc are zero, so

$$\begin{vmatrix} x_0 & x_3 \\ p_0 & p_3 \end{vmatrix} = 0, \qquad \begin{vmatrix} z_2 & z_3 \\ r_2 & r_3 \end{vmatrix} = 0, \dots .$$

To see where this comes from note that there is a natural action of S^1 on S^{2N+1} with quotient $\mathbb{C}\mathrm{P}^N$. In the $N \to \infty$ limit we see that S^∞ is contractible, so $ES^1 = S^\infty$ and $BS^1 = \mathbb{C}\mathrm{P}^\infty$. Thus $(\mathbb{C}\mathrm{P}^N)^{n+1}$ is a finite-dimensional model for BT. The condition on the vanishing of the second exterior powers implies that y_0 is proportional to x_0, etc. This means that the inverse image of a point under the natural projection $X_{p,n,N} \to (\mathbb{C}\mathrm{P}^N)^{n+1}$ is a copy of $\mathbb{C}\mathrm{P}^p$. The class h is Poincaré dual to

$$H := \{ ([x_0], \dots, [x_n]; [y_0 : \dots : y_p]) \in X_{p,n,N} \,|\, \langle x_p, y_p \rangle = 0 \} ,$$

and the class α_k is Poincaré dual to

$$A_k := \{ ([x_0], \dots, [x_n]; [y_0 : \dots : y_p]) \in X_{p,n,N} \,|\, (x_k)_N = 0 \} .$$

These formulas can serve as alternative definitions of these classes. The following exercise will prove that

$$H_T^*(X_{p,n,N}) = \mathbb{Q}[h, \alpha_0, \dots, \alpha_n] / \left(\prod_{k=0}^{n} (h - \alpha_k, \alpha_0^{N+1}, \dots, \alpha_n^{N+1}) \right) .$$

Exercise 6.3 Prove that the cohomology rings of $X_{p,n,N}$ and $\mathbb{C}\mathrm{P}^n$ take the stated form. Note for example that $X_{2,2,4}$ is a 28–real-dimensional space with one 28–cell, no 27–cells and $X^{(26)}_{2,2,4} = X_{1,2,4} \cup \bigcup_{k=0}^{4} A_k$. In fact, by properly taking complements

of unions of intersections of the A_k and H cycles, $X_{p,n,N}$ may be decomposed into a union of cells. It follows that the cohomology group of $X_{p,n,N}$ is as stated. The ring structure follows from the combinatorics of the intersections of the A_k and H cycles.

Exercise 6.4 Let q_k be the point in $\mathbb{C}\mathrm{P}^m$ with all coordinates other than z_k equal to zero. These points are fixed by the standard T–action, so there is an equivariant class ϕ_k Poincaré dual to q_k. By considering appropriate intersections of the A_k and H prove that

$$\phi_k = \prod_{j\neq k}(h-\alpha_j)\,.$$

Let $\iota_k\colon \mathrm{pt}\to\mathbb{C}\mathrm{P}^m$ be the map with image q_k. We compute

$$\iota_k^* h = \int_{\mathbb{C}\mathrm{P}^m} PD(q_k)h = \int_{\mathbb{C}\mathrm{P}^m} PD(q_k)(h-\alpha_k) + \int_{\mathbb{C}\mathrm{P}^m} PD(q_k)\alpha_k = \alpha_k\,.$$

The above computation may look a bit weird. It appears that we are restricting a 2–form to a point and getting a non-zero answer. The thing to remember here is that we are working equivariantly so we replace q_k by $ET\times_T q_k$ and $\mathbb{C}\mathrm{P}^m$ by $ET\times_T\mathbb{C}\mathrm{P}^m$, and we have tensored with $\mathbb{Q}(\alpha_0,\dots,\alpha_m)$.

7 Localization computations of Gromov–Witten invariants

As with many parts of this theory Kontsevich was the first to apply localization to Gromov–Witten invariants [89]. We are following the exposition from Hori et al [76] and Cox and Katz [43]. To see how this applies to Gromov–Witten invariants, notice that a T action on X will induce a T action on $\overline{\mathcal{M}}_{g,n}(X,\beta)$ via composition $\lambda\cdot[u,\Sigma,p] := [\lambda u,\Sigma,p]$. In order to better demonstrate how localization may be used to compute Gromov–Witten invariants we will recompute some of the Gromov–Witten invariants of $\mathbb{C}\mathrm{P}^2$. This is not a very efficient way to compute these invariants but it will allow us to explain the important points.

We should point out that thus far our discussion of localization only applies to well-behaved spaces with torus actions. In the Gromov–Witten setting this will only occur when the automorphism group of every stable map in the relevant moduli space is trivial and the obstruction space over each stable curve is zero as well, that is, $\mathrm{Aut}([u,\Sigma,p]) = 1$ and $\mathfrak{ob}([u,\sigma,p]) = 0$. When the obstruction spaces are still trivial but the automorphism groups are not, one must generalize the localization formula to the orbifold setting. When the obstruction spaces do not vanish, one must define virtual fundamental cycles and prove that localization works for these virtual cycles. In this

section we will ignore the more technical points and discuss localization in the context of Gromov–Witten invariants as though everything were smooth and automorphism-free. Take heart, we will give an introduction to virtual fundamental cycles later in this section and (virtual) localization really does work in this context as was shown by Graber and Pandharipande [66].

We have a standard T^3 action on $\mathbb{C}\mathrm{P}^2$ given by

$$(\lambda_0, \lambda_1, \lambda_2) \cdot [z_0 : z_1 : z_2] := [\lambda_0 z_0 : \lambda_1 z_1 : \lambda_2 z_2].$$

The first step is to find the fixed point set of the T action on $\overline{\mathcal{M}}_{g,n}(\mathbb{C}\mathrm{P}^2, d)$. The subtle point is that if one looks for maps as opposed to equivalence classes of maps fixed by T there will be none (unless $d = 0$ in which case a constant map with value a fixed point in $\mathbb{C}\mathrm{P}^2$ would be fixed). Thus one must remember to look for fixed equivalence classes. This means that given a stable map $[u, \Sigma, p]$ and a $\lambda \in T$ one must find an automorphism φ of the underlying marked surface so that $\lambda u = u \circ \varphi$.

Example 7.1 An example of a genus zero degree three stable map fixed by the natural T action is $[u, \mathbb{C}\mathrm{P}^1]$ where $u([x_0 : x_1]) = [x_0^3 : x_1^3 : 0]$. This is a stable map because the only automorphisms are given by $\varphi([x_0 : x_1]) = [\xi x_0 : x_1]$ where $\xi^3 = 1$. It is fixed by every $\lambda \in T$ because the map $\varphi([x_0 : x_1]) = [\xi_0 x_0 : \xi_1 x_1]$ induces an equivalence between $[\lambda \cdot u, \mathbb{C}\mathrm{P}^1]$ and $[u, \mathbb{C}\mathrm{P}^1]$ where $\xi_0^3 = \lambda_0$ and $\xi_1^3 = \lambda_1$.

Remark 7.2 A stable map fixed by T has an infinite number of left symmetries. One should not confuse this with the requirement of at most a finite number of right symmetries from the definition of a stable map. The first is the T action by post-composition. The second is pre-composition by an automorphism of the marked surface.

In general, a component of the fixed point set of the T action on $\overline{\mathcal{M}}_{g,n}(\mathbb{C}\mathrm{P}^2, d)$ can be described by a labeled graph, denoted Γ. The labels on the graphs and the correspondence between labeled graphs and components of the fixed point set will be described in Section 7.1. See Figure 7.1 for a labeled graph and element of the corresponding fixed point component.

Applying the localization formula (6) to the Gromov–Witten invariants of $\mathbb{C}\mathrm{P}^m$ gives

$$\left\langle h^{\ell_1} \dots h^{\ell_n} \right\rangle_{g,d[\mathbb{C}\mathrm{P}^1]}^{\mathbb{C}\mathrm{P}^m} = \sum_{\Gamma} \frac{1}{|\mathbb{A}_\Gamma|} \int_{\Gamma} \frac{\prod_{j=1}^n \widehat{h}|_{k(u(p_j))}^{\ell_j}}{e(N_\Gamma^{\mathrm{vir}})}. \tag{7}$$

The notations used in this formula are explained further in the next article. Here $\widehat{h}$ is an equivariant lift of the class h. The standard lift is also denoted by h, in which

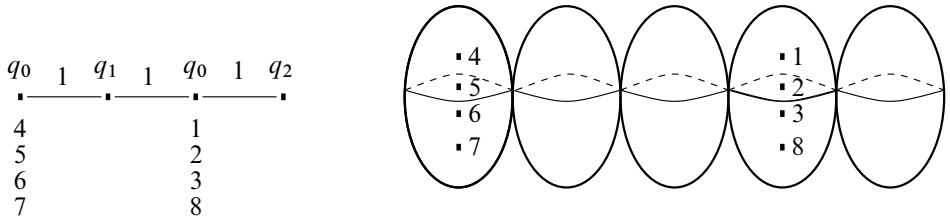


Figure 7.1: A labeled graph and element

case $\widehat{h}|_{k(u_j)} = \alpha_{k(u_j)}$. More generally, lifts of h take the form $h + \sum a_j\alpha_j$ and $(h + \sum a_j\alpha_j)|_{k(u_j)} = \alpha_{k(u_j)} + \sum a_j\alpha_j$. We have to divide each term by a group factor to take into account the fact that the moduli space is a stack (think orbifold). The group $\mathbb{A}_\Gamma$ is the automorphism group of a generic element of the fixed point set Γ. The automorphisms of the labeled graph $\mathrm{Aut}(\Gamma)$ act on the edges of the graph and therefore on the group $\prod_e \mathbb{Z}_{d(e)}$. The group $\mathbb{A}_\Gamma$ is defined as the following semidirect product:

$$\mathbb{A}_\Gamma := \mathrm{Aut}(\Gamma) \ltimes \left(\prod_e \mathbb{Z}_{d(e)} \right) .$$

Formula (7) gives the Gromov–Witten invariants of $\mathbb{C}\mathrm{P}^m$ for any genus. For genus zero one just has to include the automorphism groups; for higher genus one also has to apply virtual localization. Of course, there is still work to do to get numbers from this formula. In particular we still have to compute the Euler class $e(N_\Gamma^{\mathrm{vir}})$.

The next step in the localization computation is the computation of the Euler class of the normal bundle to each component of the fixed point set. We have seen that the tangent space to the moduli stack at a stable curve is $\mathfrak{def}([u, \Sigma, p])$. Assuming that $[u, \Sigma, p]$ is a fixed point of the T action there will be an induced action on the tangent space. This linear space decomposes into a collection of irreducible T representations. The tangent space to the fixed point set at $[u, \Sigma, p]$ is the sum of the trivial T representations because the fixed point set is well, fixed. The normal space is therefore the sum of the nontrivial representations and is denoted by $\mathfrak{def}([u, \Sigma, p])^{\mathrm{mov}}$. In fact, we will see that the T action extends to all of the spaces in the deformation-obstruction complex, so the representation $\mathfrak{def}([u, \Sigma, p])^{\mathrm{mov}}$ may be deduced from similar parts from the other terms in the deformation-obstruction complex.

Recall that given an exact sequence of vector bundles

$$0 \to E_0 \to E_1 \to \cdots \to E_{2n+1} \to 0 ,$$

one has the relation $\prod_{k=0}^{n} e(E_{2k}) = \prod_{k=0}^{n} e(E_{2k+1})$. Applying this to the moving part of the deformation-obstruction sequence (5), we obtain

$$e(\mathfrak{def}([u, \Sigma, p])^{\mathrm{mov}}) = \frac{e(\mathfrak{def}(u)^{\mathrm{mov}})e(\mathfrak{def}([\Sigma, p])^{\mathrm{mov}})e(\mathfrak{ob}([u, \Sigma, p])^{\mathrm{mov}})}{e(\mathfrak{aut}([\Sigma, p])^{\mathrm{mov}})e(\mathfrak{ob}(u)^{\mathrm{mov}})}.$$

This last formula divides the computation of the Euler class of the normal bundle required for localization calculations into more tractable parts. Instead of analyzing all deformations of a stable map we are able to analyze deformations of the map that fix the marked curve, deformations of the marked curve, etc separately. This program is carried out in detail in Sections 8.1, 8.2, 8.3 and 8.4 where we compute $e(\mathfrak{aut}([\Sigma, p])^{\mathrm{mov}})$ first, describe the torus actions second, compute $e(\mathfrak{def}([\Sigma, p])^{\mathrm{mov}})$ third, and the ratio $e(\mathfrak{def}(u)^{\mathrm{mov}})/e(\mathfrak{ob}(u)^{\mathrm{mov}})$ last. The term $e(\mathfrak{ob}([u, \Sigma, p])^{\mathrm{mov}})$ is addressed in Section 11. Before embarking on this program we collect all of the resulting formulas and provide two examples.

7.1 Representation of fixed point components by graphs

We now describe the correspondence between components of the fixed point set and labeled graphs. This article introduces the notation used throughout this and the following sections. Recall that the domain of a stable map is a prestable genus g curve with marked points. A stable map fixed by the standard torus action maps each node of the curve to one of the points $q_i := [0 : \ldots : 1 : \ldots : 0]$, which are the points fixed by the standard action on $\mathbb{C}\mathrm{P}^m$. Some irreducible components of the prestable curve are mapped entirely to a single q_i. We call such components contracted components or ghost bubbles. Note that there are no non-constant maps of higher genus curves into $\mathbb{C}\mathrm{P}^m$ fixed by the standard torus action. Therefore if a component is not contracted it has to be a copy of $\mathbb{C}\mathrm{P}^1$ that is mapped onto a projective line containing exactly two of the q_i fixed points. Notice that any two stable maps in the same component of the fixed point set have the same non-contracted $\mathbb{C}\mathrm{P}^1$ configuration.

Thus we can describe a component of the fixed point set by a graph with edges corresponding to non-contracted components and vertices corresponding to the components of the preimages of the q_i. Each edge e is labeled with a positive integer $d(e)$ indicating the degree of the map and each vertex is labeled with one of the fixed points q_i. In addition, marked points are listed in columns under the vertices (see Figure 7.1). Finally, contracted components unlike non-contracted ones, can have higher genus and in principle this has to be indicated at the vertices as well. We adopt the convention that the absence of such a label corresponds to genus 0. In particular, in this subsection we are only concerned with the Gromov–Witten invariants of rational curves and no genus labeling is necessary. Table 7.1 summarizes the notation used to label these graphs.

Remark 7.3 One can visualize a flag by drawing an arrow on the edge; the source of the arrow along with the edge is the flag. Notice that this use of the term flag agrees with the usual definition in terms of increasing sequences of subspaces (see Harris [73] and Griffiths–Harris [68]) because a vertex in the graph corresponds to a point in $\mathbb{C}\mathrm{P}^m$ which is a 1–dimensional subspace of $\mathbb{C}^{m+1}$ and an edge in the graph corresponds to a line in $\mathbb{C}\mathrm{P}^m$ which is a 2–dimensional subspace of $\mathbb{C}^{m+1}$.

Example 7.4 For the graph in Figure 7.1 let v denote the leftmost vertex. Then $\mathrm{val}(v) = 1$, $d(v) = 1$, $n(v) = 4$, $k(v) = 0$, $k(v') = 1$, $g(v)$=0, and $k(u(p_7)) = 0$.

Notation	Description
$\mathrm{val}(v)$	The valence of the vertex v, that is, the number of edges incident to it. If $\mathrm{val}(v) = 1$ we let v' denote the vertex on the other side of this single edge and if $\mathrm{val}(v) = 2$ let v_1 and v_2 denote the other two vertices on the two edges.
$n(v)$	The number of marked points listed under the vertex. Some authors draw graphs with 'legs' to indicate the marked points.
$k(v)$	The index of the fixed point in $\mathbb{C}\mathrm{P}^m$ corresponding to the vertex, for example if v is labeled by q_i then $k(v) = i$.
α_k	The first Chern class of the line bundle associated to the dual of the representation given by projection to the kth factor of the torus.
$g(v)$	The genus of the contracted component associated to the vertex. We set $g(v) = 0$ if there is no such component at the vertex.
$d(e)$	The degree of the map of the non-contracted component corresponding to edge e. If v has valence one $d(v)$ will be the degree of the unique edge meeting v. If v has valence two $d(e_1)$ and $d(e_2)$ will denote the degrees of the edges containing v_1 and v_2 respectively.
F	A flag in the graph of a fixed point component, that is, a pair of a vertex and an incident edge. We will use the flag to denote the corresponding vertex or point on the prestable curve or image point in $\mathbb{C}\mathrm{P}^m$ without comment.
$k(u(p_j))$	The label image of the jth marked point.

Table 7.1: Graph labels

We are now ready to present the formulas that are used in localization computations of Gromov–Witten invariants of $\mathbb{C}\mathrm{P}^m$.

Remark 7.5 We note that the index j in (11)–(13) runs over all possible values, not just the q_j depicted on the graph. For example, if we are in $\mathbb{C}\mathrm{P}^3$ one should take into account terms with $j=3$ even when q_3 is not on the graph. See Remarks 8.5 and 8.7 in the proofs.

7.2 Formulas used in localization

$$e(\mathfrak{aut}([\Sigma,p])^{\mathrm{mov}}) = \prod_{\substack{\mathrm{val}(v)=1\\ n(v)=0\\ g(v)=0}} \frac{\alpha_{k(v)}-\alpha_{k(v')}}{d(v)}. \tag{8}$$

$$e(\mathfrak{def}([\Sigma,p])^{\mathrm{mov}}) = \prod_{\substack{\mathrm{val}(v)=2\\ n(v)=0\\ g(v)=0}} \Big(\frac{\alpha_{k(v)}-\alpha_{k(v_1)}}{d(e_1)} + \frac{\alpha_{k(v)}-\alpha_{k(v_2)}}{d(e_2)}\Big) \prod_{\mathrm{val}(F)+n(F)+2g(v)>2} \Big(\frac{\alpha_{k(F)}-\alpha_{k(v')}}{d(F)} - \psi_F\Big). \tag{9}$$

$$\frac{e\big(\mathfrak{def}(u)^{\mathrm{mov}}\big)}{e\big(\mathfrak{ob}(u)^{\mathrm{mov}}\big)} = \frac{e\big(H^0\big(\widehat{\Sigma},\mathcal{O}_{\widehat{\Sigma}}\big(\nu^*u^*T\mathbb{C}\mathrm{P}^2\big)\big)^{\mathrm{mov}}\big)}{e\big(H^1\big(\widehat{\Sigma},\mathcal{O}_{\widehat{\Sigma}}\big(\nu^*u^*T\mathbb{C}\mathrm{P}^2\big)\big)^{\mathrm{mov}}\big)\prod_c e\big(T_{u(c)}\mathbb{C}\mathrm{P}^2\big)}. \tag{10}$$

$$\prod_c e(T_{u(c)}\mathbb{C}\mathrm{P}^2)) = \prod_F \prod_{j\neq k(F)} (\alpha_{k(F)}-\alpha_j). \tag{11}$$

$$e\big(H^0\big(\widehat{\Sigma},\mathcal{O}\big(u^*T\mathbb{C}\mathrm{P}^2\big)\big)^{\mathrm{mov}}\big) = \prod_v \prod_{j\neq k(v)} (\alpha_{k(v)}-\alpha_j) \prod_e \frac{(-1)^{d(e)}(d(e)!)^2}{d(e)^{2d(e)}}\big(\alpha_{k(v)}-\alpha_{k(v')}\big)^{2d(e)} \prod_{\substack{a+b=d\\ j\neq k(v),k(v')}} \Big(\frac{a}{d(e)}\alpha_{k(v)} + \frac{b}{d(e)}\alpha_{k(v')} - \alpha_j\Big). \tag{12}$$

$$e(H^1(\widehat{\Sigma},\mathcal{O}(\nu^*u^*T\mathbb{C}\mathrm{P}^2))^{\mathrm{mov}}) = \prod_{\mathrm{val}(v)+n(v)+2g(v)>2} \prod_{j\neq k(v)} \sum_{i=0}^{g(v)} c_i(\mathbb{E}^\vee)(\alpha_{k(v)}-\alpha_j)^{g-i}. \tag{13}$$

$$e(N_\Gamma) = e(\mathfrak{def}([u,\Sigma,p])^{\mathrm{mov}}) = \frac{e(\mathfrak{def}(u)^{\mathrm{mov}})e(\mathfrak{def}([\Sigma,p])^{\mathrm{mov}})e(\mathfrak{ob}([u,\Sigma,p])^{\mathrm{mov}})}{e(\mathfrak{aut}([\Sigma,p])^{\mathrm{mov}})e(\mathfrak{ob}(u)^{\mathrm{mov}})}. \tag{14}$$

$$\langle h^{\ell_1}\dots h^{\ell_n}\rangle^{\mathbb{C}\mathrm{P}^m}_{g,d[\mathbb{C}\mathrm{P}^1]} = \sum_{\Gamma}\frac{1}{|\mathbb{A}_\Gamma|}\int_\Gamma \frac{\prod_{j=1}^n \widehat{h}\big|^{\ell_j}_{k(u(p_j))}}{e(N_\Gamma)}\,. \tag{15}$$

Remark 7.6 In the localization formula (15) the sum is taken over all labeled graphs with n marked points having the correct genus and degree. The integral is taken over the moduli space $\mathcal{M}_\Gamma$ of all stable maps having the given graph. To evaluate these integrals one has to expand $\frac{1}{e(N_\Gamma)}$ into an infinite series in the ψ classes and integrate the (finite number of) terms of top degree. Since the values of the integral are in the equivariant cohomology ring of $\mathbb{C}\mathrm{P}^m$ the Chern classes α_k play the role of constants and can simply be factored out of the integral. Thus, actual integration is only required when one of the descendant classes ψ_F from (9) is non-zero in which case we can use the string and dilaton equations of Section 4.3 to evaluate the integral. Sample computations in the next article will clarify the details and should convince the reader that it is possible to extract useful information from these cumbersome formulas.

7.3 Small degree invariants of rational curves in $\mathbb{C}\mathrm{P}^2$

In this article we demonstrate the localization formulas by computing the genus zero invariants in degree one, two and three. We start with

$$N_1 := \langle h^2h^2\rangle^{\mathbb{C}\mathrm{P}^2}_{0,1[\mathbb{C}\mathrm{P}^1]},$$

where h is the Poincaré dual to the hyperplane class. We can lift h to an equivariant class that we denote by the same letter. Intuitively, we are looking for the number of lines through two generic points in $\mathbb{C}\mathrm{P}^2$ so you can guess that the answer is one, but it is instructive to see how localization produces this answer.

We are working with fixed stable maps in $\overline{\mathcal{M}}_{0,2}(\mathbb{C}\mathrm{P}^2,[\mathbb{C}\mathrm{P}^1])$. Since the overall degree is one, our graph can only have one edge and it will be labeled with a 1. The two vertices are labeled with two of the three points $q_0=[1:0:0]$, $q_1=[0:1:0]$, $q_2=[0:0:1]$. We also have to distribute the 2 marked points between the two vertices and there are two essentially different ways to do this: to place the marked points at the same vertex or at different vertices. This leads to two different labeled graph types. There are a total of 12 graphs to consider, 6 of type one and 6 of type two. Due to the obvious symmetry, contributions from all graphs within a type are similar.

For the first type we consider the graph labeled by q_0 and q_1 with the two marked points placed at q_0. The only automorphisms are trivial so $|\mathrm{Aut}(\Sigma_\Gamma)|=1$ and since both marked points are mapped to q_0 we have $\iota^*\mathrm{ev}_1^*h^2=\alpha_0^2$ and $\iota^*\mathrm{ev}_2^*h^2=\alpha_0^2$.

Following (8)–(14) we get

$$e(\mathfrak{aut}([\Sigma, p])^{\mathrm{mov}}) = \frac{\alpha_1 - \alpha_0}{1} = \alpha_1 - \alpha_0$$
$$e(\mathfrak{def}([\Sigma, p])^{\mathrm{mov}}) = 1 \cdot 1 \cdot \Big(\frac{\alpha_0 - \alpha_1}{1} - \psi\Big) = \alpha_0 - \alpha_1$$

Recall that the ψ–classes are the first Chern classes of the tautological bundles. In this case the components of the fixed point set are points so all ψ–classes vanish.

$$\prod_c e(T_{u(c)}\mathbb{C}\mathrm{P}^2)) = (\alpha_0-\alpha_1)(\alpha_0-\alpha_2)\cdot(\alpha_1-\alpha_0)(\alpha_1-\alpha_2)e(H^0(\widehat{\Sigma}, \mathcal{O}(u^*T\mathbb{C}\mathrm{P}^2))^{\mathrm{mov}})$$
$$= (\alpha_0-\alpha_1)(\alpha_0-\alpha_2)\cdot(\alpha_1-\alpha_0)(\alpha_1-\alpha_2)\cdot\frac{(-1)^1(1!)^2}{1^{2\cdot 1}}(\alpha_0-\alpha_1)^{2\cdot 1}(\alpha_0-\alpha_2)(\alpha_1-\alpha_2)$$
$$e(H^1(\widehat{\Sigma}, \mathcal{O}(\nu^*u^*T\mathbb{C}\mathrm{P}^2))^{\mathrm{mov}}) = 1, \qquad \text{since all vertices have genus } 0.$$

In fact we could have seen that $e(H^1(\widehat{\Sigma}, \mathcal{O}(u^*T\mathbb{C}\mathrm{P}^2))^{\mathrm{mov}}) = 1$ directly since the fiber of the obstruction bundle $\mathfrak{ob}(u)$ is $H^1(\Sigma, \mathcal{O}_\Sigma(u^*T\mathbb{C}\mathrm{P}^2)) = 0$. Using the deformation-obstruction sequence (5) this implies that $e(\mathfrak{ob}(u)^{\mathrm{mov}}) = 0$ as well. Continuing,

$$\frac{e(\mathfrak{def}(u)^{\mathrm{mov}})}{e(\mathfrak{ob}(u)^{\mathrm{mov}})} = (\alpha_0-\alpha_1)^2(\alpha_0-\alpha_2)(\alpha_1-\alpha_2)$$
$$e(N_\Gamma) = \frac{\alpha_0-\alpha_1}{\alpha_1-\alpha_0}\cdot -(\alpha_0-\alpha_1)^2(\alpha_0-\alpha_2)(\alpha_1-\alpha_2)$$
$$= (\alpha_0-\alpha_1)^2(\alpha_0-\alpha_2)(\alpha_1-\alpha_2)\,.$$

The computation of the Euler class of the normal bundle for the second graph type is analogous and we leave it as an exercise.

Exercise 7.7 Repeat the above computation for the one-edge graph labeled with q_0 and p_1 labeling one vertex and q_1 and p_2 labeling the other. You should get

$$e(N_\Gamma) = -(\alpha_0-\alpha_1)^2(\alpha_0-\alpha_2)(\alpha_1-\alpha_2).$$

The Euler classes of the normal bundles for the other graphs can be obtained from the first two examples by symmetry considerations. There are no ψ–classes to integrate so the integral in (15) can be dropped. It is convenient to first sum up all the contributions from graphs labeled with q_0, q_1. There are four such graphs and the remaining two can be obtained from the ones we already computed by switching α_0 and α_1. By (15)

the contribution to N_1 from these four graphs is

$$\frac{1}{(\alpha_0-\alpha_1)^2(\alpha_0-\alpha_2)(\alpha_1-\alpha_2)}\cdot(\alpha_0^2\cdot\alpha_0^2+\alpha_1^2\cdot\alpha_1^2)$$
$$-\frac{1}{(\alpha_0-\alpha_1)^2(\alpha_0-\alpha_2)(\alpha_1-\alpha_2)}\cdot(\alpha_0^2\cdot\alpha_1^2+\alpha_1^2\cdot\alpha_0^2)$$
$$=\frac{(\alpha_0^2-\alpha_1^2)^2}{(\alpha_0-\alpha_1)^2(\alpha_0-\alpha_2)(\alpha_1-\alpha_2)}$$

There are eight more graphs to account for, four labeled with q_0, q_2 and four labeled with q_1, q_2. The joint contributions of each four are obtained from the last expression by applying the obvious substitutions.

Exercise 7.8 Add up the three fractions and get $N_1 = 1$ as expected.

We can repeat the previous computation with different linearizations (lifts to $H_T^*(\mathbb{C}\mathrm{P}^2)$ of the class h). For example, since h is the first Chern class of the $\mathcal{O}(1)$ line bundle we can construct lifts as the equivariant Chern classes of the same line bundle with various group actions. The lift with the same name comes from the first Chern class of the standard extension of the T action on $\mathbb{C}\mathrm{P}^2$ to this bundle. We may take the tensor product with a trivial line bundle with the μ_1–action to change the action. This does not change the Chern class but it does change the equivariant lift to $\widehat{h} = h-\alpha_1$. If we use $(h-\alpha_1)(h-\alpha_2)$ as the linearization for the cohomology class corresponding to the first marked point and $(h-\alpha_0)(h-\alpha_2)$ for the second marked point then only one of the 12 graphs will contribute to the sum. Namely, the graph with one vertex labeled with q_0 and p_1 and the other vertex labeled with q_1 and p_2.

Exercise 7.9 Repeat the computation of

$$\langle h^2h^2\rangle_{0,1[\mathbb{C}\mathrm{P}^1]}^{\mathbb{C}\mathrm{P}^2}$$

using the $(h-\alpha_1)(h-\alpha_2)$ and $(h-\alpha_0)(h-\alpha_2)$ linearizations.

Generally speaking components of the fixed point set of a torus action on a moduli space of stable curves can be expressed as a finite quotient of a product of moduli spaces of stable maps into the fixed point set of the target manifold. The best way to understand this is to work out some less trivial examples.

To demonstrate a less trivial localization computation we will conclude the subsection with a computation of

$$N_2 := \langle h^2h^2h^2h^2h^2\rangle_{0,2[\mathbb{C}\mathrm{P}^1]}^{\mathbb{C}\mathrm{P}^2}$$

and some of the terms from the computation of

$$N_3 := \langle h^2h^2h^2h^2h^2h^2h^2h^2h^2\rangle^{\mathbb{C}\mathrm{P}^2}_{0,3[\mathbb{C}\mathrm{P}^1]}.$$

It is convenient to use the linearization $(h-\alpha_1)(h-\alpha_2)$ on the class associated to each marked point. This forces all of the marked points to be mapped to q_0 if a graph is going to contribute to the sum.

The graphs that contribute to the N_2 computation come in one of two types. We label these two types by $I(k_0k_1)$ and $II(k_0k_1k_2)$. The corresponding graphs are displayed in Figure 7.2 along with the graphs that contribute to the N_3 computation.

Each graph type contains all of the graphs obtained by an admissible labeling. An admissible labeling consists of assigning each k_j a value of 0, 1 or 2 such that one of the k_j is zero and no adjacent two are equal. In addition an admissible labeling includes an assignment of the five marked points to vertices labeled with a zero. The vertex labeled with k_j is really labeled by q_{k_j}; the k–labels just avoid the double index notation and this is consistent with our use of k in the formulas.

Exercise 7.10 Compute the contributions to N_2 of several graphs from Figure 7.2. You can check your answer with the results listed in Appendix B. Notice that the sum of all the contributions is equal to one as we outlined in Exercise 4.8.

This last exercise may be difficult, but we feel that the interested reader will learn more by doing it than just reading the answer. We won't feel too guilty since the answer is in an appendix. We will provide more explanation for some of the contributions to N_3.

The graphs that can contribute to N_3 come in one of four different types. These types are displayed in Figure 7.2. Once again each graph type contains all of the graphs obtained by an admissible labeling as described in the N_2 case. This time an admissible labeling includes an assignment of the eight marked points to vertices labeled with a zero.

As an example of this notation the graph and curve displayed in Figure 7.1 are of type III.

Example 7.11 To be specific the graph from the figure is $III(0102)$ with the marked points p_4, p_5, p_6 and p_7 on the k_0 vertex. In this case the moduli stack of the fixed point component is given by

$$\overline{\mathcal{M}}_\Gamma = \overline{\mathcal{M}}_{0,5}(\mathrm{pt},0)\times\overline{\mathcal{M}}_{0,6}(\mathrm{pt},0),$$

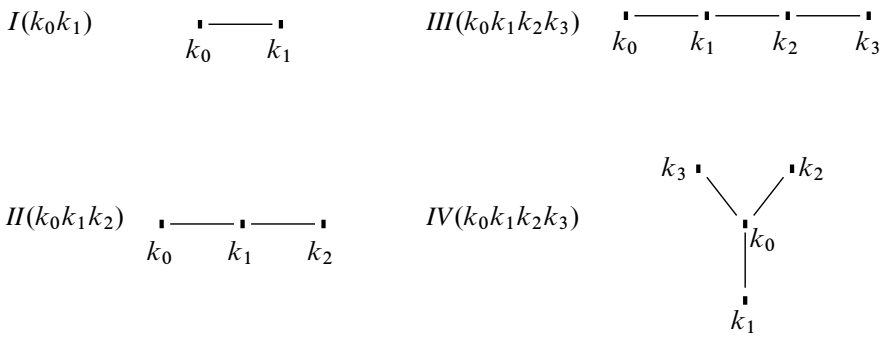


Figure 7.2: Graph types

and the inclusion map $\iota_\Gamma\colon \overline{\mathcal{M}}_\Gamma \to \overline{\mathcal{M}}_{0,8}(\mathbb{C}\mathrm{P}^2, 3[\mathbb{C}\mathrm{P}^1])$ is given by

$$\iota_\Gamma\big([\Sigma, p_1, p_2, p_3, p_4, p_5], [\Sigma', p_1', p_2', p_3', p_4', p_5', p_6']\big) = \Big[u, \big(\Sigma \cup \Sigma'\big) \bigcup_{\substack{p_1=([0:1],1)\\ p_1'=([1:0],2)\\ p_2'=([0:1],3)}} \big(\mathbb{C}\mathrm{P}^1 \times \{1,2,3\}\big)\big/ \sim, p_3', p_4', p_5', p_2, p_3, p_4, p_5, p_6'\Big],$$

where $([1:0],1) \sim ([0:1],2)$, $([1:0],2) \sim ([0:1],3)$ and the map u is given by

$$u([x_0:x_1],1) = [x_1:x_0:0], \qquad u([x_0:x_1],2) = [x_0:x_1:0],$$
$$u([x_0:x_1],3) = [x_1:0:x_0], \qquad u|_{(\Sigma\cup\Sigma')} = [1:0:0].$$

Applying formulas (8)–(14) to the labeled graph from Figure 7.1 gives

$$e(N_\Gamma)^{-1} = -\tfrac{1}{2}(\alpha_0-\alpha_1)^{-4}(\alpha_0-\alpha_2)^{-2}(\alpha_1-\alpha_2)^{-2} \int_{[\overline{\mathcal{M}}_{0,5}(\mathrm{pt},0)]^{\mathrm{vir}}} (\alpha_0-\alpha_1-\psi_1)^{-1} \cdot \int_{[\overline{\mathcal{M}}_{0,6}(\mathrm{pt},0)]^{\mathrm{vir}}} (\alpha_0-\alpha_1-\psi_1)^{-1}(\alpha_0-\alpha_2-\psi_2)^{-1}$$

Now consider the second integral that appears in the above expression. We compute

$$\begin{aligned}\int_{[\overline{\mathcal{M}}_{0,n}(\mathrm{pt},0)]^{\mathrm{vir}}} &(a_1-\psi_1)^{-1}(a_2-\psi_2)^{-1}\\ &= \int_{[\overline{\mathcal{M}}_{0,n}(\mathrm{pt},0)]^{\mathrm{vir}}} a_1^{-1}a_2^{-1}\left(\sum_{p=0}^{\infty} a_1^p\psi_1^p\right)\left(\sum_{q=0}^{\infty} a_2^q\psi_1^q\right)\\ &= a_1^{-1}a_2^{-1}\sum_{p+q=n-3} a_1^{-p}a_2^{-q}\int_{[\overline{\mathcal{M}}_{0,n}(\mathrm{pt},0)]^{\mathrm{vir}}}\psi_1^p\psi_2^q\\ &= a_1^{-1}a_2^{-1}\sum_{p+q=n-3}\binom{n-3}{p\ q}a_1^{-p}a_2^{-q} = a_1^{-1}a_2^{-1}(a_1^{-1}+a_2^{-1})^{n-3}.\end{aligned}$$

The last line in this computation used the result from Example 4.9.

The graph $III(0102)$ has no nontrivial automorphisms and each edge has degree one, so this component of the fixed point set has only trivial automorphisms. Combining the two previous computations with the $(h-\alpha_1)(h-\alpha_2)$ linearization allows us to conclude that the contribution of the graph $III(0,1,0,2)$ with the marked points p_4, p_5, p_6 and p_7 on the k_0 vertex is

$$-\tfrac{1}{2}(\alpha_0-\alpha_1)^{-4}(\alpha_0-\alpha_2)^5(\alpha_1-\alpha_2)^{-2}\big((\alpha_0-\alpha_1)^{-1}+(\alpha_0-\alpha_2)^{-1}\big)^3.$$

Of course there are $\binom{8}{4}$ ways to choose four marked points for the first vertex. This means that the contribution of such graphs is

$$-\frac{1}{2}\binom{8}{4}(\alpha_0-\alpha_1)^{-3}(\alpha_0-\alpha_2)^2(\alpha_1-\alpha_2)^{-2}(2\alpha_0-\alpha_1-\alpha_2)^3.$$

Exercise 7.12 Show that the contribution of the $III(0102)$ graphs with k marked points on the first vertex is

$$-\frac{1}{2}\binom{8}{k}(\alpha_0-\alpha_1)^{-3}(\alpha_0-\alpha_2)^{k-2}(\alpha_1-\alpha_2)^{-2}(2\alpha_0-\alpha_1-\alpha_2)^{7-k}.$$

Conclude that the total contribution to N_3 from all $III(0102)$ graphs is

$$-\tfrac{1}{2}(\alpha_0-\alpha_1)^{-3}(\alpha_0-\alpha_2)^{-2}(\alpha_1-\alpha_2)^{-2}(2\alpha_0-\alpha_1-\alpha_2)^{-1}(3\alpha_0-\alpha_1-2\alpha_2)^8.$$

Exercise 7.13 Explain why there is no graph corresponding to $II(002)$. Explain why there is no graph contribution corresponding to $I(21)$.

Exercise 7.14 Compute the contributions to N_3 of several graphs from Figure 7.2. Using some mathematical software it would be possible to push this computation all the way and get the answer $N_3 = 12$.

8 Derivation of the Euler class formulas

In this subsection we derive the general formulas that were used to compute the various Euler classes used in localization computations. We derive the formulas for $\mathbb{C}\mathrm{P}^2$; however careful inspection shows that these formulas are valid for $\mathbb{C}\mathrm{P}^m$ as well.

8.1 The Euler class of moving infinitesimal automorphisms

Before computing $e(\mathfrak{aut}([\Sigma, p])^{\mathrm{mov}})$ we need to understand the bundle of automorphisms. First consider the automorphisms of a genus zero irreducible component with no marked points mapped into $\mathbb{C}\mathrm{P}^2$ by the composition of a degree d map from $\mathbb{C}\mathrm{P}^1$ to $\mathbb{C}\mathrm{P}^1$ and an inclusion of $\mathbb{C}\mathrm{P}^1$ into $\mathbb{C}\mathrm{P}^2$. This is exactly the situation described in the following example. We assume that this component is attached to the rest of the stable map by a node at $[1:0]$.

Example 8.1 The T–action on $\mathfrak{def}(u)$ is a good example to study the induced T–action on a space in the deformation complex. Consider the T–fixed stable map given by $u([x_0:x_1]) = [x_0^d : x_1^d : 0]$. The action of T on this is just by pointwise multiplication of the coordinates. However, as we saw in Example 7.1 it is not immediately clear from this description why this map is fixed. We would like to include the reparametrization demonstrating that the class of this map is built into the definition of the action. The reparametrization requires a dth root of elements of T, but this can not be done in a consistent way. This may be corrected by considering the representation where T acts by multiplication of dth powers of the elements of T. The first Chern classes of the line bundles associated to the irreducible factors of this representation are d times the first Chern classes of the original representation. It follows that we can get the Chern classes of the line bundles in the original representation by dividing by d. The induced action on a 1–parameter family of maps (that is, a deformation) is then given by

$$(\lambda \cdot u_t)([x_0:x_1]) := \\ \big[\lambda_0^d u_t^0(\lambda_0^{-1}x_0, \lambda_1^{-1}x_1) : \lambda_1^d u_t^1(\lambda_0^{-1}x_0, \lambda_1^{-1}x_1) : \lambda_2^d u_t^2(\lambda_0^{-1}x_0, \lambda_1^{-1}x_1)\big].$$

Expressed in this way it is clear why this action fixes the map u.

One can now determine the T–action on $\mathfrak{aut}([\Sigma, p])$, and apply it to a specific bubble. Recall that the map $\mathfrak{aut}([\Sigma, p]) \to \mathfrak{def}(u)$ is given by $\varphi_t \mapsto u \circ \varphi_t$. Assuming that the domain of a stable map has a bubble with a node at $[1:0]$ and this bubble is mapped into $\mathbb{C}\mathrm{P}^2$ by a degree d map (as described in Exercise 8.3) any automorphism must restrict to this bubble to a map of the form $\varphi([x_0:x_1]) = [ax_0 + bx_1 : x_1]$. In order for

the map $\mathfrak{aut}([\Sigma, p]) \to \mathfrak{def}(u)$ to be equivariant with respect to the dth power action, we must have

$$(\lambda \cdot \varphi_t)([x_0 : x_1]) := [a_t x_0 + \lambda_0 \lambda_1^{-1} b_t x_1 : x_1] .$$

A nice way to compute the first Chern class of the corresponding subbundle is to use the Borel construction. Recall that one can combine a principal T–bundle with a T representation to get a vector bundle, as explained by the following exercise.

Exercise 8.2 Let $L_{n_0,\dots,n_1}$ be the line bundle associated to the representation $\lambda \cdot z = \lambda_0^{n_0} \dots \lambda_k^{n_k} z$ and show that

$$c_1(L_{n_0,\dots,n_1}) = n_0 c_1(L_{1,0,\dots,0}) + \dots + n_k c_1(L_{0,\dots,0,1}) .$$

We conclude that the first Chern class of the corresponding subbundle is $(\alpha_{[1:0:0]} - \alpha_{[0:1:0]})/d$, α_k is the Chern class of the bundle associated to the divisor ϕ_k described in the computation of the equivariant cohomology of $\mathbb{C}\mathrm{P}^n$.

A nice way to keep track of extra factors like $1/d$ involved in computing the Chern classes is to introduce the notion of a virtual representation. Virtual representations are just $\mathbb{Q}$–linear combinations of ordinary representations. The virtual representation on the automorphisms arising from the usual T–action is defined to be $\frac{1}{d}$ times the dth power representation.

We can split the automorphisms of a marked curve into automorphisms of each irreducible component of the marked curve and compute the Euler class of the bundle of infinitesimal automorphisms from the automorphisms of the components. In fact we can split each factor in (14) into a sum of line bundles. The irreducible representations of T are all one-dimensional. These representations induce line bundles over the components of the fixed point set of the moduli stack. The Euler class of a complex vector bundle is just the top Chern class of the bundle (see Bott and Tu [30]) and the Chern class of a sum is given by

$$c_{\mathrm{top}}(E \oplus F) = c_{\mathrm{top}}(E) c_{\mathrm{top}}(F).$$

To get to $e(\mathfrak{aut}([\Sigma, p])^{\mathrm{mov}})$ consider other irreducible components of the fixed stable map labeled by graphs as described around Figure 7.1. Each irreducible component of the stable map contributes a summand to the bundle $\mathfrak{aut}([\Sigma, p])^{\mathrm{mov}}$ and hence a factor to the Euler class. The factor of $e(\mathfrak{aut}([\Sigma, p])^{\mathrm{mov}})$ corresponding to an edge in the graph representing a component of the fixed point set that has two nodes or a node and a marked point is trivial. The induced action on the factors of $\mathfrak{aut}([\Sigma, p])$ corresponding to any contracted component (that is, one for which the stable map is

constant) is trivial. The above discussion implies that

$$e(\mathfrak{aut}([\Sigma, p])^{\mathrm{mov}}) = \prod_{\substack{\mathrm{val}(v)=1\\ n(v)=0\\ g(v)=0}} \frac{\alpha_{k(v)} - \alpha_{k(v')}}{d(e)}.$$

8.2 The T–action on the deformation complex

Before addressing the deformations of the underlying curve, we explain the T–action on all of the terms in the deformation complex in greater detail. We can use the homological algebra introduced to derive the deformation-obstruction complex to give a uniform treatment of the induced actions on the terms in the complex.

Given an action $T \times X \to X$ and a stable map fixed by this action, $u\colon \Sigma \to X$ one can construct the dth power action and the corresponding virtual T–actions on X and Σ making u equivariant as we did in Exercise 8.3.

Exercise 8.3 Let $u\colon \mathbb{C}\mathrm{P}^1 \cup_{[0:1]=[1:0]} \mathbb{C}\mathrm{P}^1 \to \mathbb{C}\mathrm{P}^2$ be given by

$$u([x_0 : x_1], 1) = [x_0^{d_1} : x_1^{d_1} : 0] \quad \text{and} \quad u([x_0 : x_1], 2) = [0 : x_0^{d_2} : x_1^{d_2}];$$

the original action on $\mathbb{C}\mathrm{P}^2$ is given by $\lambda \cdot [z_0 : z_1 : z_2] = [\lambda_0 z_0 : \lambda_1 z_1; \lambda_2 z_2]$; then the map u is equivariant with respect to the action on $\mathbb{C}\mathrm{P}^1$ given by

$$\lambda \cdot ([x_0 : x_1], 1) = ([\lambda_0^{d_2} x_0 : \lambda_1^{d_2} x_1], 1) \quad \text{and} \quad \lambda \cdot ([x_0 : x_1], 2) = ([\lambda_0^{d_1} x_0 : \lambda_1^{d_1} x_1], 2),$$

the $d_1 d_2$ power of the original action on $\mathbb{C}\mathrm{P}^2$.

Recall that the push-forward of a sheaf $\mathcal{A}$ under a map $f\colon X \to Y$ is given by $f_*\mathcal{A}(\mathcal{U}) := \mathcal{A}(f^{-1}(\mathcal{U}))$. Let L_λ represent left multiplication by λ. We will consider the push-forward by L_λ of various sheaves. There are natural transformations given by pull-back on the various sheaves,

$$\begin{aligned} L_\lambda^*\colon \Omega^1_X &\longrightarrow L_{\lambda *}\Omega^1_X \\ L_\lambda^*\colon u^*\Omega^1_X &\longrightarrow u^* L_{\lambda *}\Omega^1_X \\ L_\lambda^*\colon \mathcal{O}_\Sigma &\longrightarrow L_{\lambda *}\mathcal{O}_\Sigma \\ L_\lambda^*\colon \Omega^1_\Sigma([p]) &\longrightarrow L_{\lambda *}\Omega^1_\Sigma([p]) \end{aligned}$$

The map in the first line is just the pull-back of holomorphic 1–forms on X and the rest are analogous. The push-forward just formalizes the fact that the pull-back of a form

over U is a form over $L_\lambda^{-1}(U)$. Any of these pull-backs will induce a group action on the space of sections over any invariant set. In particular, they induce actions on the space of global sections. Our convention will be to use left group actions everywhere. Recall that a right action may be turned into a left action by taking the inverse of the group element. These actions will induce actions on the ext-groups. Recall that

$$\mathfrak{aut}([\Sigma, p]) = \mathbb{E}xt^0_{\mathcal{O}_\Sigma}(\Omega_\Sigma(p), \mathcal{O}_\Sigma) \cong \mathrm{Hom}_{\mathcal{O}_\Sigma}(\Omega_\Sigma(p), \mathcal{O}_\Sigma).$$

Given $X \in \mathrm{Hom}_{\mathcal{O}_\Sigma}(\Omega_\Sigma(p), \mathcal{O}_\Sigma)$, the left action is given by

$$(\lambda \cdot X)(\theta) := L^*_{\lambda^*}(X(L^*_\lambda \theta)).$$

Working in the $([x_0 : 1], 1)$–chart we obtain

$$(\lambda \cdot ((\dot{a}x_0 + \dot{b})\partial_x))(dx_0) = L^*_{\lambda^*}((\dot{a}x_0 + \dot{b})\partial_x(L^*_\lambda dx_0)) = \dot{a}x_0 + \lambda_0^{d_2}\lambda_1^{-d_2}\dot{b}.$$

Note that $\lambda \cdot ([x_0 : 1], 1) = ([\lambda_0^{d_2}\lambda_1^{-d_2}], 1)$ This agrees with our earlier computation of the action.

Exercise 8.4 Check that the action on $\mathfrak{def}([\Sigma, p])$ derived via homological algebra agrees with the action described earlier.

8.3 The Euler class of moving deformations of the curve

Following Cox and Katz [43], we can use local models to analyze the T–action on $\mathfrak{def}([\Sigma, p])^{\mathrm{mov}}$. Most marked points are on contracted components. The deformations corresponding to moving these points are tangent to the fixed point set of the action. The one exception is when there is exactly one marked point labeling a vertex of valence one. This corresponds to a marked point on one of the branch points of a degree d cover of a standard line in $\mathbb{C}\mathrm{P}^2$. The space of deformations corresponding to moving such a point is trivial because a genus zero curve with two marked points has no deformations. All of the normal deformations arise from resolutions of nodes. The local model here is $\Sigma = \{(x, y) \in \mathbb{C}^2 | xy = 0\}$. Let $\mathcal{I}_\Sigma = (xy)$ be the sheaf of algebraic functions on $\mathbb{C}^2$ containing a factor of xy (this is called the ideal sheaf) and notice that we have the following exact sequence,

$$0 \to \mathcal{I}_\Sigma/\mathcal{I}_\Sigma^2 \to \Omega^1_{\mathbb{C}^2}|_\Sigma \to \Omega^1_\Sigma \to 0.$$

The associated long exact sequence of ext-groups reads

$$\to \mathbb{E}\mathrm{xt}^0_{\mathcal{O}_\Sigma}(\Omega^1_{\mathbb{C}^2}|_\Sigma, \mathcal{O}_\Sigma) \to \mathbb{E}\mathrm{xt}^0_{\mathcal{O}_\Sigma}(\mathcal{I}_\Sigma/\mathcal{I}_\Sigma^2, \mathcal{O}_\Sigma) \to \mathbb{E}\mathrm{xt}^1_{\mathcal{O}_\Sigma}(\Omega^1_\Sigma, \mathcal{O}_\Sigma) \to \mathbb{E}xt^1_{\mathcal{O}_\Sigma}(\Omega^1_{\mathbb{C}^2}|_\Sigma, \mathcal{O}_\Sigma) \to \cdots.$$

Now, $\Omega^1_{\mathbb{C}^2}|_\Sigma$ is a free $\mathcal{O}_\Sigma$–module, so $\mathbb{E}\mathrm{xt}^1_{\mathcal{O}_\Sigma}(\Omega^1_{\mathbb{C}^2}|_\Sigma, \mathcal{O}_\Sigma) = 0$. It follows that

$$\begin{aligned}\mathbb{E}\mathrm{xt}^1_{\mathcal{O}_\Sigma}(\Omega^1_\Sigma, \mathcal{O}_\Sigma) &\cong \operatorname{coker}(\mathbb{E}\mathrm{xt}^0_{\mathcal{O}_\Sigma}(\Omega^1_{\mathbb{C}^2}|_\Sigma, \mathcal{O}_\Sigma) \to \mathbb{E}\mathrm{xt}^0_{\mathcal{O}_\Sigma}(\mathcal{I}_\sigma/\mathcal{I}^2_\Sigma, \mathcal{O}_\Sigma))\\ &\cong T_0(\mathbb{C}\times 0)\oplus T_0(0\times\mathbb{C})\,.\end{aligned}$$

These local results can be combined to give $\mathfrak{def}([\Sigma, p])^{\mathrm{mov}}$. We first introduce or recall some notation.

Let F refer to a flag in the graph of the stable map (that is, a pair consisting of a vertex and incident edge). We will use the flag to denote the vertex or the point on Σ or the point on $\mathbb{C}\mathrm{P}^2$ corresponding to the vertex without comment. The normalization of the surface is denoted by $\widehat{\Sigma}$ and the node branches will be denoted by b_1 and b_2. Recall that $\mathcal{L}_k$ denotes the line bundle over the moduli space whose fiber is the cotangent space of the corresponding curve.

Combining these local results gives,

$$\mathfrak{def}([\Sigma, p])^{\mathrm{mov}} = \Bigg(\bigoplus_{\substack{\mathrm{val}(v)=2\\ n(v)=0\\ g(v)=0}} \big(T_{b_1}\widehat{\Sigma}\oplus T_{b_2}\widehat{\Sigma}\big)\Bigg)\Bigg(\bigoplus_{\mathrm{val}(F)+n(F)+2g(v)>2} \big(T_F\widehat{\Sigma}\oplus\mathcal{L}_F^*\big)\Bigg).$$

This formula translates directly into a formula for the Euler class,

$$e(\mathfrak{def}([\Sigma, p])^{\mathrm{mov}}) = \Bigg(\prod_{\substack{\mathrm{val}(v)=2\\ n(v)=0\\ g(v)=0}} \Big(\frac{\alpha_{k(v)}-\alpha_{k(v_1)}}{d(e_1)} + \frac{\alpha_{k(v)}-\alpha_{k(v_2)}}{d(e_2)}\Big)\Bigg) \Bigg(\prod_{\mathrm{val}(F)+n(F)+2g(v)>2} \Big(\frac{\alpha_{k(F)}-\alpha_{k(v')}}{d(e)} - \psi_F\Big)\Bigg).$$

8.4 Euler class associated to the map

We now turn to the computation of the Euler classes of $\mathfrak{def}(u)^{\mathrm{mov}}$ and $\mathfrak{ob}(u)^{\mathrm{mov}}$. The following exact sequence serves to define holomorphic functions on a nodal curve in terms of holomorphic functions on the normalization $\nu\colon \widehat{\Sigma}\to\Sigma$:

$$0\to\mathcal{O}_\Sigma\to\nu_*\mathcal{O}_{\widehat{\Sigma}}\to\oplus_c\mathcal{O}_c\to 0.$$

Here we use c to denote the nodes (crossings) of Σ. We have a related exact sequence for holomorphic sections of the pull-back bundle $u^*T\mathbb{C}\mathrm{P}^2$:

$$0\to\mathcal{O}_\Sigma(u^*T\mathbb{C}\mathrm{P}^2)\to\nu_*\mathcal{O}_{\widehat{\Sigma}}(\nu^*u^*T\mathbb{C}\mathrm{P}^2)\to\oplus_c T_{u(c)}\mathbb{C}\mathrm{P}^2\to 0.$$

The associated cohomology long exact sequence reads,

$$\begin{aligned}0 \to H^0(\Sigma, \mathcal{O}_\Sigma(u^*T\mathbb{C}\mathrm{P}^2)) &\to H^0(\widehat{\Sigma}, \mathcal{O}_{\widehat{\Sigma}}(\nu^*u^*T\mathbb{C}\mathrm{P}^2)) \\ &\to \oplus_c H^0(\Sigma, T_{u(c)}\mathbb{C}\mathrm{P}^2) \to H^1(\Sigma, \mathcal{O}_\Sigma(u^*T\mathbb{C}\mathrm{P}^2)) \\ &\to H^1(\widehat{\Sigma}, \mathcal{O}_{\widehat{\Sigma}}(\nu^*u^*T\mathbb{C}\mathrm{P}^2)) \to \oplus_c H^1(\Sigma, T_{u(c)}\mathbb{C}\mathrm{P}^2) = 0.\end{aligned}$$

Notice that the fibers of $\mathfrak{def}(u)$ and $\mathfrak{ob}(u)$ over $[u, \Sigma, p]$ are $H^0(\Sigma, \mathcal{O}_\Sigma(u^*T\mathbb{C}\mathrm{P}^2))$ and $H^1(\Sigma, \mathcal{O}_\Sigma(u^*T\mathbb{C}\mathrm{P}^2))$ respectively. There are analogous sequences associated to any point $[u', \Sigma', p']$ in the moduli space and the spaces in these sequences glue together to define T–equivariant vector bundles over the moduli space. This is covered in more detail in Section 9 below. The sequence of linear spaces generalizes to a T–equivariant sequence of vector bundles. This implies that

$$\frac{e(\mathfrak{def}(u)^{\mathrm{mov}})}{e(\mathfrak{ob}(u)^{\mathrm{mov}})} = \frac{e(H^0(\widehat{\Sigma}, \mathcal{O}_{\widehat{\Sigma}}(\nu^*u^*T\mathbb{C}\mathrm{P}^2))^{\mathrm{mov}})}{e(H^1(\widehat{\Sigma}, \mathcal{O}_{\widehat{\Sigma}}(\nu^*u^*T\mathbb{C}\mathrm{P}^2))^{\mathrm{mov}}) \prod_c e(T_{u(c)}\mathbb{C}\mathrm{P}^2)}.$$

Recall that we are computing equivariant Euler classes here. It turns out that the possibility of a nontrivial bundle with nontrivial action does not arise in our computations. Thus, there are two different cases that we need to analyze. The first case is when the bundle is possibly nontrivial and the action is trivial. In this case the equivariant Euler class is just the usual Euler class. The second case is when the bundle is trivial and the action is nontrivial. In this case, the action on a vector space induces a bundle over $ET \times_T V$ and the equivariant Euler class is just the product of the Chern classes associated to the irreducible representations. Recall that α_k is the first Chern class of the bundle associated to the dual representation to the projection to the kth factor μ_k^*.

The standard action of T on $\mathbb{C}\mathrm{P}^2$ may be expressed as $\lambda \cdot (z_1, z_2) = (\lambda_1\lambda_0^* z_1, \lambda_2\lambda_0^* z_2)$ in the (z_1, z_2)–chart. It follows that

$$T_{[1:0:0]}\mathbb{C}\mathrm{P}^2 = \mu_0 \otimes (\mu_1^* \oplus \mu_2^*) \quad \text{and} \quad e(T_{[1:0:0]}\mathbb{C}\mathrm{P}^2) = (\alpha_0 - \alpha_1)(\alpha_0 - \alpha_2),$$

so

$$\prod_c e(T_{u(c)}\mathbb{C}\mathrm{P}^2)) = \prod_F \prod_{j \neq k(F)} (\alpha_{k(F)} - \alpha_j).$$

Remark 8.5 The number of nodes at a vertex with $n(v) = g(v) = 0$ is one less than the valence of this vertex, so when we take this product over all flags we should remove one flag from each such vertex. However, it is easier to include all flags here and include canceling terms in the expression for $e(H^0(\widehat{\Sigma}, \mathcal{O}(u^*T\mathbb{C}\mathrm{P}^2))^{\mathrm{mov}})$ (see Remark 8.7). We abuse notation by not introducing new notation for these modifications to $\prod_c e(T_{u(c)}\mathbb{C}\mathrm{P}^2))$ and $e(H^0(\widehat{\Sigma}, \mathcal{O}(u^*T\mathbb{C}\mathrm{P}^2))^{\mathrm{mov}})$.

The space $H^0(\widehat{\Sigma}, \mathcal{O}_{\widehat{\Sigma}}(v^*u^*T\mathbb{C}\mathrm{P}^2))^{\mathrm{mov}}$ splits into a sum of terms corresponding to the components of $\widehat{\Sigma}$. Clearly the bundle $u^*T\mathbb{C}\mathrm{P}^2$ is trivial over contracted components. Thus the contribution to the Euler class from the contracted components is

$$\prod_{\mathrm{val}(v)+n(v)>2} \prod_{j\neq k(v)} (\alpha_{k(v)} - \alpha_j).$$

To compute the contribution to the Euler class of the non-contracted components we will use the Euler sequence described in the next exercise.

Exercise 8.6 The subbundle of $T(\mathbb{C}^3-\{0\})$ generated by $z_k\partial_k|(z_0, z_1, z_2)$ is invariant under the multiplicative action of $\mathbb{C}^\times$. It therefore induces a bundle, say L, over the quotient $\mathbb{C}\mathrm{P}^2$. Check that the following is an exact sequence of vector bundles (it is called the Euler sequence):

$$0 \to L \to \underline{\mathbb{C}}^3 \to T\mathbb{C}\mathrm{P}^2 \to 0.$$

The Euler sequence leads to the following sequence of sheaves over $\mathbb{C}\mathrm{P}^2$:

$$0 \to \mathcal{O}_{\mathbb{C}\mathrm{P}^2} \to \mathcal{O}(1) \otimes \underline{\mathbb{C}}^3 \to T\mathbb{C}\mathrm{P}^2 \to 0$$

As a representative model of the non-contracted components, consider the map $u\colon \mathbb{C}\mathrm{P}^1 \to \mathbb{C}\mathrm{P}^2$ given by $u([x_0 : x_1]) = [x_0^d : x_1^d : 0]$. Pulling back to $\mathbb{C}\mathrm{P}^1$ via u and taking cohomology gives

$$\begin{aligned} 0 \to H^0(\mathbb{C}\mathrm{P}^1, \mathcal{O}_{\mathbb{C}\mathrm{P}^1}) \to H^0(\mathbb{C}\mathrm{P}^1, \mathcal{O}(d)) \otimes \underline{\mathbb{C}}^3 \\ \to H^0(\mathbb{C}\mathrm{P}^1, \mathcal{O}(u^*T\mathbb{C}\mathrm{P}^2)) \to H^1(\mathbb{C}\mathrm{P}^1, \mathcal{O}_{\mathbb{C}\mathrm{P}^1}) \to \cdots. \end{aligned}$$

Recall that $H^0(\mathbb{C}\mathrm{P}^1, \mathcal{O}(d))$ can be identified with the space of degree d homogeneous polynomials in x_0 and x_1. Each of the terms in the last sequence is T–equivariant. Recall that the T–action on $\mathbb{C}^3$ is given by $\lambda \cdot z = (\lambda_0 z_0, \lambda_1 z_1, \lambda_2 z_2)$ and the maps $[z_0 : z_1] \mapsto [z_0^{d-j} : z_1^j]$ generate all degree d meromorphic functions on $\mathbb{C}\mathrm{P}^1$. As a T–representation, we have

$$H^0(\mathbb{C}\mathrm{P}^1, \mathcal{O}(d)) \otimes \underline{\mathbb{C}}^3 = \left(\bigoplus_{j=0}^{d} \mu_0^{\otimes(d-j)/d} \otimes \mu_1^{\otimes j/d}\right) \otimes (\mu_0 \oplus \mu_1 \oplus \mu_2).$$

Thus taking equivariant Euler classes gives

$$\begin{aligned}e(H^0(\mathbb{C}\mathrm{P}^1,\mathcal{O}(u^*T\mathbb{C}\mathrm{P}^2))) &= \prod_{k=0}^{2}\prod_{j=0}^{d}\Big(\frac{d-j}{d}\alpha_0+\frac{j}{d}\alpha_1-\alpha_k\Big)\\ &= \prod_{j=1}^{d}\Big(\frac{j}{d}\alpha_1-\frac{j}{d}\alpha_0\Big)\prod_{j=0}^{d-1}\Big(\frac{d-j}{d}\alpha_0+\frac{j-d}{d}\alpha_1\Big)\prod_{j=0}^{d}\Big(\frac{d-j}{d}\alpha_0+\frac{j}{d}\alpha_1-\alpha_2\Big)\\ &= \frac{(-1)^d(d!)^2}{d^{2d}}(\alpha_0-\alpha_1)^{2d}\prod_{j=0}^{d}\Big(\frac{d-j}{d}\alpha_0+\frac{j}{d}\alpha_1-\alpha_2\Big).\end{aligned}$$

In general we have

$$\begin{aligned}e(H^0(\widehat{\Sigma},\mathcal{O}(u^*T\mathbb{C}\mathrm{P}^2))^{\mathrm{mov}}) &= \prod_{v}\prod_{j\neq k(v)}(\alpha_{k(v)}-\alpha_j)\\ &\prod_{e}\Bigg(\frac{(-1)^{d(e)}(d(e)!)^2}{d(e)^{2d(e)}}(\alpha_{k(v)}-\alpha_{k(v')})^{2d(e)}\prod_{\substack{a+b=d\\ j\neq k(v),k(v')}}\Big(\frac{a}{d(e)}\alpha_{k(v)}+\frac{b}{d(e)}\alpha_{k(v')}-\alpha_j\Big)\Bigg).\end{aligned}$$

Remark 8.7 The first product in this expression should be over all v such that $\mathrm{val}(v)+n(v)+2g(v)>2$ because these are the vertices with contracted components. However, taking the product over all vertices exactly cancels the extra terms introduced in $\prod_c e(T_{u(c)}\mathbb{C}\mathrm{P}^2))$ in Remark 8.5.

We have one remaining term to compute to finish our computation (10) of the equivariant Euler characteristic of the normal bundle to the fixed point set, namely

$$e(H^1(\widehat{\Sigma},\mathcal{O}(\nu^*u^*T\mathbb{C}\mathrm{P}^2))^{\mathrm{mov}}).$$

Notice that the Kodaira vanishing theorem implies that the cohomology corresponding to non-contracted components vanishes, giving

$$H^1(\widehat{\Sigma},\mathcal{O}(\nu^*u^*T\mathbb{C}\mathrm{P}^2))\cong\oplus_v H^1(\widehat{\Sigma}_v,\mathcal{O}(\nu^*u^*T\mathbb{C}\mathrm{P}^2)).$$

Now

$$\begin{aligned}&H^1(\widehat{\Sigma}_v,\mathcal{O}(\nu^*u^*T\mathbb{C}\mathrm{P}^2))\cong H^1(\widehat{\Sigma}_v,\mathcal{O}_{\Sigma_v}))\otimes T_{u(v)}\mathbb{C}\mathrm{P}^2\\ &\text{and}\qquad H^1(\widehat{\Sigma}_v,\mathcal{O}_{\Sigma_v}))\cong(H^0(\widehat{\Sigma}_v,\mathcal{O}_{\Sigma_v})))^\vee=\mathbb{E}^\vee.\end{aligned}$$

(Recall that $\mathbb{E}$ is the Hodge bundle.) Putting this together gives

$$\begin{aligned} H^1(\widehat{\Sigma}_v, \mathcal{O}(v^*u^*T\mathbb{C}\mathrm{P}^2)) &\cong \mathbb{E}^\vee \otimes T_{u(v)}\mathbb{C}\mathrm{P}^2 \\ &\cong \mathbb{E}^\vee \otimes (\mu_1 \otimes \mu_0^* \oplus \mu_2 \otimes \mu_0^*) \\ &\cong \mathbb{E}^\vee \otimes \mu_1 \otimes \mu_0^* \oplus \mathbb{E}^\vee \otimes \mu_2 \otimes \mu_0^*. \end{aligned}$$

Here we are assuming that $u(v) = [1:0:0]$.

Exercise 8.8 The splitting principle states that any formula for characteristic classes that is valid for sums of line bundles is valid for arbitrary bundles. Use the splitting principle to prove that

$$e(E \otimes L) = \sum_{i=0}^{r} c_i(E)c_1(L)^{r-i},$$

when E is a rank r vector bundle and L is a line bundle.

It follows that

$$e(H^1(\widehat{\Sigma}, \mathcal{O}(v^*u^*T\mathbb{C}\mathrm{P}^2))^{\mathrm{mov}}) = \prod_{\mathrm{val}(v)+n(v)+2g(v)>2} \prod_{j\neq k(v)} \sum_{i=0}^{g(v)} c_i(\mathbb{E}^\vee)(\alpha_{k(v)} - \alpha_j)^{g-i}.$$

Remark 8.9 We derived the formulas for the factors of the Euler class of the normal bundle to the components of the fixed point set for $\mathbb{C}\mathrm{P}^2$. However careful inspection shows that all of these formulas are valid for $\mathbb{C}\mathrm{P}^n$ without modification.

9 The virtual fundamental class

There is one new ingredient that arises in this computation. Up to now we have considered intersections in general position. It is still possible to do intersection theory when intersections are not generic. We now describe this non-generic intersection theory.

Let $\pi\colon E \to X$ be a vector bundle and let σ_0 be the zero section. Let F be a subbundle of E and $\sigma\colon X \to F$ be a generic section of F. We may also consider σ as a section of E, but it will not be transverse to the zero section. Let $Z := \sigma^{-1}(\sigma_0)$ and consider the following exact sequence of bundles,

$$0 \longrightarrow TZ \longrightarrow TX|_Z \xrightarrow{d\sigma} E|_Z \longrightarrow F^\perp|_Z \longrightarrow 0$$

Since E is a vector bundle, we have a natural map $E \to TE$. We also have a map $d\sigma_0\colon TX \to TE$. One can check that $TE|_{\sigma_0(X)} = d\sigma_0(TX) \oplus E|_{\sigma_0(X)}$. The map labeled by $d\sigma$ in the above sequence is the projection of the push-forward to $E|_{\sigma_0(X)}$. The bundle $F^\perp|_Z$ is called the obstruction bundle. It is usually denoted by $\mathfrak{ob}(Z)$. Notice that this agrees with our earlier description of the obstruction bundle. If the section σ were generic then the fundamental class of Z would be the Euler class of E. We define the virtual fundamental class of Z (denoted by $[Z]^{\text{vir}}$) to be the the Poincaré dual of the Euler class of E. We have

$$\begin{aligned}[Z]^{\text{vir}} &= \text{PD}(e(E)) = \text{PD}(e(F) \cup e(F^\perp)) \\ &= \text{PD}([Z] \cup e(F^\perp)) = [Z] \cap e(F^\perp) = [Z] \cap e(\mathfrak{ob}(Z)).\end{aligned}$$

We now consider an example of the virtual fundamental class.

Example 9.1 Consider the self-intersection of a line in $\mathbb{C}\text{P}^2$. The usual way to compute this is to perturb one copy of the line and then take the intersection, but this is not necessary. Let $L_1 = \mathbb{C}\text{P}^3 - \{[0:0:0:1]\}$ with projection $L_1 \to \mathbb{C}\text{P}^2$ be the Chern class 1 line bundle over $\mathbb{C}\text{P}^2$. Setting $\sigma_1\colon \mathbb{C}\text{P}^2 \to L_1$ to be the section $\sigma_1([x:y:z]) = [x:y:z:x]$, we see that $\sigma_1^{-1}(\sigma_0(\mathbb{C}\text{P}^2))$ is just a standard line. The self-intersection of this line is just the zeros of the section $\sigma = \sigma_1 \oplus \sigma_1$ of $L_1 \oplus L_1$. In this case it is easy to see that σ takes values in the diagonal L_1 subbundle and is transverse to the zero section of this subbundle. This is exactly the situation described above, so we have

$$\begin{aligned}\#(\mathbb{C}\text{P}^1 \cap \mathbb{C}\text{P}^1) = \int_{[\sigma^{-1}(0)]^{\text{vir}}} 1 &= \int_{[\mathbb{C}\text{P}^1] \cap e(\mathfrak{ob}(\mathbb{C}\text{P}^1))} 1 \\ &= \int_{\mathbb{C}\text{P}^1} e(\mathfrak{ob}(\mathbb{C}\text{P}^1)) = \int_{\mathbb{C}\text{P}^1} \Omega_{\mathbb{C}\text{P}^1} = 1.\end{aligned}$$

This same behavior happens in computations on many moduli spaces. Our main example is the large N dual of the 3–sphere.

Definition 9.2 The *local* P^1 or *small resolution of the conifold* is denoted by X_{S^3} or $\mathcal{O}(-1) \oplus \mathcal{O}(-1)$. It is the total space of the $\mathcal{O}(-1) \oplus \mathcal{O}(-1)$ complex vector bundle over $\mathbb{C}\text{P}^1$. We have,

$$X_{S^3} := \left\{ ([z_0:z_1], w_0, w_1, w_2, w_3) \in \mathbb{C}\text{P}^1 \times \mathbb{C}^4 \;\middle|\; \det\begin{pmatrix} w_0 & w_1 \\ z_0 & z_1 \end{pmatrix} = \det\begin{pmatrix} w_2 & w_3 \\ z_0 & z_1 \end{pmatrix} = 0 \right\}.$$

The restriction of the symplectic form on $\mathbb{C}\text{P}^1 \times \mathbb{C}^4$ is a symplectic form on X_{S^3}. Let $q_1\colon X_{S^3} \to \mathbb{C}\text{P}^1$ and $q_2\colon X_{S^3} \to \mathbb{C}^4$ be the projection maps. Clearly these maps

are holomorphic. Given a holomorphic map $u\colon \Sigma \to X_{S^3}$ we see that $q_2 \circ u$ is holomorphic, therefore constant (see Exercise 9.3.) Let (w_1, w_2, w_3, w_4) be this constant value. For positive degree d the map $q_1 \circ u$ must be surjective. Taking $x_0 \in \Sigma$ with $q_1 \circ u(x_0) = [0:1]$ we see that $w_0 = w_2 = 0$. Taking $x_\infty \in \Sigma$ with $q_1 \circ u(x_\infty) = [1:0]$ we see that $w_1 = w_3 = 0$.

Exercise 9.3 Recall the definition of the $\overline{\partial}$ operator from Section 3. Show that ∂u and $\overline{\partial} u$ are perpendicular, and that

$$(|\partial u|^2 - |\overline{\partial} u|^2)\, d\mathrm{vol}_\Sigma = -g(du, J \circ du \circ j))\, d\mathrm{vol}_\Sigma = 2u^*\omega_X .$$

Conclude that

$$\frac{1}{2}\int_\Sigma |du|^2\, d\mathrm{vol}_\Sigma = \int_\Sigma |\overline{\partial} u|^2\, d\mathrm{vol}_\Sigma + \int_\Sigma u^*\omega_X .$$

When u is holomorphic, Σ is closed and $u^*\omega_X$ is exact this implies that u is constant.

Let $\sigma_0\colon \mathbb{C}\mathrm{P}^1 \to X_{S^3}$ be the zero section. Our previous discussion implies that the induced map

$$\sigma_0\colon \overline{\mathcal{M}}_{g,0}(\mathbb{C}\mathrm{P}^1, d[\mathbb{C}\mathrm{P}^1]) \to \overline{\mathcal{M}}_{g,0}(X_{S^3}, d[\mathbb{C}\mathrm{P}^1])$$

is an isomorphism of stacks. According to Exercise 5.25,

$$\mathrm{virdim}(\overline{\mathcal{M}}_{g,0}(X_{S^3}, d[\mathbb{C}\mathrm{P}^1])) = 0,$$

and

$$\mathrm{virdim}(\overline{\mathcal{M}}_{g,0}(\mathbb{C}\mathrm{P}^1, d[\mathbb{C}\mathrm{P}^1])) = 2g - 2 + 2d\,.$$

We conclude that if the moduli space $\overline{\mathcal{M}}_{g,0}(\mathbb{C}\mathrm{P}^1, d[\mathbb{C}\mathrm{P}^1])$ is unobstructed, then we are in exactly the situation described by the excess intersection formula. In Section 5, we claimed that homological algebra would allow one to extend the spaces from the deformation-obstruction complex to bundles over the moduli space. This is similar to the situation encountered earlier of glueing the spaces $T^*_{p_k}\Sigma$ into the bundle $\mathcal{L}_k$ over the moduli space. We use the universal curve over the moduli space here as we did before. This time we need a construction from homological algebra called higher direct image functors.

Given a left exact functor $F\colon \mathcal{C} \Longrightarrow \mathcal{D}$ and an injective resolution $A \to I^0 \to I^1 \to \cdots$ one defines the right derived functors of F applied to A (denoted $R^*(F)(A)$) to be the cohomology of the complex $F(I^0) \to F(I^1) \to \cdots$. Given a map $f\colon X \to Y$ and a sheaf $\mathcal{A}$ over X one defines the direct image sheaf $f_*\mathcal{A}$ over Y by $f_*\mathcal{A}(V) := \mathcal{A}(f^{-1}(V))$. The *higher direct image functors* are just the right derived functors of the direct image functor. *Hyper-derived functors* are a generalization of derived functors

applicable to complexes of sheaves. One takes injective resolutions of each sheaf in the complex and applies the functor to obtain a double complex. The hyper-derived functor is just the cohomology of the corresponding total complex. *Hyper-higher direct image functors* $\mathbf{R}^*(f_*)$ of complexes may be defined via the total space of the resulting complexes in the same way one defines the hyper-ext functors.

We will apply the hyper-higher direct image functor to sheaves arising from vector bundles. To any vector bundle we can associate the sheaf of sections. The sheaf of sections of a holomorphic vector bundle over X is a locally free, finite rank sheaf of $\mathcal{O}_X$–modules. Conversely, given a locally free, finite rank sheaf $\mathcal{E}$ of $\mathcal{O}_X$–modules one defines an associated holomorphic vector bundle. Namely, there is an open cover $\{U_\alpha\}$ of X and isomorphisms $\varphi_\alpha\colon \mathcal{O}_X(U_\alpha)^n \to \mathcal{E}(U_\alpha)$. Define $\psi_{\alpha\beta} := \varphi_\alpha^{-1} \circ \varphi_\beta$ on the overlaps and $E := \amalg\, U_\alpha \times \mathbb{C}^n/\sim$, where $(\alpha, x, z) \sim (\beta, x, \psi_{\beta\alpha}(x)z)$.

Exercise 9.4 Show that $\Gamma(E)$ is naturally isomorphic to $\mathcal{E}$. Show that

$$\mathcal{O}_X^n \otimes_{\mathcal{O}_X} \left(\mathcal{O}_{x_0}/\mathfrak{m}_{x_0}\right) \cong \mathbb{C}^n.$$

We conclude from this exercise that the fiber of E over a point x_0 may be identified with $\mathcal{E} \otimes_{\mathcal{O}_X} \left(\mathcal{O}_{x_0}/\mathfrak{m}_{x_0}\right)$.

Now consider the universal curve over the moduli space. We have the vertical bundle

$$\mathcal{V} \longrightarrow \mathcal{U}_X \xrightarrow{\pi_X} \overline{\mathcal{M}}_{g,n}(X, \beta)$$

with sections $\rho_k\colon \overline{\mathcal{M}}_{g,n}(X, \beta) \to \mathcal{U}_X$ and the diagram

$$\begin{array}{ccc} \mathrm{ev}_X^* TX & \longrightarrow & TX \\ \downarrow & & \downarrow \\ \mathcal{U}_X & \xrightarrow[\mathrm{ev}_X]{} & X \\ \downarrow{\scriptstyle \pi_X} & & \\ \overline{\mathcal{M}}_{g,n}(X, \beta) & & \end{array}$$

Let $\mathcal{L}_{-\rho}$ be the line bundle associated to the divisor

$$-\rho_1(\overline{\mathcal{M}}_{g,n}(X, \beta)) - \cdots - \rho_n(\overline{\mathcal{M}}_{g,n}(X, \beta)).$$

Let s be a section of this bundle with simple zeros along the divisor. Define $\mathcal{V}_{-\rho}$ to be the bundle $\mathcal{V} \otimes \mathcal{L}_{-\rho}$ and a bundle map $\mathrm{ev}_*(- \otimes s)\colon \mathcal{V}_{-\rho} \to \mathrm{ev}_X^* TX$.

Definition 9.5 The bundles in the deformation-obstruction sequence of a stable map are naturally associated to the sheaves given by:

$$\begin{aligned}
\mathfrak{def}(u) &:= \mathbf{R}^0(\pi_*)(\mathrm{ev}^* TX)\\
\mathfrak{ob}(u) &:= \mathbf{R}^1(\pi_*)(\mathrm{ev}^* TX)\\
\mathfrak{aut}([\Sigma, p]) &:= \mathbf{R}^0(\pi_*)(\mathcal{V}_{-\rho})\\
\mathfrak{def}([\Sigma, p]) &:= \mathbf{R}^1(\pi_*)(\mathcal{V}_{-\rho})\\
\mathfrak{aut}([u, \Sigma, p]) &:= \mathbf{R}^0(\pi_*)(\mathcal{V}_{-\rho} \to \mathrm{ev}^* TX)\\
\mathfrak{def}([u, \Sigma, p]) &:= \mathbf{R}^1(\pi_*)(\mathcal{V}_{-\rho} \to \mathrm{ev}^* TX)\\
\mathfrak{ob}([u, \Sigma, p]) &:= \mathbf{R}^2(\pi_*)(\mathcal{V}_{-\rho} \to \mathrm{ev}^* TX)\,.
\end{aligned}$$

Before explaining why these bundles over the moduli stack have the required fibers we should explain exactly what the various maps actually are in this setting. In the above description we just treated stacks as spaces. The easiest way to understand a stack at this point is to use the definition that the moduli stack is the contravariant functor from the category of schemes to sets that associates the set of all equivalence classes of families of stable maps over a scheme to a scheme.

> In short, a stack is just the collection of all families of stable maps over a scheme.

Several observations will clarify a correct way to think about these constructions. The first observation is that the coarse moduli space is the set that the moduli stack associates to a point, that is, $\overline{\mathcal{M}}_{g,n}(X,\beta)(\mathrm{pt}) = \overline{M}_{g,n}(X,\beta)$. The second observation is that the universal curve over $\overline{\mathcal{M}}_{g,n}(X,\beta)$ is just $\mathcal{U}_X = \overline{\mathcal{M}}_{g,n+1}(X,\beta)$ even when the stable maps have nontrivial automorphisms. The map

$$\pi_X\colon \mathcal{U}_X \to \overline{\mathcal{M}}_{g,n}(X,\beta)$$

is just the natural transformation of functors that takes a family of $(n+1)$–pointed curves to the family of n–pointed curves obtained by ignoring the last point (section) and stabilizing the fibers. The third observation is that the stack associated to a space just consists of all maps from schemes into the space, and the map $\mathrm{ev}_X\colon \mathcal{U}_X \to X$ is just the natural transformation that takes a family $[w\colon \mathcal{W} \to X, \mathcal{W} \to S, \rho]$ to the map $w \circ \rho_{n+1} \to X$. The example in Section 10 will continue with the idea that the moduli stack is the collection of families of stable maps. The following exercise is good practice translating constructions into families.

Exercise 9.6 Give a definition of the vertical bundle

$$\mathcal{V} \longrightarrow \mathcal{U}_X \xrightarrow{\pi_X} \overline{\mathcal{M}}_{g,n}(X,\beta)$$

by making $\mathcal{V}$ a stack that associates appropriate families of vector bundles over families of stable maps.

We now describe why the fibers of the bundles in the deformation-obstruction sequence are the expected spaces. To apply Exercise 9.4 to the sheaves of the deformation-obstruction complex, we need a theorem of Grauert (see Hartshorne [75, page 33]). Recall the definition of flat morphism from Section 5.2.

Theorem 9.7 *If $f\colon X \to Y$ is a flat morphism, $\mathcal{F}$ is a coherent sheaf over Y and $\dim_{\mathcal{O}_y/\mathfrak{m}_y} H^k(X_y,\mathcal{F}_y)$ is constant, then $R^k(f_*)(\mathcal{F})$ is locally free of finite rank and*

$$R^k(f_*)(\mathcal{F}) \otimes_{\mathcal{O}_Y} \left(\mathcal{O}_y/\mathfrak{m}_y\right) \cong H^k(X_y,\mathcal{F}_y)\,,$$

where X_y is the fiber over y and $\mathcal{F}_y = \mathcal{F}|_{X_y}$.

> This theorem gives us a good way to think about higher direct image functors – under nice conditions the higher direct image functors associate to a family of spaces over S a vector bundle over S with fiber equal to the cohomology of the fiber in the original family.

Notice that the condition that $H^k(X_y,\mathcal{F}_y)$ have constant dimension is not satisfied in cases of interest to us. For example, a degree three map from a surface of genus three to X_{S^3} has no nontrivial infinitesimal automorphisms of the underlying surface. However there is a nodal surface in the same moduli space consisting of the one point union of a surface of genus two and a surface of genus zero mapped by degree two on the genus two part and degree one on the genus zero part. This nodal surface has a two-complex-dimensional space of infinitesimal automorphisms. The next exercise compares the virtual fibers of the bundles from the deformation-obstruction bundle complex with the spaces from the deformation-obstruction complex.

Exercise 9.8 Given an injective resolution $\mathcal{O}_\Sigma \to I^0 \to I^1 \to \cdots$, show that $u^*TX \to \operatorname{Hom}_{\mathcal{O}_\Sigma}(u^*\Omega_X, I^0) \to \cdots$ is an injective resolution of u^*TX. Conclude that

$$R^1(\pi_*)(\mathrm{ev}^*TX) \otimes_{\mathcal{O}/\mathfrak{m}} (\mathcal{O}/\mathfrak{m}) \cong H^1(\Sigma, u^*TX) = \mathbb{E}xt^1(u^*\Omega_X, \mathcal{O}_\Sigma)\,.$$

Repeat this computation with the other bundles.

Assume for the moment that the fiber

$$\mathbb{E}\mathrm{xt}^2(u^*\Omega_X, \mathcal{O}_\Sigma) \cong \mathbb{H}^2(\Sigma, T\Sigma \to u^*TX)$$

has constant dimension as $[u, \Sigma]$ varies in $\overline{\mathcal{M}}_{g,0}(X_{S^3}, d[\mathbb{C}\mathrm{P}^1])$. (Here $\mathbb{H}^k$ represents hypercohomology.) Then $\mathfrak{ob}([u, \Sigma]) = \mathbf{R}^2(\pi_*)(\mathcal{V} \to \mathrm{ev}^*TX)$ satisfies the assumptions of Grauert's theorem and therefore satisfies the assumptions required for our definition of the virtual fundamental class. Using the map $\sigma_0\colon \overline{\mathcal{M}}_{g,0}(\mathbb{C}\mathrm{P}^1, d[\mathbb{C}\mathrm{P}^1]) \to \overline{\mathcal{M}}_{g,0}(X_{S^3}, d[\mathbb{C}\mathrm{P}^1])$, we write

$$\begin{aligned}
N_{g,d} := \langle 1\rangle^X_{g,d[\mathbb{C}\mathrm{P}^1]} &= \int_{[\overline{\mathcal{M}}_{g,0}(X_{S^3}, d[\mathbb{C}\mathrm{P}^1])]^{\mathrm{vir}}} 1 \\
&= \int_{[\overline{\mathcal{M}}_{g,0}(X_{S^3}, d[\mathbb{C}\mathrm{P}^1])]} e(\mathbf{R}^2(\pi_*)(\mathcal{V} \to \mathrm{ev}^*TX)) \\
&= \int_{[\overline{\mathcal{M}}_{g,0}(\mathbb{C}\mathrm{P}^1, d[\mathbb{C}\mathrm{P}^1])]} \sigma_0^* e(\mathbf{R}^2(\pi_*)(\mathcal{V} \to \mathrm{ev}^*TX)) \\
&= \int_{[\overline{\mathcal{M}}_{g,0}(\mathbb{C}\mathrm{P}^\cdot d[\mathbb{C}\mathrm{P}^1])]} e(\sigma_0^*\mathbf{R}^2(\pi_*)(\mathcal{V} \to \mathrm{ev}^*TX)) \,.
\end{aligned}$$

We continue with the change of base theorem for higher direct image functors. See Hartshorne [75, page 255].

Theorem 9.9 *Given a commutative diagram,*

$$\begin{array}{ccc} W & \xrightarrow{v} & X \\ {\scriptstyle g}\downarrow & & \downarrow{\scriptstyle f} \\ Z & \xrightarrow[u]{} & Y \end{array}$$

with u flat and a complex of coherent sheaves $\mathcal{A}$ one has

$$u^*\mathbf{R}^*(f_*)(\mathcal{A}) = \mathbf{R}^*(g_*)(v^*\mathcal{A}) \,.$$

We conclude that

$$\sigma_0^*\mathbf{R}^2(\pi_*^X)(\mathcal{V} \to \mathrm{ev}_X^*TX) = \mathbf{R}^2(\pi_*^{\mathbb{C}\mathrm{P}^1})(\hat{\sigma}_0^*\mathcal{V} \to \hat{\sigma}_0^*\mathrm{ev}_X^*TX) \,.$$

Here $\hat{\sigma}_0\colon \mathcal{U}_{\mathbb{C}\mathrm{P}^1} \to \mathcal{U}_X$ is the natural map. Since $\sigma_0 \circ \mathrm{ev}_{\mathbb{C}\mathrm{P}^1} = \mathrm{ev}_X \circ \hat{\sigma}_0$, we know

$$\mathbf{R}^2(\pi_*^{\mathbb{C}\mathrm{P}^1})(\hat{\sigma}_0^*\mathcal{V} \to \hat{\sigma}_0^*\mathrm{ev}_X^*TX) = \mathbf{R}^2(\pi_*^{\mathbb{C}\mathrm{P}^1})(\hat{\sigma}_0^*\mathcal{V} \to \mathrm{ev}_{\mathbb{C}\mathrm{P}^1}^*\sigma_0^*TX).$$

To go further, consider the exact sequence of bundles,

$$0 \to \mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1} T\mathbb{C}\mathrm{P}^1 \to \mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1} \sigma_0^* TX \to \mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}(\mathcal{O}(-1)\oplus\mathcal{O}(-1)) \to 0\,.$$

By the horseshoe lemma from homological algebra there are injective resolutions $\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1} T\mathbb{C}\mathrm{P}^1 \to I^*$, $\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}\sigma_0^* TX \to J^*$ and $\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}(\mathcal{O}(-1)\oplus\mathcal{O}(-1))$ that form a short exact sequence of complexes

$$0 \longrightarrow I^* \longrightarrow J^* \longrightarrow K^* \longrightarrow 0.$$

Exercise 9.10 Let $0 \to A \to B \to C \to 0$ be a short exact sequence of modules, and let $A \to I^*$ and $C \to K^*$ be injective resolutions. Show that there are maps $\varepsilon\colon B \to I^0 \oplus K^0$ and $d^k\colon I^k \oplus K^k \to I^{k+1} \oplus K^{k+1}$ such that $B \to I^* \oplus K^*$ is an injective resolution that fits into a short exact sequence of complexes (see Weibel [156]).

Let $\mathcal{V} \to L^*$ be an injective resolution and form the following short exact sequence of complexes,

$$\begin{array}{ccccccccc}
0 & \longrightarrow & \boldsymbol{L^0} & \longrightarrow & \boldsymbol{L^0} & \longrightarrow & \boldsymbol{0} & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & \boldsymbol{L^1 \oplus I^0} & \longrightarrow & \boldsymbol{L^1 \oplus J^0} & \longrightarrow & \boldsymbol{K^0} & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & \boldsymbol{I^1} & \longrightarrow & \boldsymbol{J^1} & \longrightarrow & \boldsymbol{K^1} & \longrightarrow & 0 \\
 & & & & & & \downarrow & & \\
 & & & & & & 0 & &
\end{array}$$

Applying the direct image functor $\pi_*^{\mathbb{C}\mathrm{P}^1}$ to each term and writing out the associated long exact sequence of cohomology groups produces a long exact sequence containing

$$\begin{aligned}
&\to \mathbf{R}^0(\pi_*)\big(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}(\mathcal{O}(-1)\oplus\mathcal{O}(-1))\big) \to \mathbf{R}^2(\pi_*)\big(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}\mathcal{V} \to \mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1} T\mathbb{C}\mathrm{P}^1\big) \\
&\qquad \to \mathbf{R}^2(\pi_*)\big(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}\mathcal{V} \to \mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}\sigma_0^* TX\big) \to \mathbf{R}^1(\pi_*)\big(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}(\mathcal{O}(-1)\oplus\mathcal{O}(-1))\big) \to 0.
\end{aligned}$$

Now look at the deformation-obstruction sequence for $\overline{\mathcal{M}}_{g,0}(\mathbb{C}\mathrm{P}^1, d[\mathbb{C}\mathrm{P}^1])$,

$$\begin{aligned}
0 \longrightarrow \mathfrak{aut}([v,\Sigma]) \longrightarrow \mathfrak{aut}([\Sigma]) \longrightarrow \mathfrak{def}(v) \longrightarrow \mathfrak{def}([v,\Sigma]) \longrightarrow \mathfrak{def}([\Sigma]) \\
\longrightarrow \mathfrak{ob}(v) \longrightarrow \mathfrak{ob}([v,\Sigma]) \longrightarrow 0.
\end{aligned}$$

The fiber of $\mathfrak{ob}(v)$ is $H^1(\Sigma, v^*T\mathbb{C}\mathrm{P}^1)$. For every stable map in the genus zero case this cohomology group is zero (this property is called convexity). For $\mathbb{C}\mathrm{P}^1$ convexity follows from the Kodaira vanishing theorem (see Griffiths and Harris [68]). This implies that

$$0 = \mathfrak{ob}([v, \Sigma]) = \mathbf{R}^2(\pi_*)\big(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}\mathcal{V} \to \mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}T\mathbb{C}\mathrm{P}^1\big),$$

so that

$$\mathbf{R}^2(\pi_*)\big(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}\mathcal{V} \to \mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}\sigma_0^*TX\big) \cong \mathbf{R}^1(\pi_*)\big(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}(\mathcal{O}(-1)\oplus\mathcal{O}(-1))\big),$$

by the previous exact sequence. Now,

$$\dim_{\mathbb{C}} H^1(\Sigma, v^*(\mathcal{O}(-1)\oplus\mathcal{O}(-1))) = 2d + 2g - 2$$

by the Riemann–Roch theorem together with the Kodaira vanishing theorem. Since this is constant, we conclude that the excess intersection formula holds and

$$N_{g,d} = \int_{[\overline{\mathcal{M}}_{g,0}(\mathbb{C}\mathrm{P}^1, d[\mathbb{C}\mathrm{P}^1])]} c_{2d+2g-2}\big(\mathbf{R}^1(\pi_*)\big(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}(\mathcal{O}(-1)\oplus\mathcal{O}(-1))\big)\big).$$

The argument we used to get here was valid for the $g = 0$ case. In fact, one can show that this formula is valid for all genera. When the dimension of the virtual fibers of $\mathfrak{ob}([u, \Sigma, p])$ is not constant, one can still define a virtual fundamental class. This was done independently and almost simultaneously in the spring of 1996 by Fukaya–Ono, Hofer–Salamon, Li–Tian, Ruan and Siebert, see Salamon's lectures [133],5 and references therein. All of the authors originally worked in a symplectic setting. Our description is closest to Li–Tian [97] who later extended their results to an algebraic setting [96]. See also Behrend [29], Cox and Katz [43] and Liu [98].

Virtual fundamental classes are important in the computation of degree zero Gromov–Witten invariants as well. The next exercises address this situation.

Exercise 9.11 Show that the obvious map $\sigma_0\colon \overline{\mathcal{M}}_{g,n} \times X \to \overline{\mathcal{M}}_{g,n}(X, 0)$ is an isomorphism of stacks.

Exercise 9.12 Let $\pi_M\colon \overline{\mathcal{M}}_{g,0}\times X \to \overline{\mathcal{M}}_{g,n}$ and $\pi_X\colon \overline{\mathcal{M}}_{g,0}\times X \to X$ be the projection maps and show that

$$\int_{[\overline{\mathcal{M}}_{g,n}(X,0)]^{\mathrm{vir}}} \gamma = \int_{[\overline{\mathcal{M}}_{g,n}]^{\mathrm{vir}}\times X} c_{\mathrm{top}}(\pi_M^*\mathbb{E}^\vee \otimes \pi_X^*TX) \cup \sigma_0^*\gamma\,.$$

Exercise 9.13 Use the splitting principle to show that $\dim_{\mathbb{C}} X = 3$ and $c_1(TX) = 0$ imply

$$N_{g,0}(X) := \int_{[\overline{\mathcal{M}}_{g,0}(X,0)]^{\mathrm{vir}}} 1 = (-1)^g\chi(X)\int_{[\overline{\mathcal{M}}_{g,0}]^{\mathrm{vir}}} c_{g-1}(\mathbb{E})^3/2\,.$$

10 The multiple cover formula in degree two

In this subsection we describe how to compute the Gromov–Witten invariants of the manifold X_{S^3}. We begin with a direct computation of

$$N_{0,2}(X) := \langle 1 \rangle^X_{0,2[\mathbb{C}\mathrm{P}^1]}.$$

A localization computation of this same number is presented in Cox and Katz [43]. We begin by analyzing the corresponding coarse moduli space.

Recall from the previous subsection that any stable map into X_{S^3} factors through $\mathbb{C}\mathrm{P}^1$, so the moduli stack of stable maps to X_{S^3} is isomorphic to the moduli stack of stable maps to $\mathbb{C}\mathrm{P}^1$. Given a stable map, one can construct a graph with vertices corresponding to the maximal contracted components, edges corresponding to the non-contracted components, and labels corresponding to the marked points, images of the contracted components, genera of the contracted components, and degrees of the non-contracted components. Since the curves in our case have genus zero, the corresponding graphs must be trees. Computing the Euler characteristic of the resulting graph gives $1 = \sum_v (1 - \frac{1}{2}\mathrm{valence}(v))$. It follows that we must either have a vertex of valence zero or two vertices of valence one. In order to be stable the map must have positive degree on each vertex with valence less than three. It follows that the only stable curves in the genus zero degree two moduli space have domain $\mathbb{C}\mathrm{P}^1$ or two copies of $\mathbb{C}\mathrm{P}^1$ joined by a single node.

Now consider a degree two holomorphic map $u\colon \mathbb{C}\mathrm{P}^1 \to \mathbb{C}\mathrm{P}^1$. Locally such a map has a power series representation, so it must be a branched cover. This implies that

$$2 = \chi(\mathbb{C}\mathrm{P}^1) = 2\chi(\mathbb{C}\mathrm{P}^1) - \sum_{y \in S(u)} (2 - |u^{-1}(y)|) = 4 - |S(u)|\,.$$

It follows that such a map must have exactly two critical points and two critical values. We will use the critical values to parametrize these maps. Notice that pre-composition with a linear fractional transformation cannot change the locations of the critical values, so two maps with different critical values are different. Given any two distinct points p, q in $\mathbb{C}\mathrm{P}^1$ we can take a linear fractional transformation taking $[0:1]$ to p and $[1:0]$ to q. Composing the map $[z:w] \mapsto [z^2:w^2]$ with this linear fractional transformation gives a degree two map with the desired critical values. Notice that this is not an equivalence of stable maps because such equivalences must be by pre-composition. We will map a nodal curve to the image of the node.

Exercise 10.1 Show that two stable maps with the same critical values are equivalent, and show that the automorphism group of any stable curve in this space is $\mathbb{Z}_2$. Conclude

that the coarse moduli space is isomorphic to the symmetric product of two copies of $\mathbb{C}\mathrm{P}^1$. (This is a reasonable example to use to understand the Gromov topology, or the algebraic structure of the coarse moduli space.)

The second symmetric power of $\mathbb{C}\mathrm{P}^1$ may be identified with the space of degree two polynomials up to scale. One associates the roots of the polynomial to the polynomial. This gives the isomorphism $\mathrm{Sym}^2\mathbb{C}\mathrm{P}^1 \to \mathbb{C}\mathrm{P}^2$ taking $([z_0 : z_1], [w_0 : w_1])$ to $[z_1 w_1 : -z_0 w_1 - z_1 w_0 : z_0 w_0]$. The same map is an explicit isomorphism

$$\Phi\colon \overline{M}_{0,0}(\mathbb{C}\mathrm{P}^1, 2[\mathbb{C}\mathrm{P}^1]) \cong \mathbb{C}\mathrm{P}^2,$$

when the singular set of the map is $([z_0 : z_1], [w_0 : w_1])$. The boundary divisor is just $D = \{[a : b : c] \in \mathbb{C}\mathrm{P}^2 | b^2 - 4ac = 0\}$.

At the level of stacks the universal curve is just the moduli space $\overline{\mathcal{M}}_{0,1}(\mathbb{C}\mathrm{P}^1, 2[\mathbb{C}\mathrm{P}^1])$ with evaluation as the map to $\mathbb{C}\mathrm{P}^1$ and projection as the map to $\overline{\mathcal{M}}_{0,0}(\mathbb{C}\mathrm{P}^1, 2[\mathbb{C}\mathrm{P}^1])$. It is interesting to compute the corresponding coarse moduli space. We claim that $\overline{M}_{0,1}(\mathbb{C}\mathrm{P}^1, 2[\mathbb{C}\mathrm{P}^1])$ is isomorphic to $\mathbb{C}\mathrm{P}^2 \times \mathbb{C}\mathrm{P}^1$ with isomorphism taking $[u, \Sigma, p]$ to $(\Phi(u), u(p))$. Given a point in $\mathbb{C}\mathrm{P}^2 \times \mathbb{C}\mathrm{P}^1$ one can take a stable map corresponding to the $\mathbb{C}\mathrm{P}^2$–component and then pick a point in the inverse image of the $\mathbb{C}\mathrm{P}^1$–component as the marked point. Such a point exists because the map has degree two. (If the inverse image is a node, we add a ghost bubble containing the marked point at the node.) If the $\mathbb{C}\mathrm{P}^1$–component is one of the critical values there is a unique choice for the marked point. Otherwise there are two possibilities. However one finds that the resulting marked stable curves are equivalent. For example, the point $[0 : 1 : 0], [1 : 1]$ gives a stable map with critical points $[0 : 1]$ and $[1 : 0]$ (the roots of $bz_0z_1 = 0$.) This map is just $[z : w] \mapsto [z^2 : w^2]$. The inverse image of $[1 : 1]$ is just $[\pm 1 : 1]$. The reparametrization $[z : w] \mapsto [-z : w]$ takes one to the other. Most points in $\overline{M}_{0,1}(\mathbb{C}\mathrm{P}^1, 2[\mathbb{C}\mathrm{P}^1])$ have trivial automorphism group. This is because the underlying stable map has automorphism group $\mathbb{Z}_2$ and only the trivial automorphism will fix the marked point unless the marked point is at a critical value. It follows that the points along the divisor

$$D_1 = \{[a : b : c], [z : w] | az^2 + bzw + cw^2 = 0\}$$

have automorphism group $\mathbb{Z}_2$.

For the universal curve $\mathcal{U} \to \mathcal{M}$ one should have that the inverse image of a point $s_0 \in \mathcal{M}$ is isomorphic to s_0. This fails with the coarse moduli spaces. The inverse image of a point in $\mathbb{C}\mathrm{P}^2$ is a copy of $\mathbb{C}\mathrm{P}^1$ and the restriction of the evaluation map to this copy is just a degree one map. This should be a degree two map. The reason for this failure is that each map in this moduli space has automorphism group $\mathbb{Z}_2$. It appears that one should take a double cover of $\mathbb{C}\mathrm{P}^2 \times \mathbb{C}\mathrm{P}^1$ branched along the divisor

D_1. The fiber of such a cover over a point in $\mathbb{C}\mathrm{P}^2$ would be a two-fold branched cover of $\mathbb{C}\mathrm{P}^1$ branched over two points. This is also a copy of $\mathbb{C}\mathrm{P}^1$, but the induced evaluation map would have degree two as it should. The only problem with this is that no such cover exists in the category of schemes - such a map would restrict to a two-fold connected cover over a simply-connected space. The map

$$\overline{\mathcal{M}}_{0,1}(\mathbb{C}\mathrm{P}^1, 2[\mathbb{C}\mathrm{P}^1]) \to \overline{\mathcal{M}}_{0,0}(\mathbb{C}\mathrm{P}^1, 2[\mathbb{C}\mathrm{P}^1])$$

has exactly this structure in the category of stacks.

To go further we have to use the power of stacks. Recall that one description of a Deligne–Mumford stack is a contravariant functor from SCHEME to SET. Any scheme produces such a functor. For $\mathbb{C}\mathrm{P}^2$ one gets the functor $\underline{\mathbb{C}\mathrm{P}^2}$ taking a scheme S to $\mathrm{Mor}(S, \mathbb{C}\mathrm{P}^2)$. To analyze the local symmetry groups we need to use the fibered category structure. The objects in the associated fibered category are ordered pairs consisting of an element of the set associated to a scheme and the scheme. The points in the stack are just those objects corresponding to the one point scheme. For example,

$$[\![0:1:0]\!] := ([0:1:0] : \mathrm{pt} \to \mathbb{C}\mathrm{P}^2, \mathrm{pt}) \in \mathrm{Ob}(\mathbf{D}^{\underline{\mathbb{C}\mathrm{P}^2}}),$$

has $\mathrm{Hom}([\![0:1:0]\!], [\![0:1:0]\!]) = \{\mathrm{id}\}$, so the local automorphism group of a point in the stack associated to $\mathbb{C}\mathrm{P}^2$ is trivial as expected.

Let $\mathcal{M} := \overline{\mathcal{M}}_{0,0}(\mathbb{C}\mathrm{P}^1, 2[\mathbb{C}\mathrm{P}^1])$. The argument showing that the corresponding coarse moduli space is isomorphic to $\mathbb{C}\mathrm{P}^2$ shows that the points of this stack correspond to the points of $\mathbb{C}\mathrm{P}^2$. Consider the local automorphisms of a point of this stack. Let

$$z^2 := (u([z:w]) = [z^2 : w^2] \colon \mathbb{C}\mathrm{P}^1 \to \mathbb{C}\mathrm{P}^1, \pi \colon \mathbb{C}\mathrm{P}^1 \to \mathrm{pt}, \mathrm{pt}) \in \mathrm{Ob}(\mathbf{D}^{\underline{\mathcal{M}}}).$$

We have $\mathrm{Hom}(z^2, z^2) = \{\mathrm{id}, n\}$, where $n([z:w]) = [-z:w]$. A similar thing is true for every point in $\mathcal{M}$, so every point in this stack has automorphism group $\mathbb{Z}_2$.

Exercise 10.2 Do a similar analysis in $\overline{\mathcal{M}}_{0,1}(\mathbb{C}\mathrm{P}^1, 2[\mathbb{C}\mathrm{P}^1])$ to show that the points in D_1 have automorphism group $\mathbb{Z}_2$ and all others have trivial automorphism group.

We can interpret the fundamental cycle of the stack $\mathcal{M}$ to be $\frac{1}{2}[\overline{M}_{0,0}(\mathbb{C}\mathrm{P}^1, 2[\mathbb{C}\mathrm{P}^1])]$ or just $\frac{1}{2}[\mathbb{C}\mathrm{P}^2] \in H_4(\mathbb{C}\mathrm{P}^2; \mathbb{Q})$. The factor of $1/2$ here is due to the $\mathbb{Z}_2$ automorphism group. To compute the virtual fundamental class, we need to compute

$$c_{2d+2g-2}(\mathbf{R}^1(\pi_*)(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}(\mathcal{O}(-1) \oplus \mathcal{O}(-1)))).$$

In this case $2d + 2g - 2 = 2$, so the Whitney sum formula gives

$$c_2(\mathbf{R}^1(\pi_*)(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}(\mathcal{O}(-1) \oplus \mathcal{O}(-1)))) = c_1(\mathbf{R}^1(\pi_*)(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}(\mathcal{O}(-1))))^2.$$

Attempting to construct a two-fold cover of $\mathbb{C}P^2 \times \mathbb{C}P^1$ branched along D_1 provides good motivation for the computation of the above first Chern class. In general to construct a p–fold cyclic cover branched along a divisor D, one constructs the line bundle associated to D, say L_D, and takes a section vanishing along D, say σ_D. If $L^{1/p}$ is a pth root of this bundle in the sense that $(L^{1/p})^{\otimes p} \cong L_D$, the desired cover will be

$$\{\xi \in L^{1/p} | \xi^2 = \sigma_D(\pi(\xi))^p\}.$$

Applying this idea to the divisor D_1 in $\mathbb{C}P^2 \times \mathbb{C}P^1$, we cover $\mathbb{C}P^2 \times \mathbb{C}P^1$ by charts V_a, V_b, and V_c corresponding to $a = 1$, $b = 1$ and $c = 1$. Define a bundle $L_c^{1/2} = \mathbb{C}^3 \times (\mathbb{C}^2 - \{0\})/\sim$ where $(a, b, \gamma, z_0, z_1) \sim (a, b, \lambda\gamma, \lambda z_0, \lambda z_1)$, and similar bundles over V_a and V_b. We have a section of the tensor square of this bundle taking $(a, b, [z_0 : z_1])$ to $[a, b, az_0^2 + bz_0z_1 + z_1^2, z_0, z_1]$. The problem is that there is no reasonable way to glue these pieces into a global bundle. Continue anyway and define

$$Q_c := \{[a, b, \gamma, z_0, z_1] \in L_c^{1/2} | \gamma^2 = az_0^2 + bz_0z_1 + z_1^2\}$$

with natural projection map $\pi\colon Q_c \to \mathbb{C}^2$ and $v\colon Q_c \to \mathbb{C}P^1$ given by

$$v([a, b, \gamma, z_0, z_1]) = [z_0 : z_1].$$

Exercise 10.3 Show that π is a flat morphism.

We can see that Q_c is a flat family of stable, genus zero, degree two maps to $\mathbb{C}P^1$. We just need to check that the restriction of v to the inverse image of any point in $\mathbb{C}^2$ is such a stable map. For example, we have an isomorphism $\mathbb{C}P^1 \to \pi^{-1}(1, 0)$ given by $[s : t] \mapsto [1, 0, s^2 - t^2, 2st, s^2 + t^2]$. The composition of this map with the restriction of v is a degree two map with critical values at $[\pm i : 1]$ as expected.

Exercise 10.4 Identify the restriction of v to $\pi^{-1}(0, 0)$.

Stacks should be considered as generalizations of schemes that include orbifold information. Just as a manifold is defined via a maximal atlas while specific manifolds are usually described by a finite atlas, a specific stack can be described by a finite cover while the general definition adds a condition analogous to maximality. One can cover the stack $\overline{\mathcal{M}}_{0,0}(\mathbb{C}P^1, 2[\mathbb{C}P^1])$ by the family Q_c together with two other analogous families denoted by Q_a and Q_b.

Exercise 10.5 Construct analogous families Q_a and Q_b and use these three families to conclude that $\overline{\mathcal{M}}_{0,0}(\mathbb{C}P^1, 2[\mathbb{C}P^1])$ is locally representable.

We have been treating stacks as contravariant functors from the category of schemes to sets, so for example the stack associated to a scheme T is the functor that takes a scheme S to the set of morphisms from S to T. As explained in Appendix A, stacks can also be viewed as fibered categories. The objects of the fibered category $\mathbf{D}^{\underline{T}}$ are just morphisms $u\colon R \to T$, similarly the objects of the fibered category version $\mathbf{D}^{\mathcal{M}}$ of the stack $\overline{\mathcal{M}}_{0,0}(\mathbb{C}\mathrm{P}^1, 2[\mathbb{C}\mathrm{P}^1])$ are just families of stable maps $[v\colon V \to \mathbb{C}\mathrm{P}^1, \pi\colon V \to R]$.

Notice that any flat family Q over a scheme T (really, $[q\colon Q \to \mathbb{C}\mathrm{P}^1, \pi\colon Q \to T]$) living in $\overline{\mathcal{M}}_{0,0}(\mathbb{C}\mathrm{P}^1, 2[\mathbb{C}\mathrm{P}^1])(T)$ defines a map of stacks (covariant functor) taking $\mathbf{D}^{\underline{T}}$ to $\mathbf{D}^{\mathcal{M}}$. This map takes a morphism $u\colon R \to T$ in $\mathbf{D}^{\underline{T}}$ to the fiber product family $[q \circ \mathrm{pr}_1, Q \times_T R \to R]$ where $Q \times_T R := \{(x, y) \in Q \times R | \pi(x) = u(y)\}$ in$\mathbf{D}^{\mathcal{M}}$.

Exercise 10.6 Define the action of the functor on a morphism from $u\colon R \to T$ to $v\colon S \to T$, that is, $w\colon R \to S$ that satisfies $u = v \circ w$.

> An alternate way to think of stacks is as a categorical construction of terminal objects. The moduli space should be a terminal object in the category of families of stable maps. The problem is that generally no such terminal object exists. When we stackify we replace a scheme by a contravariant functor and a morphism by a natural transformation and arrive at a category in which the collection of all families itself turns into a terminal object.

We now turn to the computation of the Chern classes in the formula for the Gromov–Witten invariants from the last subsection. Recall that Chern classes satisfy $c(f^*E) = f^*c(E)$, so we could compute a Chern class of a bundle by computing the Chern class of the pull-back under a finite branched cover and dividing by the degree of the cover. Since $H^2(\mathbb{C}\mathrm{P}^2; \mathbb{Z}) \cong \mathbb{Z}$, to compute the first Chern class of any bundle over $\mathbb{C}\mathrm{P}^2$ it suffices to compute the first Chern class of the restriction of the bundle to $\mathbb{C}\mathrm{P}^1$. Even though there is no bundle corresponding to the square root of the line bundle associated to D_1, there is an object corresponding to the result of taking the pull-back of the restriction to $\mathbb{C}\mathrm{P}^1$ of such a bundle under a two-fold cover. We apply this to the $\mathbb{C}\mathrm{P}^1$ at $a = 0$.

Define

$$\widehat{Q} := \left\{[\alpha, b, c, z_0, z_1] \in \mathbb{C} \times (\mathbb{C}^2 - \{0\})^2 \;\middle|\; \alpha^2 = b^2 z_0 z_1 + c^2 z_0 z_1\right\} / \sim,$$

with $[\alpha, b, c, z_0, z_1] \sim [\lambda\mu\alpha, \lambda b, \lambda c, \mu z_0, \mu z_1]$. This is the total space of a family with projection $\pi\colon \widehat{Q} \to \mathbb{C}\mathrm{P}^1$ taking $[\alpha, b, c, z_0, z_1]$ to $[b : c]$ and evaluation $v\colon \widehat{Q} \to \mathbb{C}\mathrm{P}^1$ given by $v([\alpha, b, c, z_0, z_1]) := [z_0 : z_1]$. Notice that the variables b and c here do not

correspond to the same variables as used earlier. It might have been clearer to use b_1, c_1 here with the relations $b = b_1^2$ and $c = c_1^2$ defining the two-fold branched cover of the $\mathbb{C}\mathrm{P}^1$ cycle. This would just complicate the notation a bit further. The space $\widehat{Q}$ is also a flat family of genus zero, degree two stable maps. We wish to compute $c_1(\mathbf{R}^1(\pi_*)(v^*(\mathcal{O}(-1))))$. Recall that $\mathbf{R}^1(\pi^*)$ is just a vector bundle over $\mathbb{C}\mathrm{P}^1$ with fibers isomorphic to $H^1(\pi^{-1}(-), v|^*(\mathcal{O}(-1))))$. It helps to remember how to compute sheaf cohomology at this point.

Exercise 10.7 Define $L_n := \mathbb{C}^2 \times (\mathbb{C} - \{0\})/\sim$, with $(z_0, z_1, \zeta) \sim (\lambda z_0, \lambda z_1, \lambda^n \zeta)$. Taken with the natural projection to $\mathbb{C}\mathrm{P}^1$, this is a line bundle. Let $\mathcal{O}(n)$ be the associated sheaf of sections. Using the standard $z_1 \neq 0$, $z_0 \neq 0$ cover of $\mathbb{C}\mathrm{P}^1$, compute the Čech cohomology groups $\check{H}^*(\mathbb{C}\mathrm{P}^1; \mathcal{O}(n))$ for various positive and negative values of n.

To compute $\mathbf{R}^1$ we use the following description of it from Hartshorne's book [75].

Theorem 10.8 *If $\mathcal{A}$ is a sheaf over X and $f\colon X \to Y$, then $R^k(f_*)(\mathcal{A})$ is the sheaf associated to the presheaf taking V to $H^k(f^{-1}(V); \mathcal{A}|_{f^{-1}(V)})$.*

One can check that the v pull-back of the L_{-1} bundle over $\mathbb{C}\mathrm{P}^1$ to $\widehat{Q}$ is the bundle defined by

$$v^*L_{-1} := \{[\zeta, \alpha, b, c, z_0, z_1] \in \mathbb{C}^2 \times (\mathbb{C}^2 - \{0\})^2 \mid \alpha^2 = b^2 z_0 z_1 + c^2 z_0 z_1\}/\sim,$$

with $[\zeta, \alpha, b, c, z_0, z_1] \sim [\mu^{-1}\zeta, \lambda\mu\alpha, \lambda b, \lambda c, \mu z_0, \mu z_1]$. Now work over the $c \neq 0$ chart U_c of $\mathbb{C}\mathrm{P}^1$. We can take an open cover of the π–inverse image of this chart consisting of $z_1 \neq 0$ and $z_0 \neq 0$. Computing $R^1(\pi_*)(\mathcal{O}(v^*L_{-1}))(U_c)$ amounts to computing the Čech cohomology of the inverse image of U_c. A 0–Čech cochain consists of an algebraic section of v^*L_{-1} over the $z_1 \neq 0$ chart and a section over the $z_0 \neq 0$ chart. A section over the first takes the form $[f(\alpha, b, z_0), \alpha, b, 1, z_0, 1]$ where $f(\alpha, b, z_0)$ is polynomial and we are using the obvious coordinates obtained by setting c and z_1 to one. A algebraic section over the $z_0 \neq 0$ chart takes the form $[g(\alpha, b, z_0), \alpha, b, 1, 1, z_1]$ with g polynomial. Using the relation from the definition of $\widehat{Q}$ we can eliminate all second and higher powers of α and write $f(\alpha, b, z_0) = f_0(b, z_0) + \alpha f_1(b, z_0)$. The polynomial $g(\alpha, b, z_1)$ can be expressed similarly. A Čech 1–cochain is just a algebraic section on the overlap. The Čech coboundary is just the difference of the restrictions of the sections from the two large charts. In order to compute this difference we must write each section in the same coordinates. Using the $z_1 \neq 0$ chart we can write the Čech coboundary as

$$\delta(f, g) := \big[(f_0(b, z_0) - z_0^{-1} g_0(b, z_0^{-1})) + \alpha(f_1(b, z_0) - z_0^{-2} g_1(b, z_0^{-1})), \alpha, b, 1, z_0, 1\big].$$

Notice that on the overlap $z_0 \neq 0$ and $\alpha^2 = bz_0 + 1$, so any algebraic function on this overlap may be written in the form $z_0^{-N}(F_0(b, z_0) + F_1(b, z_0)\alpha)$, with polynomial F_0 and F_1. Combining this with the expression for the coboundary implies that the cokernel of δ is the $\mathbb{C}[b]$–module generated by $z_0^{-1}\alpha$ and this is the definition of $\mathbf{R}^1(\pi_*)(v^*(\mathcal{O}(-1)))(\{c \neq 0\})$. This confirms our theoretical arguments from the previous subsection that $\mathbf{R}^1(\pi_*)(\mathrm{ev}^*\mathcal{O}(-1))$ is a locally free, finite-rank sheaf of $\mathcal{O}_S$–modules of the correct dimension. On the $b \neq 0$ chart we will use c, z_0 (z_1) and α with $b = 1$ as our coordinates.

Exercise 10.9 Show that $\mathbf{R}^1(\pi_*)(v^*(\mathcal{O}(-1)))(\{b \neq 0\})$ is the $\mathbb{C}[c]$–module generated by $z_0^{-1}\alpha$ in the given coordinates.

We must be careful when making the identifications between the $b \neq 0$ chart and the $c \neq 0$ chart. It might help to use a different variable, say β in place of α when describing the $b \neq 0$ chart. The correct way is to scale the $b = 1$ answer to a $c = 1$ answer using the equivalence from the definition with $\lambda = c^{-1}$. It follows that the section of $\mathbf{R}^1(\pi_*)(v^*(\mathcal{O}(-1)))$ over the $c \neq 0$ chart given by $z_0^{-1}\alpha$ extends to a meromorphic section over all of $\mathbb{C}\mathrm{P}^1$ given by $c^{-1}z_0^{-1}\alpha$ in the $b \neq 0$ chart. This meromorphic section has exactly one simple pole, and no other poles or zeros. It follows that $c_1(\mathbf{R}^1(\pi_*)(v^*(\mathcal{O}(-1))))([\mathbb{C}\mathrm{P}^1]) = -1$. We conclude that

$$c_1(\mathbf{R}^1(\pi_*)(\mathrm{ev}^*(\mathcal{O}(-1))))([\mathbb{C}\mathrm{P}^1]) = -\tfrac{1}{2}.$$

Putting everything together gives

$$\begin{aligned} N_{0,2} &= \int_{[\overline{\mathcal{M}}_{0,0}(\mathbb{C}\mathrm{P}^1, 2[\mathbb{C}\mathrm{P}^1])]} c_1(\mathbf{R}^1(\pi_*)(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}\mathcal{O}(-1)))^2 \\ &= c_1\big(\mathbf{R}^1(\pi_*)\big(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}\mathcal{O}(-1)\big)\big)^2\big(\tfrac{1}{2}[\mathbb{C}\mathrm{P}^2]\big) \\ &= \tfrac{1}{2}\big(c_1\big(\mathbf{R}^1(\pi_*)\big(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}\mathcal{O}(-1)\big)\big)([\mathbb{C}\mathrm{P}^1])\big)^2 = \tfrac{1}{8}\,. \end{aligned}$$

While this direct computation clarifies all of the ingredients in the definition of the Gromov–Witten invariants, it is not very practical for computing the general case. To compute the answer in the general case, we return to localization.

11 The full multiple cover formula via localization

In this section we apply localization to compute the Gromov–Witten invariants of the resolved conifold X_{S^3}. We should really say that we are using virtual localization. This is a generalization of the localization formula that we have already explained in

two different directions. First one must apply localization in the stack setting, and second one must apply it with virtual fundamental classes. The correct generalization is given by Graber and Pandharipande [66].

The standard torus action on $\mathbb{C}\mathrm{P}^1$ extends to a torus action on all of the spaces involved in the multiple cover formula for the Gromov–Witten invariants of $\mathcal{O}(-1)\oplus\mathcal{O}(-1)$. In fact this action extends to the bundle $\mathcal{O}(-1)$ in several ways. Just as line bundles are classified by the first Chern class, equivariant line bundles are classified by the equivariant first Chern class. An action on a vector bundle compatible with an action on the base is called a linearization. Using the classification of equivariant bundles it is standard to label linearizations by the associated equivariant first Chern class. Thus we label the linearizations of $\mathcal{O}(-1)$ by $n\alpha_0+m\alpha_1-h$.

Exercise 11.1 Construct group actions on $\mathcal{O}(-1)$ corresponding to the equivariant class $n\alpha_0+m\alpha_1-h$.

The numbers that we need to compute are

$$N_{g,d}=\int_{[\overline{\mathcal{M}}_{g,0}(\mathbb{C}\mathrm{P}^1,d[\mathbb{C}\mathrm{P}^1])]} c_{2d+2g-2}\big(\mathbf{R}^1(\pi_*)\big(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}(\mathcal{O}(-1)\oplus\mathcal{O}(-1))\big)\big).$$

Based on the computation of N_3 via localization in Section 7.3 one might worry that this localization computation is going to be very complicated. In fact, the computation is fairly straightforward provided that one chooses the proper linearizations. Faber and Pandharipande discovered that if one chooses α_0-h as the linearization on the first factor of $\mathcal{O}(-1)$ and α_1-h on the second factor then only components of the fixed point set corresponding to one graph contribute to $N_{g,d}$ [55].

To see why only one graph can contribute let $[u,\Sigma,p]$ be a stable map fixed by the group action. We start with the normalization sequence

$$0\to\mathcal{O}_\Sigma\to\nu_*\mathcal{O}_{\widehat{\Sigma}}\to\oplus_c\mathcal{O}_c\to 0.$$

We already used this sequence in computing the Euler class of the normal bundle associated to deformations of the map in Section 8.4. As before we use c to denote the nodes of Σ. We have a related exact sequence for holomorphic sections of the pull-back bundle $u^*\mathcal{O}(-1)$:

$$0\to u^*\mathcal{O}(-1)\to\nu_*\nu^*u^*\mathcal{O}(-1)\to\oplus_c L_{-1}|_{u(c)}\to 0.$$

The associated long exact sequence on cohomology reads,

$$\cdots\to H^0(\widehat{\Sigma},\nu^*u^*\mathcal{O}(-1))\to\oplus_n L_{-1}|_{u(n)}\to H^1(\Sigma,u^*\mathcal{O}(-1))\to\cdots.$$

There are bundles over the moduli space with fibers isomorphic to the spaces in this exact sequence. We will use the fibers as names for these bundles so for example, we will denote the bundle $\mathbf{R}^1(\pi_*)\big(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}(\mathcal{O}(-1))\big)$ by $H^1(\Sigma, u^*\mathcal{O}(-1))$.

The normalization is a union of smooth components $\widehat{\Sigma} = \coprod \Sigma_k$ and the cohomology $H^0(\widehat{\Sigma}, \nu^*u^*\mathcal{O}(-1))$ is isomorphic to $\oplus H^0(\Sigma_k, \nu^*u^*\mathcal{O}(-1))$. The contribution $H^0(\Sigma_k, \nu^*u^*\mathcal{O}(-1))$ is trivial unless the component Σ_k is contracted to a point under the stable map in which case it is isomorphic to $\mathbb{C}$. Since the original prestable curve is connected we see that there is at least one node attached to each contracted component and there are extra nodes if there is any vertex in the graph associated to the stable curve with valence grater than one.

It follows that the cokernel of the map

$$H^0(\widehat{\Sigma}, \nu^*u^*\mathcal{O}(-1)) \to \oplus_c L_{-1}|_{u(c)}$$

is only trivial if there are no vertices of valence greater than one.

In the case the cokernel is nontrivial the long exact sequence of bundles implies that $e\big(\mathbf{R}^1(\pi_*)\big(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}(\mathcal{O}(-1))\big)\big)$ contains a factor of $e(L_{-1}|_{u(c)})$. In order for the stable map to be equivariant $u(c)$ must either equal $q_0 = [1:0]$ or q_1. Without loss of generality it is equal to q_0. Now consider the factor $e\big(\mathbf{R}^1(\pi_*)\big(\mathrm{ev}^*_{\mathbb{C}\mathrm{P}^1}(\mathcal{O}(-1))\big)\big)$ corresponding to the $\alpha_0 - h$ linearization. The corresponding linearization of $L_{-1}|_{u(c)}$ is trivial, implying that $e(L_{-1}|_{u(c)})$ is trivial, so such a stable map cannot contribute to the Gromov–Witten invariant.

It follows that the only components of the fixed point set that contribute to the Gromov–Witten invariant correspond to graphs consisting of one straight edge. Consider such a graph with genus g_k on the q_k vertex and compute the Euler class in the integrand and the Euler class of the normal bundle to the fixed point set.

Start with the Euler class in the integrand. We have seen that the cokernel of the map into the fibers over the nodes is trivial in this case. On the other hand, the higher cohomology groups of any skyscraper sheaf like $L_{-1}|_{u(c)}$ vanish. If follows from the long exact cohomology sequence that

$$\begin{aligned}(16)\quad H^1(\Sigma, u^*\mathcal{O}(-1)) &\cong H^1(\widehat{\Sigma}, \nu^*u^*\mathcal{O}(-1))\\ &\cong H^1(\Sigma_{g_0}; L_{-1}|_{q_0}) \oplus H^1(\Sigma_{g_1}; L_{-1}|_{q_1}) \oplus H^1(\mathbb{C}\mathrm{P}^1; u^*\mathcal{O}(-1))\,.\end{aligned}$$

We now need to compute the equivariant Euler classes of these bundles. For this we need to describe the group actions on the relevant spaces. The degree d line bundle over $\mathbb{C}\mathrm{P}^1$

may be described as equivalence classes $[z_0, z_1, \xi]$ where $(az_0, az_1, a^d\xi) \sim (z_0, z_1, \xi)$ for $a \in \mathbb{C} - \{0\}$. One can check that the action

$$\lambda \cdot [z_0, z_1, \xi] = [\lambda_0 z_0, \lambda_1 z_1, \lambda_0^{n_0}\lambda_1^{n_1}\xi]$$

is the linearization labeled by $dh - n_0\alpha_0 - n_1\alpha_1$. This is the answer to Exercise 11.1.

Exercise 11.2 Show that the action on the fiber over $q_0 := [1:0]$ is $(d-n_0)\alpha_0 - n_1\alpha_1$. Use this to show that the natural linearization on $T\mathbb{C}\mathrm{P}^1$ is $2h - \alpha_0 - \alpha_1$. Also show that the linearization on a tensor product of line bundles is the sum of the corresponding linearizations.

Exercise 11.3 Recall that the holomorphic sections of the degree d line bundle over $\mathbb{C}\mathrm{P}^1$ are just degree d polynomials. Use this to show that the linearization $dh - n_0\alpha_0 - n_1\alpha_1$ on $\mathcal{O}(d)$ turns $H^0(\mathbb{C}\mathrm{P}^1; \mathcal{O}(d))$ into the representation

$$\oplus_{k=0}^{2}[(k-n_0)\alpha_0 + (d-k-n_1)\alpha_1].$$

Finally, recall that to get an action on the domain of the degree d map $[z_0 : z_1] \mapsto [z_0^d : z_1^d]$ one must pass to the d th power action on the codomain. This is kept track of by dividing by d at the end.

Kodaira–Serre duality implies that

$$H^1(\mathbb{C}\mathrm{P}^1; u^*\mathcal{O}(-1)) \cong H^0(\mathbb{C}\mathrm{P}^1; \mathcal{O}(d) \otimes \mathcal{O}(-2))^\vee,$$

where the $\mathcal{O}(d)$ inherits a linearization from $\mathcal{O}(-1)$ and $\mathcal{O}(-2)$ has the natural linearization arising as the cotangent bundle (dualizing sheaf) on $\mathbb{C}\mathrm{P}^1$. Using the $\alpha_0 - h$ linearization on $\mathcal{O}(-1)$ the previous two exercises together with the above remarks give the Euler class of the last summand of equation (16) as

$$e(H^1(\mathbb{C}\mathrm{P}^1; u^*\mathcal{O}(-1))) = \prod_{k=0}^{d-2}[(d-k-1)\alpha_0/d + (k-d+1)\alpha_1/d] = (d-1)!d^{1-d}(\alpha_0-\alpha_1)^{d-1}. \tag{17}$$

Similarly, with the $\alpha_1 - h$ linearization we obtain

$$e(H^1(\mathbb{C}\mathrm{P}^1; u^*\mathcal{O}(-1))) = (-1)^{d-1}(d-1)!d^{1-d}(\alpha_0-\alpha_1)^{d-1}.$$

Now Kodaira–Serre duality implies that

$$H^1(\Sigma_{g_0}; L_{-1}|_{q_0}) \cong H^0(\Sigma_{g_0}; \omega_{\Sigma_{g_0}})^\vee \otimes L_{-1}|_{q_0} \cong \mathbb{E}^\vee \otimes L_{-1}|_{q_0}.$$

Recall that $\mathbb{E}$ is the Hodge bundle, which is by definition the bundle over the moduli space with fiber over a point isomorphic to the first cohomology of the curve representing

the point with coefficients in the dualizing sheaf ω_Σ. By Exercise 8.8 we conclude that

$$e(H^1(\Sigma_{g_0}; L_{-1}|_{q_0})) = \sum_{i=0}^{g_0} c_i(\mathbb{E}^\vee) t^{g_0-i}\,, \tag{18}$$

where t is the linearization (equivariant first Chern class) of $L_{-1}|_{q_0}$. When the linearization on L_{-1} is $\alpha_0 - h$ we have $t = 0$; when the linearization is $\alpha_1 - h$ we have $t = \alpha_1 - \alpha_0$.

Now turn to the equivariant Euler class of the normal bundle to the fixed point set. Consider a component of the fixed point set with $0 < g_0 < g$. Formula (8) implies that $e(\mathfrak{aut}([\Sigma, p])^{\mathrm{mov}}) = 1$, while formula (9) gives

$$e(\mathfrak{def}([\Sigma, p])^{\mathrm{mov}}) = ((\alpha_0 - \alpha_1)/d - \psi_0)\,((\alpha_1 - \alpha_0)/d - \psi_1)\ .$$

Similarly formula (11) gives

$$\prod_c e(T_{u(c)}\mathbb{C}\mathrm{P}^2)) = -(\alpha_0 - \alpha_1)^2\,,$$

formula (12) gives

$$e(H^0(\widehat{\Sigma}, \mathcal{O}(u^* T\mathbb{C}\mathrm{P}^2))^{\mathrm{mov}}) = -(-1)^d (d!)^2 d^{-2d} (\alpha_0 - \alpha_1)^{2d+2}\,,$$

and formula (13) gives

$$e(H^1(\widehat{\Sigma}; \mathcal{O}(\nu^* u^* T\mathbb{C}\mathrm{P}^2))^{\mathrm{mov}}) = \left(\sum_{i=0}^{g_0} c_i(\mathbb{E}^\vee)(\alpha_0 - \alpha_1)^{g_0-i}\right)\left(\sum_{i=0}^{g_1} c_i(\mathbb{E}^\vee)(\alpha_1 - \alpha_0)^{g_1-i}\right).$$

Localization also works on stacks with virtual fundamental cycles (see Graber and Pandharipande [66]). The factor $e(\mathfrak{ob}([u, \Sigma, p])^{\mathrm{mov}})$ does not need to be computed because it is accounted for in the virtual fundamental cycle. Automorphisms of generic elements of the fixed point components do have to be taken into account. The virtual localization formula reads (compare to (6) and (7)):

$$\int_{M^{\mathrm{vir}}} \phi = \sum_F \frac{1}{|\mathbb{A}_F|} \int_{F^{\mathrm{vir}}} \frac{\iota_F^* \widehat{\phi}}{e(N(F)^{\mathrm{vir}})}\,.$$

Combining all of the above formulas together with (14) and (10) shows that the contribution to the Gromov–Witten invariant coming from the g_0, g_1 fixed point

component is

$$d^{-1}(-1)^d(d!)^{-2}d^{2d}(\alpha_0-\alpha_1)^{-2d}(d-1)!d^{-1}(\alpha_0-\alpha_1)^{d-1}(-1)^{d-1}(d-1)!d^{-1}(\alpha_0-\alpha_1)^{d-1}$$
$$\left(\int_{[\overline{\mathcal{M}}_{g_0,1}]^{\mathrm{vir}}}\sum_{i=0}^{g_0}c_i(\mathbb{E}^\vee)(\alpha_0-\alpha_1)^{g_0-1}c_{g_0}(\mathbb{E}^\vee)\sum_{i=0}^{g_0}c_i(\mathbb{E}^\vee)(\alpha_1-\alpha_0)^{g_0-i}((\alpha_0-\alpha_1)/d-\psi_0)^{-1}\right)$$
$$\left(\int_{[\overline{\mathcal{M}}_{g_1,1}]^{\mathrm{vir}}}\sum_{i=0}^{g_1}c_i(\mathbb{E}^\vee)(\alpha_0-\alpha_1)^{g_1-1}c_{g_1}(\mathbb{E}^\vee)\sum_{i=0}^{g_1}c_i(\mathbb{E}^\vee)(\alpha_1-\alpha_0)^{g_1-i}((\alpha_1-\alpha_0)/d-\psi_1)^{-1}\right).$$

Using the fact that $c_i(E^\vee)=(-1)^ic_i(E)$ in general together with the relation $c(\mathbb{E})c(\mathbb{E}^\vee)=1$ proved by Mumford for the Hodge bundle [115] this contribution can be simplified to

$$d^{2g-3}\int_{[\overline{\mathcal{M}}_{g_0,1}]^{\mathrm{vir}}}c_{g_0}(\mathbb{E})\psi^{2g_0-2}\int_{[\overline{\mathcal{M}}_{g_1,1}]^{\mathrm{vir}}}c_{g_1}(\mathbb{E})\psi^{2g_1-2}.$$

Exercise 11.4 Prove that the same formula is valid when either g_0 or g_1 is zero provided $\int_{[\overline{\mathcal{M}}_{g_0,1}]^{\mathrm{vir}}}c_{g_0}(\mathbb{E})\psi^{2g_0-2}$ is interpreted to be one when $g_0=0$.

Expressions such as

$$b_{g_0}:=\int_{[\overline{\mathcal{M}}_{g_0,1}]^{\mathrm{vir}}}c_{g_0}(\mathbb{E})\psi^{2g_0-2}$$

are called Hodge integrals. The classes $c_k(\mathbb{E})$ are called Hodge classes and are often denoted by λ_k.

If a different pair of linearizations was chosen for the two $\mathcal{O}(-1)$ factors, one would obtain a very different looking expression for this Gromov–Witten invariant. This generates relations between the various Hodge integrals that can be used together with similar relations coming from an integral with a $\mathcal{O}(1)$ factor to compute the Hodge integrals. This is the approach taken by Faber and Pandharipande [55]. The result obtained there is

$$\sum_{g=0}^{\infty}b_gs^{2g}=\left(\frac{s/2}{\sin(s/2)}\right).$$

Relations in the cohomology of moduli space allow one to express the cubic Hodge integral from Exercise 9.13 in terms of the above Hodge integrals. The answer obtained in [55] for $g\geq 2$ is

$$\int_{[\overline{\mathcal{M}}_{g,0}]^{\mathrm{vir}}}c_{g-1}(\mathbb{E})^3=\frac{(1-2g)B_{2g}B_{2g-2}}{(2g-2)(2g)!}.$$

This can be combined with Exercise 9.13 to obtain the following formula for the degree zero invariants of Calabi–Yau 3–folds:

$$N_{g,0}(X) = \frac{(-1)^{g-1}(2g-1)B_{2g}B_{2g-2}\,\chi(X)}{2(2g-2)(2g)!}\,. \tag{19}$$

12 The Gromov–Witten free energy

In this final section on the Gromov–Witten invariants, we cover a way to package the Gromov–Witten invariants in case all of the unmarked moduli spaces have zero virtual dimension. For a Calabi–Yau 3–fold, this assumption is true. This implies that the most interesting invariants are of the form

$$N_{g,\beta}(X) := \langle\,\rangle^X_{g,\beta} = \int_{[\overline{\mathcal{M}}_{g,n}(X,\beta)]^{\mathrm{vir}}} 1\,.$$

These are combined in the following definition.

Definition 12.1 The Gromov–Witten free energy of a Calabi–Yau 3-fold X is the following formal function depending on a complex parameter y and a cohomology class $t \in H^2(X;\mathbb{C})$.

$$F^{GW}_X(t,y) := \sum_{g=0}^{\infty}\sum_{\beta} N_{g,\beta}(X)e^{-\langle t,\beta\rangle}y^{2g-2}\,.$$

The restricted Gromov–Witten free energy is the sum taken over all non-zero homology classes β.

It might be interesting to formulate a free energy for more general symplectic manifolds. The Gromov–Witten invariants of X_{S^3} as originally computed in the previous article are given by

$$N_{g,d} := N_{g,d[\mathbb{C}\mathrm{P}^1]} = d^{2g-3}\sum_{g_0+g_1=g} b_{g_0}b_{g_1},$$

where b_g are the Hodge integrals computed by Faber and Pandharipande presented in the form of the generating function

$$\sum_{g=0}^{\infty} b_g s^{2g} = \left(\frac{s/2}{\sin(s/2)}\right).$$

Exercise 12.2 Combine the last three displayed formulas to prove that the restricted Gromov–Witten free energy of X_{S^3} is

$$\widehat{F}^{GW} X_{S^3} = \sum_{d=1}^{\infty} \frac{1}{d} \left(2 \sin \frac{dy}{2}\right)^{-2} e^{-td} .$$

It is the full Gromov–Witten free energy that appears in the gauge-string duality. The full Gromov–Witten free energy is simply the sum of the restricted free energy with the degree zero invariant computed earlier. After we give an overview of path integral techniques we provide a heuristic argument for combining the Gromov–Witten invariants into this generating function. This is contained at the end of Section 15.

We can think of t as a complex parameter if we identify $H^2(X_{S^3}; \mathbb{C})$ with $\mathbb{C}$ via evaluation on $[\mathbb{C}\mathrm{P}^1]$. The Gopakumar–Vafa integrality conjecture states that the Gromov–Witten invariants are uniquely specified by a set of integer invariants called Gopakumar–Vafa invariants (or BPS states) [65]. These BPS states are denoted by n_β^g and are supposed to be a count of embedded genus g, J–holomorphic curves in the homology class β. The Gopakumar–Vafa integrality conjecture takes the exact form

$$F^{GW}(X) = \sum_{g=0}^{\infty} \sum_{\beta} \sum_{d=1}^{\infty} n_\beta^g \frac{1}{d} \left(2 \sin \frac{dy}{2}\right)^{2g-2} e^{-d\langle t, \beta \rangle} .$$

Clearly X_{S^3} satisfies the Gopakumar–Vafa conjecture with $n^0_{[\mathbb{C}\mathrm{P}^1]} = 1$ and the rest of the BPS states equal to zero.

Part II Witten–Chern–Simons theory

There is a vast amount of literature on Witten's (quantum) Chern–Simons theory, conformal field theory, quantum groups and so on. We will try to outline the basic notions and definitions that are needed to address Large N Duality. The first subsection below reviews background material about surgery and 3–manifolds. The easiest definition of Witten's Chern–Simons invariants is based on skein theory (Appendix C). It is however, difficult to to compute the resulting invariants or to show that they are well-defined. For this reason we will work with the definition based on quantum group. Section 15 describes the physical motivation for these invariants. Path integral motivation leads to perturbative invariants. Quantum field theory motivation leads to the Reshetikhin–Turaev (exact) invariants. Definitions of the Chern–Simons partition function and free energy based on the latter are given in Sections 16.7, 18.1 and 19.2 after several sections on background and motivation.

13 Framed links and 3–manifolds

Chern–Simons theory provides topological invariants of 3–manifolds and framed links in 3–manifolds. There is a close connection between framed links and 3–manifolds, namely any closed oriented 3–manifold may be obtained by surgery on a framed link in the 3–sphere.

Definition 13.1 A framed link is an embedding of a finite disjoint union of copies of $S^1 \times D^2$ into a 3–manifold. Two framed links are considered equivalent if they are related by an ambient isotopy.

Any ambient isotopy can be decomposed into elementary isotopies called Reidemeister moves. Two framed links are isotopic if and only if they are related by a finite sequence of Reidemeister moves. Framed links in $\mathbb{R}^3$ can be cut into elementary pieces called

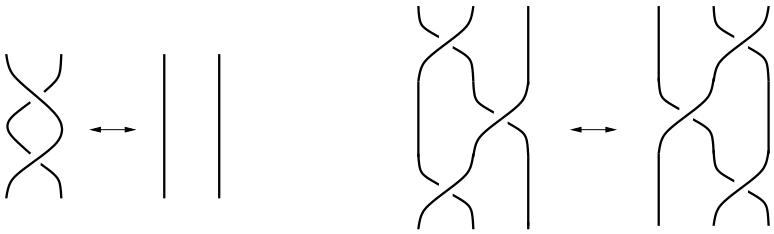


Figure 13.1: Reidemeister moves II and III

tangles (see Figure 16.2). It is often easier to analyze elementary tangles.

Definition 13.2 A framed (or ribbon) tangle is an embedding of a finite disjoint union of copies of $S^1\times D^2$ and $[0,1]\times D^2$ into $[0,1]\times\mathbb{R}^2$ taking $\{0,1\}\times D^2$ into $\{0,1\}\times\mathbb{R}^2$. The embedding of $\{0,1\}\times D^2$ into $\{0,1\}\times\mathbb{R}^2$ must be standard, depending on the number of components landing on each boundary component so that ribbon tangles may be stacked. Tangles are also considered up to isotopy.

Figure 13.2: Framed link projection and ribbon link

In order to move everything into closed 3–manifolds it is standard to work with the one point compactification of $\mathbb{R}^3$. This is homeomorphic to the 3–sphere. By general position we may assume that any framed link in S^3 misses the point at infinity. All framed links in S^3 can therefore be brought back to $\mathbb{R}^3$ and represented by a projection of the cores $S^1\times\{0\}$ to a plane keeping track of over-crossings and under-crossings.

To recover a framed link from the projection one first pushes the over-crossings slightly above the plane to obtain an embedding $\gamma\colon \coprod S^1\hookrightarrow\mathbb{R}^3$. This is then extended to an embedding $\widehat{\gamma}\colon \coprod S^1\times D^2\hookrightarrow\mathbb{R}^3$ by $\widehat{\gamma}(t,x,y):=\gamma(t)+\mathbf{k}x+\dot{\gamma}(t)\times\mathbf{k}y$. This convention is called the blackboard framing. The same technique may be used to represent ribbon tangles. Figure 13.2 displays the projection of a framed link and the image of $S^1\times(\{0\}\times[0,1])$. This second picture justifies the name 'ribbon'.

Framed links in S^3 can be used to construct more complicated 3–manifolds by a process called surgery.

Definition 13.3 Surgery on a framed link $\widehat{\gamma}$ refers to removing the image of the $S^1\times B^2$'s and attaching the same number of $S^1\times D^2$'s in such a way as to glue $\{1\}\times S^1$ to $\widehat{\gamma}(S^1\times\{(0,1)\})$.

If you have never seen surgery before, the book by Rolfsen is a good reference [130]. Other good references that are relevant to this exposition are Prasolov–Sossinsky [124] and Kassel [82].

By a theorem of Lickorish and Wallace any closed oriented 3–manifold can be obtained by surgery on a framed link in S^3 [130]. It is not difficult to see that surgery on isotopic framed links produces homeomorphic 3–manifolds. What is less obvious but

still not difficult is that surgery on two framed links related by an additional move called the Kirby move (Figure 13.3) still produces homeomorphic 3–manifolds. In fact, Kirby proved that two framed links represent the same 3–manifold if and only if they are related by a sequence of Reidemeister and Kirby moves [84; 130]. Actually, what we are calling the Kirby move was introduced by Fenn and Rourke [56]. Kirby himself used an equivalent pair of moves: blow up/down and handle slide [84]. Blow up/down adds or removes an unlinked, (± 1)–framed circle and handle slide tubes one component of a link to a parallel copy of a second component. The names come from descriptions of 3–manifolds as boundaries of 4–manifolds and the corresponding moves on 4–manifolds.

Exercise 13.4 Show that surgery on the framed link with projection a simple circle in the plane gives $S^2 \times S^1$.

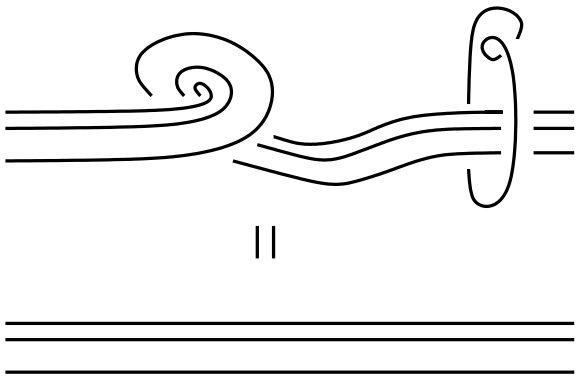


Figure 13.3: Left Kirby move

A framed link in an arbitrary closed oriented 3–manifold can thus be represented by a pair of disjoint framed links in S^3. One of them represents a surgery description of the 3–manifold and the other represents the framed link in it. It follows that to define an invariant of 3–manifolds or framed links in 3–manifolds it suffices to define it on a pair of framed links. Of course one then has to prove that this link invariant is also invariant under the Kirby moves. There are many invariants of links and framed links, some are described in Appendix C. Most of them are not invariant under the Kirby moves and therefore do not produce invariants of 3–manifolds. But Chern–Simons theory does generate 3–manifold invariants along these lines.

14 Physical and heuristic descriptions

The definition of the Chern–Simons invariants of 3–manifolds was motivated by quantum field theory (QFT) (see Deligne et al [44] and Zee [162]). We can afford to be sketchy because this will only serve as motivation for the mathematically rigorous definition of the Reshetikhin–Turaev invariants that we use later. The mathematical foundations of quantum field theory are not completely developed, but the existing machinery and conjectural structure of QFT has produced and motivated many remarkable mathematical theorems. The invariants that we are discussing in this section are the expectation values of observables of the QFT associated to the Chern–Simons action. These values do not depend on any additional geometric structure and are therefore topological invariants of the underlying 3–manifold space-time. Such QFT's are called topological quantum field theories TQFT's.

First recall the Lagrangian approach to quantum mechanics. In this approach the partition function is equal to the integral of the exponential of the action that is, $Z = \int e^{i\hbar^{-1}S}$. For the quantum mechanics of a classical particle under the influence of a conservative force with potential V the action is $S = \int \frac{1}{2}m\dot{x}(t)^2 - V(x(t))\,dt$. The classical equations of motion are just the stationary curves of this functional. It is exactly these stationary points that contribute to highest order in the stationary phase approximation of Z. This generalizes in an obvious way for extra degrees of freedom. The same framework generalizes to continuum mechanics and field theory.

The fields in Chern–Simons theory are connections. A connection can be viewed as a Lie algebra valued 1–form on the 3–manifold. Think of the sum $A = A_i\theta^i$ where the matrices A_i live in the Lie algebra and the θ^i are 1–forms. If R is a representation of a Lie algebra we define a trace function on the algebra by $\mathrm{Tr}_R(A) = \mathrm{Tr}(R(A))$. The Chern–Simons action of a $U(N)$ connection is given by

$$CS(A) = \frac{1}{4\pi}\int_M \mathrm{Tr}_\square\big(A\wedge dA + \tfrac{2}{3}A\wedge A\wedge A\big)$$

where $\square$ is the defining representation of $U(N)$. Witten suggested the simple idea that the average of a function of this Chern–Simons action taken over all connections should be a topological invariant. Indeed, nothing in the definition of CS depends on any geometric data [159] except the connection which is integrated out. This average is called the Chern–Simons partition function. It is formally written as

$$Z_k(M) = \int_{\mathcal{A}} e^{\frac{i}{2x}CS(A)}\mathcal{D}A\,,$$

where $\mathcal{D}A$ is an as of yet undefined measure on the space of connections and x is the so-called string coupling constant. In the path integral expression one takes $x = \frac{2\pi}{k}$

where k is a positive integer called the level. It turns out that after performing formal perturbative expansion in x one needs to 'renormalize' it to $x = \frac{2\pi}{k+N}$, where N is the rank of $U(N)$ to get the 'correct' answer. The explanation for this shift is not fully understood mathematically and underscores subtleties of infinite-dimensional integration. This shift comes from the interpretation of the signature of an operator as an eta invariant, see Witten [159] and Atiyah [14]. Because of this some authors call the level k and others will call it $k+N$, so one must be careful when comparing different results in the literature. We call the level k.

This formal representation of the partition function is an example of a Feynman path integral (see Etingof [54]). This invariant or various normalizations of it is more often denoted by $\tau(M)$ in the mathematical literature. The problem is that this 'average' is not well-defined because the space of connections is infinite-dimensional and a translation-invariant measure does not exist. However, there are ways to formally define invariants that have most of the properties expected of this 'average'. It turns out however that they do depend on an additional geometric structure on a 3–manifold known as 2–framing (trivialization of $T(TM)$ up to homotopy) [14; 159]. This phenomenon is called gravitational anomaly by physicists and is sometimes explained by the 'measure' $\mathcal{D}A$ not being purely topological [162]. Gravitation in physics is represented by a background metric and a metric in its turn determines many additional structures including a 2–framing. Realistic quantum field theories such as quantum chromodynamics do depend on metric or in physical terms, are coupled to gravity. In (almost) topological quantum Chern–Simons theory framing dependence can be seen as a lingering ghost of this metric dependence.

This being said, 2–framing is a very weak structure, so weak in fact that every 3–manifold admits a canonical one. Using it one can normalize the partition function so as to cancel out the framing dependence altogether. This is exactly the Reshetikhin–Turaev normalization of invariants that we adopt in Definitions 16.35, 19.19. It differs from the physical normalization used by Witten [159] and this has implications for Large N Duality. For instance, Ooguri and Vafa [121] find a much better agreement between the gauge and string partition functions than we do. Unfortunately, there is no consistent definition of the 'physical normalization'. In examples it is usually derived ad hoc by comparing the exact answers to perturbative expansions, see eg Rozansky [132].

One can also 'average' holonomies around colored framed links. A colored framed link is a framed link with a group representation associated to each component. The Chern–Simons invariants of colored framed links are the expectation values of observables constructed from the holonomy of connections. In physical language these observables are called Wilson loop operators. The holonomy is given by $\mathrm{Hol}_A(\gamma) := \mathrm{P}\exp\oint_\gamma A$. More explicitly this means that one solves the system of ODE's given by

$\frac{d}{dt}X(t) + A(\dot\gamma(t)) = 0$ with initial data $X(0) = I$. Given this the holonomy is given by $\mathrm{Hol}_A(\gamma) := X(1)$. Here we are assuming that $\gamma(0) = \gamma(1)$.

A Wilson loop operator for one component is the trace of the holonomy to a given connection along that component in a given representation. That is $W_R^K(A) := \mathrm{Tr}_R(\mathrm{Hol}_A(K))$. The link invariant associated to these Wilson loop operators is just the vacuum expectation value (vev) or correlation function:

$$W_{R_1,\dots,R_c}(L) := \frac{1}{Z_k(M)}\int_{\mathcal{A}} e^{\frac{i}{2x}CS(A)}\prod W_{R_i}^{L_i}(A)\mathcal{D}A.$$

The invariants defined mathematically based on this motivation (up to various different normalizations) are called colored Jones polynomials for $\mathrm{SU}(2)$ and the colored THOMFLYP polynomials for $\mathrm{SU}(N)$. They are sometimes also denoted by $J(L, R_1, \dots, R_c)$ for $M = S^3$ or by $\tau(M, L)$ in general.

Remark 14.1 This $W_{R_1,\dots,R_c}(L)$ is an invariant of oriented framed links, as changing the orientation inverts the holonomy.

After a physical construction of invariants, Witten went further and outlined ideas that led to one way of making these invariants mathematically rigorous. He argued using skein relations that the expectation values of Wilson loop operators are given by the Jones polynomial of the corresponding links. Moreover, he gave an explicit prescription for computing Chern–Simons partition functions of 3–manifolds based on their link surgery presentation, see Witten [159] and Axelrod–Della Pietra–Witten [19]. This was an amazing insight, but Witten's surgery prescription is a long way from a mathematically rigorous definition of the invariants. Reshetikhin and Turaev were first to devise a rigorous definition based on quantum groups [128; 129]. This is the definition that we will ultimately use.

There are two philosophically different ways to interpret the expressions for these invariants: the perturbative approach and the TQFT approach. Each of these approaches can be formalized in different ways. The perturbative approach is outlined in Section 15 and the exact approach is outlined in Section 16. Witten's original idea is also explained in the book of Atiyah [14].

15 Perturbative Chern–Simons theory

Here we outline the perturbative approach to Chern–Simons theory because it motivates many of the definitions and conjectures that appear later. Numerous other authors have written expositions on perturbative expansions (see Sawon [136] in this volume,

Bar-Natan [23] and Polyak [123]). We include an overview here because it helps motivate Large N Duality. The perturbative approach is a generalization of two ideas for finite-dimensional integrals: the stationary phase approximation for oscillatory integrals, and a graphical calculus due to Feynman for evaluating Gaussian integrals (see Etingof [54]).

Recall the stationary phase expansion,

$$\int_M e^{i\lambda H}\, d\mathrm{vol}_M = \sum_{dH|_p=0} (2\pi\lambda^{-1})^{n/2} e^{\pi i \mathrm{sgn}(D^2 H_p)} |\det D^2 H_p|^{-1/2} e^{i\lambda H(p)} + O(\lambda^{-n/2-1}).$$

A theorem of Duistermaat and Heckman asserts that this is exact (with no $O(\lambda^{-n/2-1})$ term) when M is a symplectic manifold and H is an invariant Hamiltonian with only non-degenerate critical points [52]. One nice proof of this is based on the localization formula discussed in the section on Gromov–Witten invariants. In fact, the localization formula was originally discovered in an attempt to better understand why the stationary phase approximation was exact (see Atiyah and Bott [15]).

Similarly, there is an infinite-dimensional symplectic structure on the space of connections and the Chern–Simons action is invariant under the action of the gauge group. So one might expect that the stationary phase approximation is exact in this setting. The critical points of the Chern–Simons action are flat connections (see Baez and Muniain [20]) and one can define a perturbative expansion about these flat connections by analogy to the stationary phase approximation (see Bar-Natan[23; 22]). For the unknot in S^3 the agreement between perturbative and exact invariants has been verified, see Bar-Natan–Garoufalidis–Rozansky–Thurston [25]. There are still interesting open questions related to the appropriate interpretation of the full expansion on nontrivial manifolds since such manifolds admit nontrivial flat connections as critical points for the perturbative expansion.

To better understand the structure of the perturbative expansion we consider a finite-dimensional Gaussian integral analog.

Exercise 15.1 Recall that $f(a) = \int_{-\infty}^{\infty} e^{-ax^2}\, dx$ may be evaluated by squaring it and then converting to polar coordinates. By taking successive derivatives of $f(a)$ evaluate $\int_{-\infty}^{\infty} x^{2n} e^{-ax^2}\, dx$.

The expressions in the previous exercise become more complicated as n grows and are even more complicated for integrals over higher-dimensional Euclidean spaces.

Feynman added some slick book-keeping machinery to produce an efficient method for computing higher-dimensional Gaussian integrals. These integrals are analogous to the path integrals that arise in Quantum Field Theory.

For finite-dimensional integrals the method is as follows. Let Q be a symmetric bilinear form on $\mathbb{R}^n$ and let V be a trilinear form on $\mathbb{R}^n$. There is an obvious analogy between the following Gaussian integral,

$$Z = \int_{\mathbb{R}^n} e^{-\hbar^{-1}(Q(x,x)/2+V(x,x,x)/6)}\, d^n x\,,$$

and the path integral formally defining the Chern–Simons partition function,

$$Z_k(M) = \int_{\mathcal{A}} e^{\frac{i}{8\pi x}\left(\int_M \mathrm{Tr}_\square(A\wedge dA)+\int_M \mathrm{Tr}_\square(\frac{2}{3}A\wedge A\wedge A)\right)}\mathcal{D}A\,.$$

Substituting $x = \sqrt{\hbar}\,y$ and expanding the second exponential gives

$$\begin{aligned}
&= \hbar^{\frac{n}{2}} \int_{\mathbb{R}^n} e^{-Q(y,y)/2}\cdot e^{-\sqrt{\hbar}V(y,y,y)/6}\, d^n y \\
&= \hbar^{\frac{n}{2}} \sum_{m=0}^{\infty} \int_{\mathbb{R}^n} e^{-Q(y,y)/2} \frac{1}{6^{2m}(2m)!}(-\sqrt{\hbar}V(y,y,y))^{2m}\, d^n y,
\end{aligned}$$

the odd-order terms are missing from the above expression since their integrals evaluate to 0. A typical term in this sum may be evaluated by the trick described in Exercise 15.1 by diagonalizing the quadratic form Q or generalizing the trick to higher-dimensional Gaussian integrals (see Sawon [136] and Etingof [54]). The result has the form

$$\hbar^{\frac{n}{2}} \frac{(2\pi)^{\frac{n}{2}}}{(\det Q)^{\frac{1}{2}}} \quad \frac{\hbar^m}{6^{2m}(2m)!} \sum_{\sigma} W_\sigma, \tag{20}$$

where σ represents a partition of the set $\{1,\dots,6m\}$ into two-element subsets encoded as a permutation on $1,\dots,6m$ and the W_σ have the form

$$\sum_{j_1=1}^{n} \cdots \sum_{j_{6m}=1}^{n} \prod_{k=0}^{2m-1} V(e_{j_{3k+1}}, e_{j_{3k+2}}, e_{j_{3k+3}}) \prod_{k=0}^{3m-1} Q^{-1}(e_{j_{\sigma(2k+1)}}, e_{j_{\sigma(2k+2)}}). \tag{21}$$

with $e_1,\dots,e_n$ being the standard basis in $\mathbb{R}^n$. For example, when $m = 1$

$$\sum_{i_1=1}^{n} \cdots \sum_{i_6=1}^{n} V(e_{i_1}, e_{i_2}, e_{i_3}) V(e_{i_4}, e_{i_5}, e_{i_6}) Q^{-1}(e_{i_1}, e_{i_2}) Q^{-1}(e_{i_3}, e_{i_6}) Q^{-1}(e_{i_4}, e_{i_5}) \tag{22}$$

is a typical term. Terms with $Q^{-1}(e_{i_1}, e_{i_2})$ and $Q^{-1}(e_{i_2}, e_{i_1})$ are not distinguished, so in total we get $\frac{1}{(3m)!}\binom{6m}{2,\dots,2} = \frac{(6m)!}{(3m)!2^{3m}}$ summands. This is cumbersome. Fortunately, there is a way due to Feynman to represent such terms diagrammatically.

We construct a trivalent graph Γ for each term W_σ with a vertex for each V in (21) and an edge for each Q^{-1}. The edge labeled $Q^{-1}(e_a, e_b)$ will connect to the vertex or vertices containing a or b. The graph corresponding to example (22) is then represented by the graph (Feynman diagram) in Figure 15.1. It is easy to see that different summands in (20) give the same contributions as long as they have isomorphic graphs. So instead of summing over all partitions σ as in (20) we may sum over graphs as long as we factor in the number of different partitions associated with each graph properly. Counting the number of partitions associated to a given trivalent graph is an elementary combinatorial problem with the answer (see Etingof [54])

$$\#\text{ partitions associated to } \Gamma = \frac{6^{2m}(2m)!}{|\mathrm{Aut}(\Gamma)|}, \tag{23}$$

where $\mathrm{Aut}(\Gamma)$ is the automorphism group of the graph. Here we view a graph as a one-dimensional complex and an automorphism must restrict to a linear map on each edge.

Example 15.2 There are 15 partitions of $1, \dots, 6$ into two-element sets, so the sum in (20) for $m = 1$ would have 15 terms. However, there are exactly two trivalent graphs with $2m = 2$ vertices (see Figures 15.1 and 15.2) so the corresponding sum over graphs would only have two terms. The automorphism groups of the graphs in Figures 15.1 and 15.2 have orders 8 and 12 respectively.

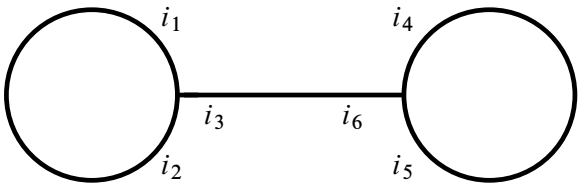


Figure 15.1: Glasses

Notice that

$$\chi(\Gamma) := \#\text{ vertices}(\Gamma) - \#\text{ edges}(\Gamma) = 2m - 3m = -m.$$

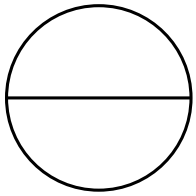


Figure 15.2: Theta

Thus the partition function Z can be rewritten as

$$Z = \frac{(2\pi\hbar)^{n/2}}{(\det Q)^{1/2}} \sum_{\Gamma\in\Xi} \frac{\hbar^{-\chi(\Gamma)}}{|\mathrm{Aut}(\Gamma)|} W(\Gamma), \tag{24}$$

where the sum is taken over the set of all trivalent graphs (including disconnected ones) Ξ and $W(\Gamma)$ is the contribution or *Feynman amplitude* W_σ of any partition with graph Γ.

The sum

$$F = \sum_{\Gamma\in\Xi'} \frac{\hbar^{-\chi(\Gamma)}}{|\mathrm{Aut}(\Gamma)|} W(\Gamma),$$

taken over the set of all connected trivalent graphs Ξ' is called the *free energy*.

Exercise 15.3 Show that if $\Gamma_1\Gamma_2$ represents the disjoint union of Γ_1 and Γ_2 then one has

$$\begin{aligned} W(\Gamma_1\Gamma_2) &= W(\Gamma_1)W(\Gamma_2), \\ \chi(\Gamma_1\Gamma_2) &= \chi(\Gamma_1) + \chi(\Gamma_2) \\ |\,\mathrm{Aut}(\Gamma_1^{n_1}\dots\Gamma_e^{n_e})| &= \prod_{i=1}^{e} |\,\mathrm{Aut}(\Gamma_i)|^{n_i}(n_i)!. \end{aligned}$$

Use exponentiation and series expansion to conclude that $F = \ln(Z/Z_0)$ where $Z_0 = (2\pi\hbar)^{n/2}/(\det Q)^{1/2}$.

We conclude that it will be helpful to consider the natural logarithm of the exact Chern–Simons invariants.

Of course the notions of partition function and free energy were introduced in the field of statistical thermodynamics before they ever appeared in Quantum Field Theory (see Schrödinger [137]). In statistical mechanics, the probability that a state at energy E in a given system will be occupied is given by $e^{-E/(kT)}$. The partition function in this

context is defined to be the following integral over phase space: $Z = \int_{T^*Q} e^{-\frac{H}{kT}}\, d\mathrm{vol}$, where H is the Hamiltonian. In this context the free energy is defined by $F = -kT \ln Z$.

The finite-dimensional analogy may be taken further and used to motivate definitions of similar invariants of 3–manifolds and framed links in 3–manifolds from Chern–Simons theory. The common feature of all of these invariants is that they are expressed as sums over graphs analogous to equation (24). Thus it makes sense to introduce the free algebra generated by all trivalent graphs with multiplication being disjoint union as in Exercise 15.3. Two different combinations of graphs can have the same contribution so we introduce relations in the algebra to identify such combinations.

Example 15.4 One such relation is the IHX relation displayed in Figure 15.3. The meaning of this relation is that the contribution of any graph containing the piece on the left can be replaced by the contribution of the difference of the two graphs containing the pieces on the right. As Figure 15.3 shows this relation implies that the contribution of the 'glasses' graph is trivial.

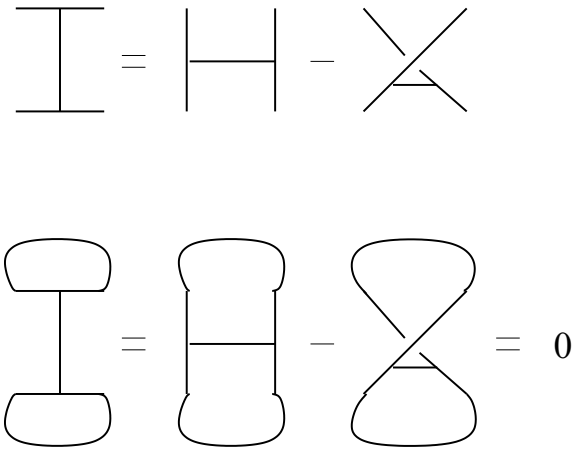


Figure 15.3: The IHX relation

It is natural to consider the quotient algebra by these relations to reduce the expression for the invariants even further. In fact this sum can be taken over a basis for the quotient algebra. The resulting expression for the free energy takes the form

$$F(M) = \sum_{\Gamma \in B} \hbar^{-\chi(\Gamma)} W^G(\Gamma)\, \mathrm{FA}(M, \Gamma), \tag{25}$$

where $\mathcal{W}^G$ is a homomorphism from the algebra of graphs to the complex numbers called a weight system, B is a basis for the algebra of graphs and $\mathrm{FA}(M,\Gamma)$ is the Feynman amplitude associated to the graph and the 3–manifold.

We are particularly interested in the weight system for $U(N)$ (see Bar-Natan [22]). This weight system applied to a graph can be computed as a sum over labelings. Given a graph Γ label each of the vertices with 0 or 1, then fatten the graph according to the rules in Figure 15.4. The graph then turns into what is called a 'fat graph' which

Figure 15.4: Fat Graphs

topologically represents a Riemann surface with boundary. Let g be the genus of the surface and h the number of boundary components. Also, let ℓ denote a labeling of the graph, $\Gamma_{(\ell)}$ the labeled graph, $\Lambda_{(\ell)}$ its fattened version, and $|\ell|$ the sum of all labels in ℓ; then

$$\mathcal{W}^{U(N)}(\Gamma)=\sum_{\ell}(-1)^{|\ell|}N^{h(\Lambda_{(\ell)})}. \tag{26}$$

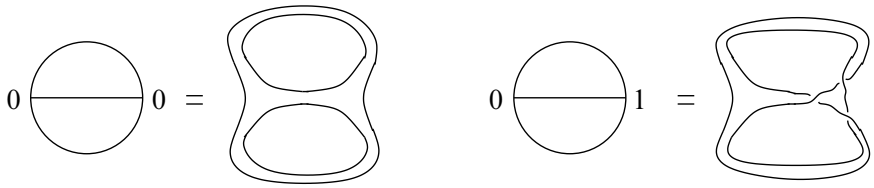


Figure 15.5: Fat theta

Example 15.5 To compute the $\mathrm{U}(N)$ weight of the 'theta' graph consider the fat graphs corresponding to the labels $(0,0)$ and $(0,1)$ displayed in Figure 15.5. We see that the $(0,0)$ labeled graph has genus zero and three boundary components, and the $(0,1)$ labeled graph has genus one and one boundary component. The $(1,1)$ and $(1,0)$ labeled graphs can be constructed similarly. Thus, for the 'theta' graph $\mathcal{W}^{U(N)}(\Gamma)=N^3-N-N+N^3=2N^3-2N$.

Since the Euler characteristic is a homotopy invariant, $2g-2+h=-\chi(\Lambda_\ell)=-\chi(\Gamma)$ for any labeling. Combining equations (25) and (26) one sees that the $U(N)$ Chern–Simons free energy has the form

$$F_M^{\mathrm{CS}}=\sum_{g=0}^{\infty}\sum_{h=1}^{\infty}x^{2g-2+h}N^h F_{g,h}(M)\,.$$

Here x is playing the role of $\hbar$ in our finite-dimensional Gaussian integral. The free energy takes the form of the free energy of an open string theory. One expects that there is such a theory of 'instantons at infinity' (degenerate curves) but there is no mathematical definition of this theory.

The fat graphs lead to closed surfaces as explained/conjectured by 't Hooft and these in turn lead to the J–holomorphic curves that are counted on the Gromov–Witten side. See Ooguri–Vafa [121] for this idea applied to Chern–Simons theory. 't Hooft suggested to 'sum over all holes' in this sum to obtain a 'closed string' expansion [148]. This means introducing a new parameter $t=xN$ and combining all summands with like powers of h. Denoting

$$F_g(M)=\sum_{h=1}^{\infty}t^h F_{g,h}(M)\,,$$

we obtain

$$F_M^{\mathrm{CS}}=\sum_{g=0}^{\infty}x^{2g-2}F_g(M)\,.$$

> This expression for the Chern–Simons free energy has the structure of the free energy of a closed string theory and is one reason to believe that there may be some relationship between string theory and Chern–Simons invariants.

We explain the structure of the free energy of a closed string theory in more detail at the end of this section. On the string theory side, the 'instantons at infinity' live in the cotangent bundle to the 3–manifold and are open strings. The cotangent bundle then undergoes a geometric transition where the boundaries of the open strings are collapsed to points giving closed strings on the manifold on the other side of the transition. The manifold across the transition from the cotangent bundle to S^3 is the resolved conifold. Thus, one expects that the Chern–Simons free energy is the same as the free energy of a closed string theory on the resolved conifold. There is a mathematically defined closed string theory on the resolved conifold, namely the Gromov–Witten theory. Identifying

it as the correct dual theory completes the physical derivation of the duality and was the major contribution of Gopakumar and Vafa [65], see also Ooguri–Vafa [121].

> To summarize, the first step is to describe the Chern–Simons invariants via fat graphs that should be considered as open strings. The second step is to sum over the 'holes' via a geometric transition to obtain a closed string theory.

If the expression for F_M^{CS} is rewritten using N as a parameter in place of $x = tN^{-1}$ one obtains a $(1/N)$–expansion and the $g = 0$ terms will dominate for large N. When the gauge-string duality holds for the leading terms (genus zero contributions) it is said to hold in the large N limit. This was the case originally studied by 't Hooft. Witten realized that the duality would be exact without considering the large N limit for topological (metric independent) field theories, see Witten [161].

The physical ideas here have been encoded into various mathematically defined perturbative Chern–Simons invariants [22]. Kontsevich defined a universal Vassiliev invariant for links taking the form of a rational linear combination of trivalent graphs in an algebra generated by trivalent graphs with a few simple relations, [90; 88]. See Bar-Natan [24] for a good time reading about these invariants and see [22] by the same author for a more typical overview. Schematically this invariant takes the form

$$Z(L) = \sum_{m=0}^{\infty} \sum_{(z_1,z_1')\ldots,(z_m,z_m')} E((z_1,z_1'),\ldots,(z_m,z_m');L)\Gamma((z_1,z_1'),\ldots,(z_m,z_m');L),$$

where $(z_1,z_1'),\ldots,(z_m,z_m')$ are m pairs of points on the link, and

$$\Gamma((z_1,z_1'),\ldots,(z_m,z_m');L)$$

is the chord diagram representing the locations of these points on the circles in L. Furthermore, $E((z_1,z_1'),\ldots,(z_m,z_m');L)$ is some expression computed as an integral or via intersection theory from the link L and pairs of the points. To get numerical invariants one just applies an algebra homomorphism from the algebra of chord diagrams to the complex numbers. Such homomorphisms are called weight systems [22].

There is a similar set of 3–manifold invariants. A universal 3–manifold invariant of this type (now called the LMO invariant) was introduced by Le, Murakami and Ohtsuki [93], and expressed like the Feynman expansion of a Gaussian integral in [26] by Bar-Natan, Garoufalidis, Rozansky and Thurston. As in the case of links, the universal

3–manifold invariant is a weighted sum of graphs. This invariant is an element in an algebra obtained as a quotient of the algebra freely generated by all trivalent graphs.

It is interesting to see how string theory considerations suggest that the Gromov–Witten free energy will take the form $F^{\mathrm{GW}} = \sum F_g y^{2g-2}$. The action for a simple model of a vibrating string is

$$S = \frac{1}{2} \iint u_{tt} - u_{ss}\, ds\, dt\,.$$

Notice the close similarity between this action and the Dirichlet functional

$$D = \frac{1}{2} \iint u_{tt} + u_{ss}\, ds\, dt\,.$$

As we saw in Exercise 9.3, the minima of the Dirichlet functional are exactly the J–holomorphic maps. One common feature of all string theories is the existence of an internal degree of freedom s in addition to the time t appearing in the action. This means that one must consider collections of surfaces in string theory where one considered paths in ordinary field theory.

To discretize a path one just subdivides the interval. To develop a discrete model of surfaces it is natural to triangulate the surfaces. Notice that the dual graph of a triangulation is a 3–valent graph similar to the Feynman diagrams encountered in path integrals. In fact neighborhoods of these dual graphs are the 'fat graphs' that we just discussed above. One can turn the process that we used in this subsection backwards and write out a partition function that would have these 'fat graphs' in its perturbative expansion. The most obvious candidate is the matrix integral

$$Z = \int e^{-N\mathrm{Tr}(\frac{1}{2}M^2 + wM^3)}\, dM\,,$$

where the integral is taken over the space of all $N \times N$ Hermitian matrices.

When one performs a perturbative expansion on this partition function the important things to notice about each term in the expansion are

(1) Each vertex contributes a factor of wN.

(2) Each edge corresponds to a propagator and contributes a factor of N^{-1}.

(3) Each face (of the dual complex to the triangulation) contributes a factor of N (for the sum of the indices).

It follows that each term in the expansion has order $w^V N^{V-E+F} = w^V N^{2-2g}$. Thus the free energy is also a sum of terms of order $w^V N^{V-E+F} = w^V N^{2-2g}$ because it

is just the sum of the contributions from the connected graphs. It follows that the free energy can be written as

$$F = \sum_g F_g N^{2-2g}.$$

Let us summarize the main conclusions of this section. It is natural to package Chern–Simons perturbative invariants into a formal partition function, then take the natural logarithm, introduce a new variable $t = xN$ and expand into a power series in x. Furthermore, this function called the free energy will take the form $F = \sum F_g x^{2g-2}$ identical to the form of the free energy in any closed string theory.

In the next section we review physical motivations behind Topological Quantum Field Theory (TQFT) and describe in detail a simplified version of it that can be nicely packaged into the language of ribbon and modular categories invented by N Reshetikhin and V Turaev. For a particular choice of categories coming from the theory of quantum groups these constructions produce the celebrated Reshetikhin–Turaev or quantum invariants of framed links and 3–manifolds that can be seen as a mathematical formalization of Witten's Chern–Simons path integrals.

16 Modular categories and topological invariants

Chern–Simons theory is a special type of quantum field theory called a topological quantum field theory (TQFT). Just as classical mechanics may be described in the Lagrangian or Hamiltonian frameworks any quantum field theory may be described in these two frameworks. The Lagrangian framework leads to path integrals and the Hamiltonian approach leads to canonical quantization (see Simms and Woodhouse [142]). We have already discussed the Lagrangian approach and the resulting perturbative Chern–Simons invariants. We will use the Hamiltonian framework for formal definitions and computations that we do from here on.

16.1 The Hamiltonian approach to TQFT

In the Hamiltonian approach, one begins with a symplectic manifold called the phase space. This is the collection of all positions and momenta. The mechanical system is specified by specifying an energy or Hamiltonian function denoted by H on this space. The evolution of the mechanical system is given by Hamilton's equations (see Arnol'd [13]). A quantization of a classical mechanical system is a map from the space of smooth functions on the phase space to the Hermitian operators on a Hilbert space. The relation between the Hamiltonian and Lagrangian is given by

$$\langle \phi_0 | e^{itH} | \phi_1 \rangle = \int_{\phi(0)=\phi_0, \phi(1)=\phi_1} e^{iL(\phi)} \mathcal{D}\phi,$$

Here $|\phi_k\rangle$ are elements of the Hilbert space, $\langle\phi_k|$ are the functionals obtained from the elements via the pairing on the Hilbert space (bras and kets), H is the Hamiltonian, L is the Lagrangian and the right hand side is a path integral (that is, the formal integral over the space of paths).

One can reinterpret this last expression by calling $|\phi_1\rangle$ a state and considering the Hilbert space as the space of states. The operator e^{itH} just represents evolution through t units of time. Then one sees that the last displayed expression expresses the evolution of the state $|\phi_1\rangle$ through time as a path integral.

Geometrically this suggests considering manifolds with two boundary components and associating a Hilbert space to each boundary component. The Hilbert space associated to one boundary component corresponds to the initial states and the Hilbert space associated to the other component corresponds to the states after evolving through time. The path integral should give an operator from the Hilbert space associated to one boundary component to the Hilbert space associated to the other boundary component. Gluing two manifolds along a common boundary as in Figure 16.1 corresponds to taking the composition of the corresponding operators. The Hilbert space associated to an empty boundary should just be $\mathbb{C}$; thus operators corresponding to closed manifolds can be interpreted as complex numbers.

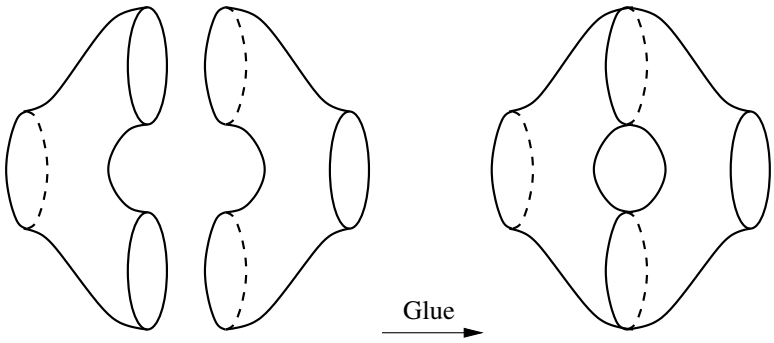


Figure 16.1: Gluing for topological quantum field theories

Formalizing these ideas leads one to the notion of a topological quantum field theory. Slices of 3–manifolds have 2–dimensional boundary components. Here slice means cobordism (that is, a 3–manifold with boundary components $\Sigma_\pm$). One views a product cobordism $[0,1]\times\Sigma$ as a space-time with two spatial dimensions. Hence such theories are called (2+1)–dimensional TQFT's. A (2+1)–dimensional TQFT associates a Hilbert space $\mathcal{H}_\Sigma$ to any closed surface and a bounded linear map, $\mathcal{H}_{\Sigma_-}\to\mathcal{H}_{\Sigma_+}$ to every cobordism. This formalism was suggested by Segal [138; 140; 139] and axiomatized by Atiyah [14].

Aside 16.1 Recall that the TQFT approach has origins in the Hamiltonian framework. This is a brief outline of the Hamiltonian description of Chern–Simons theory. In Chern–Simons theory the Lagrangian is just the Chern–Simons functional and the corresponding Hamiltonian is zero. The phase space used for the Hamiltonian description of Chern–Simons theory is the space of flat connections modulo gauge equivalence over a surface Σ of genus g. Equivalently this is the space of all connections modded out by the complexified gauge group (the geometric invariant theory picture) or the symplectic quotient of the space of connections by the gauge group (the symplectic reduction picture). Denote this moduli space of flat connections by $\mathcal{M}_\Sigma$.

In order to construct the associated Hilbert space for Chern–Simons theory, we need to introduce a line bundle over the moduli space $\mathcal{M}_\Sigma$. Quillen's determinant line bundle is the complex line bundle over $\mathcal{M}_\Sigma$ with fiber

$$\mathcal{L}_A = \bigwedge^{\text{top}} (\ker(d_A))^* \otimes \bigwedge^{\text{top}} (\operatorname{coker}(d_A)),$$

where $d_A\colon \Omega^{0,0}(\Sigma, E) \to \Omega^{0,1}(\Sigma, E)$ is the covariant derivative associated to the flat connection A on the bundle E. The associated Hilbert space is then $\mathcal{H}_\Sigma := H^0(\mathcal{M}_\Sigma, \mathcal{L}^{\otimes k+N})$. One should note that sometimes the quantity $k+N$ is called the level and is sometimes denoted by k. For general Lie groups the two different notions of level are related by the so-called dual Coxeter number. In this paper we will always use k to be the level as used in the definition of the string coupling constant. We discuss this in greater detail in Appendix E. More information on the above description of Witten–Chern–Simons theory may be found in Axelrod–Della Pietra–Witten [19], Hu [77], Di Francesco–Mathieu–Sénéchal [45] and Kohno [86].

While the motivation for this approach is fairly straightforward, formally constructing invariants in this manner is very complicated. There is an alternative approach that is more difficult to motivate but slightly easier technically.

16.2 Link invariants in a $U(1)$ theory

Instead of cutting 3–manifolds into cobordisms we use the fact that every 3–manifold can be expressed as surgery on a framed link to reduce our considerations to framed links. Just as every 3–manifold can be cut into cobordisms every framed link can be cut into elementary framed tangles. This is easier to draw and conceptualize.

To pass from a TQFT describing 3–manifold invariants to a corresponding theory for framed tangles the notion of a $(2+1)$–dimensional TQFT was enhanced by Reshetikhin, Turaev and many others [128]. One includes Wilson loops (framed links) into the manifolds. Formalizing the entire picture with Wilson loops in general is fairly complicated because one has to consider framed tangles in general 3–manifolds with surfaces of arbitrary genus as boundaries.

It is easier to formalize this picture for the special case of framed tangles in $S^2 \times [0,1]$. These are usually viewed as framed tangles in $\mathbb{R}^2 \times [0,1]$ as in Figure 16.2. This proves to be sufficient because any framed link can be obtained by stacking such tangles and any 3–manifold can be obtained by surgery on a framed link.

Before describing the general case in the next section we are going to introduce the cutting and pasting idea in this section in the setting of $\mathrm{U}(1)$ Chern–Simons theory. In the $\mathrm{U}(1)$ theory the cubic term in the Chern–Simons action vanishes because purely imaginary 1–forms anti-commute. Thus one expects considerable simplification in the $\mathrm{U}(1)$ case.

According to the discussion on topological quantum field theories with framed tangles, one should be able to cut a framed link into elementary pieces and associate invariants to each piece. This is indeed the case. Any framed link can be cut into cups, caps, right crossings, left crossings and twists (see Figure 16.2 for an example and Section 16.4 for formal definitions).

The simple $\mathrm{U}(1)$ theory at level $2m+1$ produces an invariant of colored, oriented framed links. Appropriately drawn link diagrams can be oriented by putting an upward pointing arrow on the left branch of each cup. Right and left crossings are then defined so that the crossings in the simple diagram for the left Hopf link from Figure 16.2 are left crossings and the opposite crossings are right crossings. Left and right refer to the direction of the strand that goes under the crossing. Coloring is captured by labeling every cup with a number p from $\mathbb{Z}_{2m+1}$ subject to a compatibility condition. The evaluation rule is very simple. In the obvious analogy with Feynman rules, every

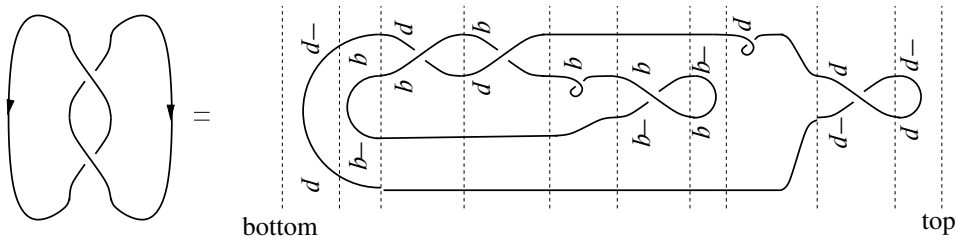


Figure 16.2: The left Hopf link

p–labeled cup on a framed link creates a pair $\pm p$ with p on the left and $-p$ on the right. The rest of the link receives labels by extension once we choose a 'color' p for each cup (see Figure 16.2). The compatibility condition is that every cap annihilates such a $\pm p$ pair. Once the link is colored, the $\mathrm{U}(1)$ theory at level $2m+1$ associates the following numbers to elementary pieces:

$\mid$	$=$	1	id	$\times_{p,q}$	$=$	$e^{2\pi ipq/(2m+1)}$	right crossing
$\cup$	$=$	1	cup	$\times^{-1}_{p,q}$	$=$	$e^{-2\pi ipq/(2m+1)}$	left crossing
$\cap$	$=$	1	cap	θ_p	$=$	$e^{2\pi ip^2/(2m+1)}$	twist

The colors p, q can be any natural numbers from 0 to $2m$. The final colored invariant $J_{p_1,\dots,p_l}(L)$ of the framed link is just the product of all numbers assigned to the pieces. In analogy to Feynman rules, the quantum invariant $F(L)$ is just the sum of the colored Jones polynomials over all possible ways to color the link.

Example 16.2 Consider the left Hopf link L from Figure 16.2 with the larger component colored by p and the smaller one colored by q. Multiplying from the bottom up we compute

$$J_{p,q}(L) = \times^{-1}_{p,q} \cdot \times^{-1}_{q,p} \cdot \theta_q \cdot \times_{q,-q} \cdot \theta_p \cdot \times_{p,-p} = e^{-4\pi ipq/(2m+1)}\,.$$

Therefore, $F(L) = \sum_{p,q=0}^{2m} e^{-4\pi ipq/(2m+1)} = 2m+1$.

The 'color and multiply' rule is not very sophisticated. In particular, it implies that any two links with the same elementary pieces have the same invariant no matter how those pieces are assembled. Of course, this makes it easier to prove independence of the presentation of a link by a regular projection but it also yields rather weak invariants. We want evaluation rules that are still independent of the projection but 'see' much more structure of the link. The right balance for link invariants is struck in the notion of ribbon categories. There colors are replaced by simple objects (think irreducible representations), numbers by morphisms (think linear maps) and multiplication by composition and tensor product. Accordingly, the evaluation rules become more involved.

To get invariants of 3–manifolds one takes weighted sums of colored invariants such as the one computed in the previous example. The hard part is to make sure that the final answer does not depend on how the manifold is expressed as a framed link and how the framed link is decomposed into elementary tangles. The required axioms are formalized in the notion of a strict modular category. In the rest of this subsection we introduce ribbon categories first, then discuss modular categories and evaluation rules, and finally the quantum invariants that arise from these categories.

16.3 Ribbon categories

As explained in the previous subsection the invariants that we are considering are naturally defined for framed tangles in $\mathbb{R}^2 \times [0,1]$ and satisfy formal gluing axioms.

We will follow the version of these invariants due to Reshetikhin and Turaev [129] as explained by Bakalov and Kirillov [21] and Turaev [152].

It is important to keep the basic idea in mind. Every 3–manifold may be expressed as surgery on a framed link.

> Associate a simple invariant to each elementary piece of a framed link diagram and define the final invariant to be an algebraic combination of all of the elementary pieces. Of course, one must know that different ways to assemble the same manifold give the same final invariant.

Axiomatizing assembly rules with outcomes independent of a link presentation leads to the notion of modular categories.

The basic outline is as follows. Any framed link may be constructed out of elementary building blocks that look like $\times_{U,V}$, θ_V, $\cap_V$ and $\cup_V$ from Figure 16.3. The trick is to associate an invariant to each of these and then write out all of the axioms that correspond to changing the height function isotopy of the link, or Kirby moves. In particular, one colors and orients the link. To each generic horizontal line one associates a tensor product of objects and their duals according to the sign of the intersection of the link with the horizontal line. The elementary building blocks between the horizontal lines induce morphisms between the associated objects. We use the standard convention that combining elementary building blocks side-by-side corresponds to taking the tensor product of the associated morphisms.

There are many types of categories related to modular categories. It is helpful to consider a related category with an easier definition first. The categories related to framed link invariants (as opposed to 3–manifold invariants) are ribbon categories.

Definition 16.3 A strict ribbon category is a category with a unit object $\mathbb{1}$, tensor product functor $\otimes$, families of isomorphisms $\times$, θ called the braiding and the twist respectively, and a duality triple $(*, \cup, \cap)$ satisfying axioms 1 through 12 under 16.5 below.

Remark 16.4 We are slightly changing notation from Turaev [152] and other references. The correspondence is $\times_{U,V} = c_{U,V}$, $\theta_V = \theta_V$, $\cup_V = b_V$ and $\cap_V = d_V$. In fact, our notation is often used to represent objects in a colored ribbon category. There is a natural functor taking the colored ribbon morphisms to morphisms in a strict modular category. We feel that no confusion will arise by using the same notation for both, and the ribbon notation is more descriptive.

Remark 16.5 One can obtain a new strict ribbon category by replacing the braiding and twist by their inverses.

The first example of a strict ribbon category is the category of representations of a group.

Example 16.6 Let REP_G be the category of representations of a Lie group. The objects are representations $\rho\colon G \to \mathrm{Aut}(V)$. The unit object is the trivial representation. Morphisms are equivariant linear maps $f(gv) = gf(v)$. The dual representation V^* has the dual space to V as the representation space and the action is given by $(g\varphi)(v) := \varphi(g^{-1}v)$. The pairing $\cup$ is the standard duality pairing for vector spaces and $\cap$ the copairing given by $1 \mapsto \sum e_k \otimes e^k$ where $\{e_k\}$ and $\{e^k\}$ are dual bases. The image of 1 under the copairing is sometimes called the Casimir element. The tensor product is the standard one in the category of finite-dimensional vector spaces with action given by tensor product of actions as described in Appendix D. The braiding is given by $\times_{V,W}(v \otimes w) = w \otimes v$, and the twist is given by $\theta_V = \mathrm{id}_V$. As is, this is not a strict ribbon category because equality signs in the axioms are only canonical isomorphisms. For example V is not equal to $V \otimes \mathbb{C}$. In truth, we should be talking about equivalence classes of representations rather than representations themselves. This requires some tweaking in the notion of morphisms and ribbon operations that can be done in a standard way by the Mac Lane coherence theorem [100].

The category of representations cannot be used to construct nontrivial link invariants following the procedure given in the next article because the braiding is its own inverse (this category cannot detect the difference between right and left crossings.) The nontrivial invariants described in the previous subsection can be seen to arise from a strict ribbon category.

Example 16.7 Construct a category $\mathcal{U}(1)_{2m+1}$ with objects elements of $\mathbb{Z}_{2m+1}$, unit object equal to 0, only the zero morphism between unequal objects and the endomorphisms of any object equal to $\mathbb{C}$. The tensor product is given by

$$(f\colon p \to q) \otimes (g\colon r \to s) := fg\colon p + r \to q + s,$$

and the braiding given by

$$\times_{p,q} = e^{2\pi i pq/(2m+1)}\colon p + q \to q + p.$$

The twist is given by

$$\theta_p = e^{2\pi i p^2/(2m+1)}\colon p \to p.$$

The duality pairing is $\cap_p\colon\ p+(-p)\to 0$ and the copairing is $\cup_p = 1\colon\ 0\to p+(-p)$. One sees that this weird category is strict and satisfies all of the axioms for a strict ribbon category.

Exercise 16.8 Prove that in a strict ribbon category $\times_{V,\mathbb{1}} = \mathrm{id}_V = \times_{\mathbb{1},V}$ and the following Yang–Baxter equation holds

$$(\times_{V,W}\otimes \mathrm{id}_U)\circ(\mathrm{id}_V\otimes\times_{U,W})\circ(\times_{U,V}\otimes \mathrm{id}_W)$$
$$= (\mathrm{id}_W\otimes\times_{U,V})\circ(\times_{U,W}\otimes \mathrm{id}_V)\circ(\mathrm{id}_U\otimes\times_{V,W})\,.$$

There is a way to represent the axioms and other formulas with ribbon operations graphically making them much more intuitive. The correspondence between elementary

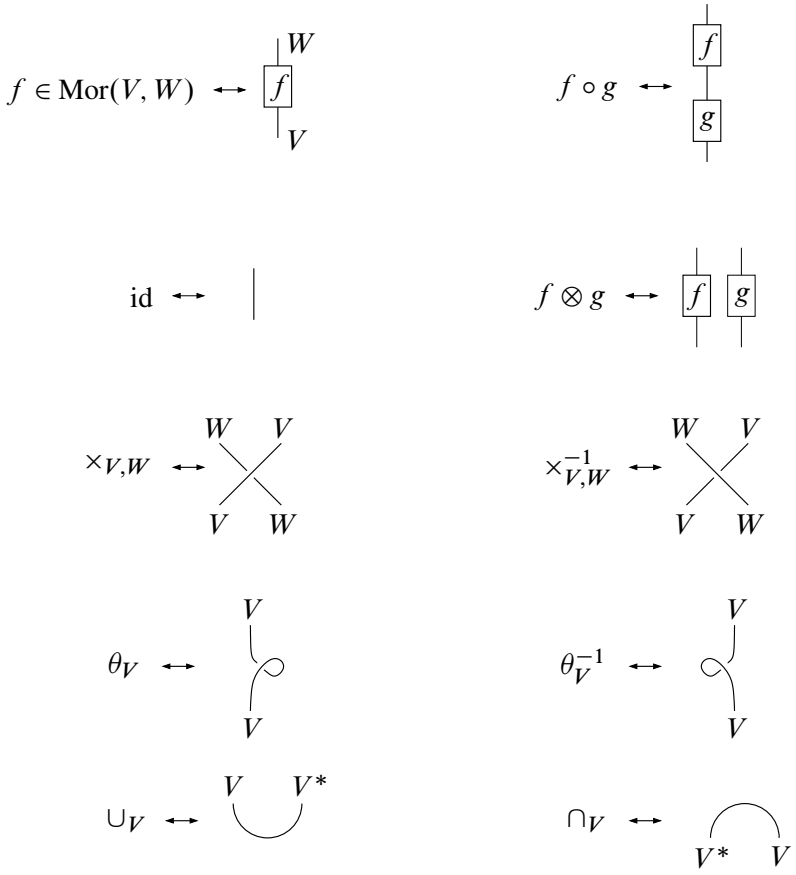


Figure 16.3: Strict ribbon category

operations and graphs is depicted on Figure 16.3. To draw the picture corresponding to a

formula start from the right (as in reading Arabic) and draw the pieces corresponding to each expression between two compositions on the same horizontal level while moving from the bottom up and putting the obtained pieces on top of each other. Conversely, given a labeled graph draw horizontal lines not intersecting any crossings, twists, maxima or minima (see Figure 16.2). Next write the expression for each elementary piece on the same level and tensor them. Finally assemble the expressions right to left by compositions moving from the bottom to the top of the graph. The graphs cannot be labeled arbitrarily; the labels on successive horizontal levels must match. If a ribbon expression produces a graph with mismatching labels it is nonsensical: in categorical language you would be trying to compose morphisms with targets and sources that do not match.

The main advantage of using graphs is that one can immediately see if two expressions in a ribbon category are equal.

> If two (correctly composed) expression graphs represent isotopic framed tangles the expressions are in fact equal and the equality can be established using the axioms; see Bakalov and Kirillov [21].

Framing is important here. Even though we draw pictures with strands one should actually think of them as very thin ribbons so that the twists (depicted as curls on the strands) cannot be undone. Later we will slightly enhance the notation to allow arrows on the strands but for now this will suffice.

The motivating example of a ribbon category is the category of ribbon tangles.

Example 16.9 The objects of the category of ribbon tangles are just non-negative integers, with zero representing the unit object. The morphisms between n and m are just the isotopy classes of framed tangles from n points to m points. Our graphs of expressions are just plane projections (with indication of under- and over-crossings and twists) of framed tangles with labels being natural numbers. The tensor product is the sum and the dual of n is n itself. The braiding, twist and duality for single strands are displayed in Figure 16.4; in general one just has to put n and m strands parallel to the one or two depicted. Invariants produced by this category are complete but useless: the invariant of an isotopy class is the isotopy class itself. Fortunately, there are more interesting examples.

The pictures for axioms 6 and 7 are displayed in Figure 16.5. These axioms correspond exactly to elementary isotopies. For example, a combination of axioms 6 and 7 implies the third Reidemeister move as shown in Figure 16.5.

Figure 16.4: Category of ribbon tangles

Axiom 6 Axiom 7

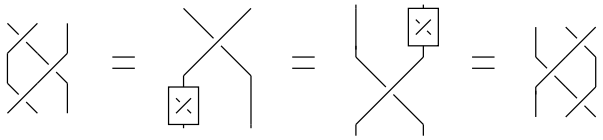


Figure 16.5: Axioms 6 and 7

Exercise 16.10 Draw diagrams representing each of the braiding, twist and duality axioms. Remember that it is standard to compose maps from the bottom up in a diagram.

16.4 Modular categories

While looking at axioms 1–12 you may have noticed that the remaining ones 13–17 have a different flavor. Rather than describing algebraic properties of operations they describe the global structure of a category. The idea is that the first twelve axioms take care of the Reidemeister moves and produce link invariants. Surgeries on isotopic links produce diffeomorphic 3–manifolds but diffeomorphic 3–manifolds are also produced by non-isotopic links related by Kirby moves. The invariants we are looking for must take the same values on such pairs of links. The Kirby move has a more complex structure than the Reidemeister moves and one requires the modular axioms to account for it.

> The point is that any ribbon category generates many framed link invariants. There is a framed link invariant for each object in the category or more generally one can define invariants of framed links colored by objects in the category. The hope is that an appropriate linear combination of the resulting framed link invariants will be invariant under the Kirby move and thus define a 3–manifold invariant.

Before stating the axioms of a strict modular category we need to define a few terms used in the axioms. Axiom 13 simply says that we can add morphisms with the same sources and targets and this addition behaves the same way as linear operators on vector spaces. Categories satisfying axiom 13 are called preabelian. Denote $\mathrm{End}(V) := \mathrm{Mor}(V, V)$ and notice that $(\mathrm{End}(\mathbb{1}), +, \circ)$ is a ring. Moreover, it is a commutative ring.

Exercise 16.11 Use $f \circ g = (f \otimes \mathrm{id}_{\mathbb{1}}) \circ (\mathrm{id}_{\mathbb{1}} \otimes g)$ to prove that $\mathrm{End}(\mathbb{1})$ is commutative based on axioms 1–13. Hint: recall that $\otimes$ is functorial.

Exercise 16.12 Prove that $\theta_{\mathbb{1}}^2 = \theta_{\mathbb{1}}$ and

$$\theta_{V \otimes W} = \times_{W,V} \circ \times_{V,W} \circ (\theta_V \otimes \theta_W)\,.$$

In the light of the above we will sometimes omit the composition sign between morphisms when composing them. The analogy with vector spaces goes further: one can define 'traces' of endomorphisms and 'dimensions' of objects.

Definition 16.13 The quantum trace of $f \in \mathrm{End}(V)$ is

$$\mathrm{Tr}_q(f) := \cap_V \circ \times_{V,V^*} \circ (\theta_V \otimes \mathrm{id}_V) \circ (f \otimes \mathrm{id}_V) \cup_V \in \mathrm{End}(\mathbb{1})\,.$$

The corresponding graph is shown in Figure 16.6. The quantum dimension of an object is $\dim_q(V) := \mathrm{Tr}_q(\mathrm{id}_V)$.

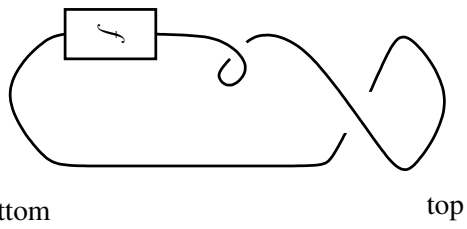


Figure 16.6: The quantum trace

Exercise 16.14 Prove that $\mathrm{Tr}_q(fg) = \mathrm{Tr}_q(gf)$ for all f and g where the compositions make sense, $\mathrm{Tr}_q(f) = f$ for all $f \in \mathrm{End}(\mathbb{1})$, and $\mathrm{Tr}_q(f \otimes g) = \mathrm{Tr}_q(f)\mathrm{Tr}_q(g)$ for all morphisms f and g. Conclude that for any pair of objects $\dim_q(V \otimes W) = \dim_q(V)\dim_q(W)$.

Example 16.15 In the category REP_G of Example 16.6 we have

$$\begin{aligned}\mathrm{Tr}_q(f) &= \cap_V \circ \times_{V,V^*} \circ (\theta_V \otimes \mathrm{id}_V) \circ (f \otimes \mathrm{id}_V)\Big(\sum\nolimits_k e^k \otimes e_k\Big)\\ &= \cap_V \circ \times_{V,V^*} \circ (\theta_V \otimes \mathrm{id}_V)\Big(\sum\nolimits_k f(e^k) \otimes e_k\Big)\\ &= \cap_V\Big(\sum\nolimits_k e_k \otimes f(e^k)\Big) = \sum\nolimits_k \langle e_k, f(e^k)\rangle = \mathrm{Tr}(f)\end{aligned}$$

so the quantum trace reduces to the ordinary one and $\dim_q(V) = \dim(V)$.

Exercise 16.16 In the category $\mathcal{U}(1)_{2m+1}$ of Example 16.7 show that $\mathrm{Tr}_q(f) = f$.

In preabelian categories one can define an analog of irreducible representations.

Definition 16.17 To say that an object V is *simple* means that the map $\mathrm{End}(\mathbb{1}) \to \mathrm{Mor}(V, V)$ given by $f \mapsto f \otimes \mathrm{id}_V$ is an isomorphism. To say that such a category is *dominated by* $\{V_\lambda\}_{\lambda \in I}$ means that for every $V \in \mathrm{Ob}(\mathcal{V})$ there are morphisms $f_r\colon V_{\lambda_r} \to V$ and $g_r\colon V \to V_{\lambda_r}$ such that $\mathrm{id}_V = \sum f_r \circ g_r$.

Remark 16.18 Domination is the preabelian analog of semi-simplicity in Abelian categories. The point is that while one can take sums of objects in Abelian categories one can only take sums of morphisms in preabelian categories.

The Schur lemma implies that in the category of representations of a Lie group REP_G the simple objects are the irreducible representations. However, any nontrivial Lie group has infinitely many simple objects. The category of finite-dimensional representations of a semi-simple Lie group is dominated by the simple objects because any representation decomposes into irreducible ones. In the category from Example 16.7 every object is simple and there are only finitely many objects. Clearly, this category is dominated by its simple objects.

The fact that $\mathrm{End}(\mathbb{1})$ is a commutative ring allows one to define the matrix with the following entries (to be used in axiom 17):

$$\widetilde{s}_{\lambda\mu} := \mathrm{Tr}_q(\times_{\mu,\lambda} \circ \times_{\lambda,\mu}) \tag{27}$$

From Exercise 16.14 we see that $\widetilde{s}$ is symmetric.

Strict modular category ingredients	Example
A category $\mathcal{V}$	$\mathrm{Ob}(\mathcal{U}(1)_{2m+1}) = \mathbb{Z}_{2m+1}$
	$\mathrm{Mor}(p,q) = \mathbb{C}$ if $p = q$ and 0 otherwise
A tensor product $\otimes\colon \mathcal{V}\times\mathcal{V} \Rightarrow \mathcal{V}$	$(f\colon p \to q) \otimes (g\colon r \to s) = fg\colon p + r \to q + s$
A unit $\mathbb{1} \in \mathrm{Ob}(\mathcal{V})$	$0 \in \mathbb{Z}_{2m+1}$
A braiding $\times_{U,V}\colon U \otimes V \to V \otimes U$	$\times_{p,q} = e^{2\pi i pq/(2m+1)}\colon p + q \to q + p$
A twist $\theta_V\colon V \to V$	$\theta_p = e^{2\pi i p^2/(2m+1)}\colon p \to p$
A duality pairing $\cap_V\colon V^* \otimes V \to \mathbb{1}$	$\cap_p\colon p + (-p) \to 0$
A copairing $\cup_V\colon \mathbb{1} \to V \otimes V^*$	$\cup_p = 1\colon 0 \to p + (-p)$
A finite collection of simple objects $\{V_\lambda\}_{\lambda\in I}$	$V_p = p$, $I = \mathbb{Z}_{2m+1}$

Table 16.1: Strict modular category ingredients

Exercise 16.19 Show that the $\widetilde{s}$–matrix defined in (27) is

$$\widetilde{s}_{\lambda\mu} = \cap_{\lambda\otimes\mu} \circ \times_{\lambda\otimes\mu,\mu^*\otimes\lambda^*} \circ (\theta_{\lambda\otimes\mu} \otimes \mathrm{id}_{\mu^*\otimes\lambda^*}) \circ (\times_{\mu\otimes\lambda} \otimes \mathrm{id}_{\mu^*\otimes\lambda^*}) \circ (\times_{\lambda\otimes\mu} \otimes \mathrm{id}_{\mu^*\otimes\lambda^*}) \circ \cup_{\lambda\otimes\mu} \in \mathrm{End}(\mathbb{1}).$$

A graphical representation of the $\widetilde{s}$–matrix is given by the right Hopf link as displayed in Figure 16.10.

We are now ready to define a strict modular category. We refer our readers to Bakalov and Kirillov [21] and to Turaev [152] for further description, properties and related definitions. The reference [21] uses modular tensor categories as opposed to strict modular categories. The difference is that the underlying category in a modular tensor category is Abelian rather than just preabelian, that is, direct sums of objects are defined. Even though our main example is in fact a modular tensor category, at this point it is expedient to ignore the additional structure.

Definition 16.20 A strict modular category is a category with a tensor product functor, natural isomorphisms $\times$, θ, a duality $(*, \cup, \cap)$ together with an indexed collection of special objects $\{V_\lambda\}_{\lambda\in I}$ satisfying (all 17 of) the axioms under Section 16.5 below.

In particular any strict modular category is a strict ribbon category. The ingredients in a strict modular category are listed in Table 16.1. This table also includes the ingredients of the $\mathcal{U}(1)_{2m+1}$ category from Example 16.7 as an illustration.

16.5 Axioms defining a strict modular category

Tensor axioms

Axiom 1: $V \otimes \mathbb{1} = \mathbb{1} \otimes V = V$.

Axiom 2: $U \otimes (V \otimes W) = (U \otimes V) \otimes W$.

Axiom 3: $f \otimes \mathrm{id}_{\mathbb{1}} = \mathrm{id}_{\mathbb{1}} \otimes f = f$.

Axiom 4: $f \otimes (g \otimes h) = (f \otimes g) \otimes h$.

Braiding axioms

Axiom 5: $\times_{U,V\otimes W} = (\mathrm{id}_V \otimes \times_{U,W}) \circ (\times_{U,V} \otimes \mathrm{id}_W)$.

Axiom 6: $\times_{U\otimes V,W} = (\times_{U,W} \otimes \mathrm{id}_V) \circ (\mathrm{id}_U \otimes \times_{V,W})$.

Axiom 7: $(g \otimes f) \circ \times_{U,W} = \times_{V,Z} \circ (f \otimes g)$ for any morphisms $f\colon U \to V$, $g\colon W \to Z$.

Twist axioms

Axiom 8: $\theta_{V\otimes W} = \times_{W,V} \circ \times_{V,W} \circ (\theta_V \otimes \theta_W)$.

Axiom 9: $f \circ \theta_U = \theta_V \circ f$ for any morphism $f\colon U \to V$.

Duality axioms

Axiom 10: $(\mathrm{id}_V \otimes \cap_V) \circ (\cup_V \otimes \mathrm{id}_V) = \mathrm{id}_V$.

Axiom 11: $(\cap_V \otimes \mathrm{id}_{V^*}) \circ (\mathrm{id}_{V^*} \otimes \cup_V) = \mathrm{id}_{V^*}$.

Axiom 12: $(\theta_V \otimes \mathrm{id}_{V^*}) \circ \cup_V = (\mathrm{id}_V \otimes \theta_{V^*}) \circ \cup_V$.

Modular axioms

Axiom 13: $\mathrm{Mor}(V, W)$ are abelian groups and $\circ\colon \mathrm{Mor}(V, W) \times \mathrm{Mor}(U, V) \to \mathrm{Mor}(U, W)$ is bilinear.

Axiom 14: $\mathcal{V}$ is dominated by a finite collection $\{V_\lambda\}_{\lambda\in I}$.

Axiom 15: There is $0 \in I$ such that $V_0 = \mathbb{1}$.

Axiom 16: For every $\lambda \in I$ there is a $\lambda^* \in I$ such that $V_{\lambda^*} \cong V_\lambda^*$.

Axiom 17: The matrix $\widetilde{s}_{\lambda,\mu} = \mathrm{Tr}_q(\times_{\mu,\lambda} \circ \times_{\lambda,\mu})$ is non-singular.

The tensor axioms are not difficult to understand if one keeps the example of vector spaces in mind in which case $\mathbb{1}$ is just the underlying field. Graphical representations help one understand the braiding, twist and duality axioms. As we emphacized in Section 16.3 they simply catalogue elementary transformations of tangle diagrams. In fact, one can forget about them and work with diagrams directly. Axioms 7 and 9 just restate the naturality of $\times$ and θ but we included them for the sake of diagrammatic interpretation. For instance, axiom 9 means that one can slide the twist through any morphism. The next example is intended to help illustrate the modular axioms.

Example 16.21 The category REP_G for an infinite group G fails to be strict modular for the trivial reason of having infinitely many simple objects (irreducible representations). But even when $|G| < \infty$ this category is not strict modular. It comes very close though: the only thing that goes wrong is the non-degeneracy axiom 17. Indeed, for any pair of objects (representation spaces) we have by Example 16.15:

$$\mathrm{Tr}_q(\times_{W,V} \circ \times_{V,W}) = \mathrm{Tr}_q(\mathrm{id}_{V\otimes W}) = \dim_q(V \otimes W) = \dim(V)\dim(W)$$

and the structure matrix $\widetilde{s}_{\lambda\mu} = \dim(V_\lambda)\dim(V_\mu)$ always has rank 1. We get non-degeneracy for $|G| = 1$ that is, the trivial group, but this is a rather trivial example.

On the other hand, the category $\mathcal{U}(1)_{2m+1}$ from Example 16.7 is both strict modular and nontrivial. In fact, every object is simple so it is definitely dominated by simple objects, and there are only $2m + 1 < \infty$ of them. By Example 16.15

$$\widetilde{s}_{pq} = \mathrm{Tr}_q(e^{2\pi ipq/(2m+1)} \circ e^{2\pi ipq/(2m+1)}) = e^{4\pi ipq/(2m+1)} = z^{pq}$$

with $z := e^{4\pi i/(2m+1)}$.

Exercise 16.22 Verify that the category $\mathcal{U}(1)_{2m+1}$ from Example 16.7 and the Table 16.1 satisfies the definition of a a strict modular category. Hint: notice that $\det z^{pq}$ is a Vandermonde determinant.

We extend the honor of being named a number to the elements of $\mathrm{End}(\mathbb{1})$ which is, after all, a commutative ring. In all examples of interest to us $\mathrm{End}(\mathbb{1}) = \mathbb{C}$ anyway.

Definition 16.23 The *characteristic numbers* of a strict modular category are given by $d_\lambda := \mathrm{Tr}_q(\mathrm{id}_\lambda)$, $p^\pm_\lambda := \mathrm{Tr}_q(\theta^{\pm 1}_\lambda)$, where λ indexes simple objects. The numbers $p^\pm := \sum_{\lambda\in I} p^\pm_\lambda = \sum_{\lambda\in I} \theta^{\pm 1}_\lambda d_\lambda$ are called the twists and $\mathcal{D} := \left(\sum d^2_\lambda\right)^{1/2}$ the quantum diameter (also rank or dimension; see Bruguières [35] and Müger [111]) of a category.

Remark 16.24 Note that the characteristic 'numbers' are defined in any ribbon category not just modular categories. The numbers $p^{\pm}, \mathcal{D}$ are instrumental in making sure that framed link invariants defined by a modular category are invariant under Kirby moves and hence define 3–manifold invariants. The square root we need to take to define $\mathcal{D}$ may not exist in the ring $\mathrm{End}(\mathbb{1})$ and, when it does may not be unique. It is possible to extend an arbitrary strict modular category so that this square root does exist and there is no important difference between choosing any of the two roots (see Turaev [152]). Numerical values of invariants do however depend on a choice of quantum diameter. In examples of interest to us $\mathrm{End}(\mathbb{1}) = \mathbb{C}$ and $\mathcal{D}^2$ is a positive real number so we agree to always choose the positive square root. For instance, in $\mathcal{U}(1)_{2m+1}$ there are $2m+1$ objects of dimension one each so $\mathcal{D} = (2m+1)^{1/2}$. For REP_G with G finite the sum of the squares of dimensions of irreducible representations is $|G|$ by the Burnside theorem (see Fulton and Harris [62]) so $\mathcal{D} = |G|^{1/2}$.

16.6 Coloring, double duals and the arrow convention

Before introducing invariants of links and 3–manifolds we augment the graphic notation introduced in Figure 16.4. You may notice that it is impossible to label a simple circle coherently with the rules we have so far because $\cup_V$ and $\cap_V$ put together have mismatching labels V, V^*. We can isotope the circle as in Figure 16.6 (without f) so that it can be labeled consistently. Isotoping links into shapes that allow labeling every time leads to cumbersome pictures as in Figure 16.2. Instead we can simply add a dual pairing and copairing (cups and caps) to the notation. Namely, define $\cup_V^*\colon \mathbb{1} \to V^* \otimes V$ and $\cap_V^*\colon V \otimes V^* \to \mathbb{1}$ by (see Figure 16.7):

$$\cup_V^* = (\theta_{V^*}^{-1} \otimes \mathrm{id}_V) \circ \times_{V,V^*}^{-1} \circ \cup_V$$
$$\cap_V^* = \cap_V \circ \times_{V,V^*} \circ (\theta_V \otimes \mathrm{id}_{V^*}).$$

Using the dual pairing and copairing we can re-express the quantum trace of $f\colon V \to V$

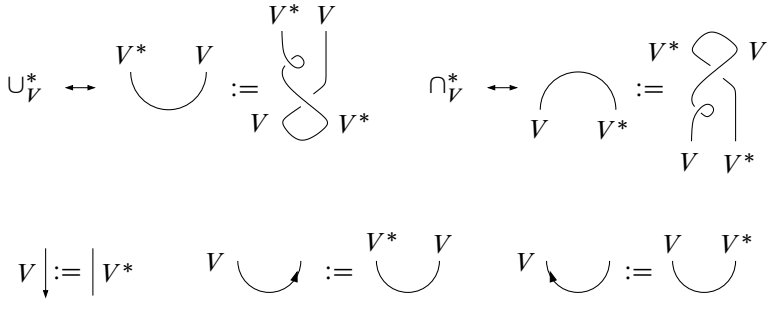


Figure 16.7: Dual cups and caps and the arrow convention

as $\mathrm{Tr}_q(f) = \cap_V^* \circ (f \otimes \mathrm{id}_{V^*}) \circ \cup_V$ and simplify all the definitions that use it.

Exercise 16.25 Show that $\cup_V^*$ and $\cap_V^*$ in the category $\mathcal{U}(1)_{2m+1}$ from Example 16.7 are multiplications by 1.

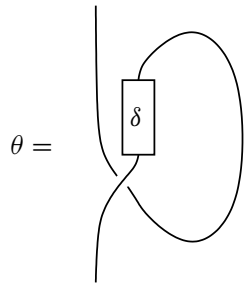


Figure 16.8: Twist from the double dual isomorphism

Using $\cup_V^*$ and $\cap_V^*$ we can label any diagram consistently with V and V^* only. The final simplification is to get rid of V^* as well by placing arrows on the strands and agreeing that an up arrow on a strand in a V–labeled component corresponds to the V label and a down arrow on the same component corresponds to the V^* label. With the arrow convention all we have to do in order to label an entire link or tangle is to orient every component and label it with an object in a single place, that is, *color* it. To evaluate a colored graph one simply has to transform it into the ribbon expression according to the rules on Figures 16.7 and 16.4 and tensor and compose all the morphisms.

Exercise 16.26 Show that for the category $\mathcal{U}(1)_{2m+1}$ from Example 16.7 this reduces to the 'color and multiply' rule of Section 16.2.

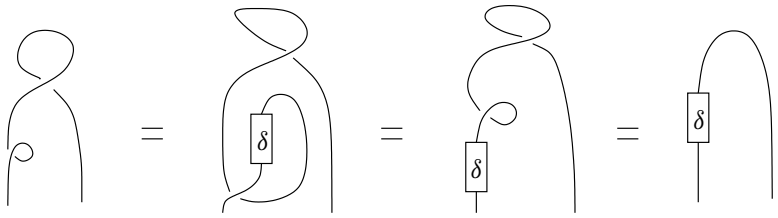


Figure 16.9: Dual pairing via the double dual isomorphism

One may wonder why double and higher duals in the category have not been discussed. What we are doing of course is implicitly identifying V^{**} with V. In any strict modular category it is possible to do this explicitly.

Exercise 16.27 Show that there exists a natural isomorphism $V \xrightarrow{\delta_V} V^{**}$ in any strict modular category such that the twist is given by Figure 16.8 or

$$\theta_V = (\mathrm{id}_{V^*} \otimes \cap_V) \circ (\mathrm{id}_V \otimes \delta_V \otimes \mathrm{id}_{V^*}) \circ (\times_{V,V} \otimes \mathrm{id}_{V^*}) \circ (\mathrm{id}_V \otimes \cup_V).$$

The right way to do this exercise is to draw a colored ribbon tangle representing a map from V to V^{**} built from elementary pieces. The answer is

$$\begin{aligned}\delta_V = (\cap_V \otimes \mathrm{id}_{V^{**}}) \circ (\times_{V,V^*} \otimes \mathrm{id}_{V^{**}}) \circ (\theta_V \otimes \mathrm{id}_{V^{**}}) \circ (\mathrm{id}_V \otimes \times_{V^{**},V^*}) \\ \circ (\mathrm{id}_V \otimes \theta^{-1}_{V^{**}} \otimes \mathrm{id}_{V^*}) \circ (\mathrm{id}_V \otimes \times_{V^*,V^{**}}) \circ (\mathrm{id}_V \otimes \cup_{V^*}) \circ \theta_V.\end{aligned}$$

We call δ from the exercise the double dual isomorphism. Using this isomorphism we can streamline the definitions of the dual pairing and copairing (see Figure 16.9)

$$\begin{aligned}\cup^*_V &= (\mathrm{id}_{V^*} \otimes \delta_V^{-1}) \circ \cup_{V^*} \\ \cap^*_V &= \cap_{V^*} \circ (\delta_V \otimes \mathrm{id}_{V^*}).\end{aligned}$$

The best part about knowing δ_V explicitly is that by using it one can evaluate a number of graphs without evaluating any braidings in the process. We we will take full advantage of this fact when discussing the modular categories coming from quantum groups because as in many other nontrivial modular categories, it is the braiding that is the hardest to compute. In particular, we get the following braiding-free formula for the quantum trace (Figure 16.6):

$$\mathrm{Tr}_q(f) := \cap_{V^*} \circ (\delta_V f \otimes \mathrm{id}_{V^*}) \circ \cup_V.$$

Exercise 16.28 Show that the double dual in $\mathcal{U}(1)_{2m+1}$ from Example 16.7 is the identity map $\mathbb{C} \to \mathbb{C}$ and the double dual in REP_G is the standard isomorphism between a vector space and its double dual $\delta_V(v)(\varphi) := \varphi(v)$.

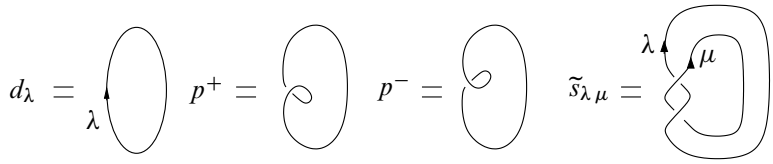


Figure 16.10: Characteristic numbers of a strict modular category

Our conventions thus far only allow evaluation of graphs that are completely colored (labeled). But usual framed links do not have representation labels attached to their

components. One can make sense of unlabeled graphs as well. The trick is to come up with evaluation rules such that the resulting expressions do not change at least under the Reidemeister moves (for link invariants) and even better, do not change under the Kirby moves (for 3–manifold invariants). There are many different ways to color a particular graph and the hope is that an appropriate linear combination of the resulting evaluations will be invariant under the appropriate moves. It turns out that there is an essentially unique way to evaluate unlabeled graphs or unlabeled closed components in a graph that guarantees invariance under the Kirby moves (see Turaev and Wenzl [153]). We therefore introduce the following important convention.

> In any diagram we sum over all ways of labeling unlabeled closed components by simple objects λ with multiplicity d_λ.

Examples using our extended graphic notation to visualize the definitions of the $\widetilde{s}$–matrix and the characteristic numbers of a modular category are shown on Figure 16.10. The next article explains in detail how to construct framed link invariants from strict ribbon categories and how to construct 3–manifold invariants from strict modular categories.

16.7 Invariants from modular categories

To see the correspondence between the Chern–Simons invariants and strict modular categories one needs to describe how to build 3–manifold invariants from a strict modular category. We first define the colored Jones polynomial of an oriented framed link L with $c(L)$ components. Color the components with objects $V_1, \dots, V_{c(L)}$ from the strict ribbon category. Notice that a diagram with only closed components represents the element of $\mathrm{End}(\mathbb{1})$ obtained by taking the composition of all of the basic morphisms corresponding to the elementary framed tangles occurring between horizontal lines as in Figure 16.2. The axioms of a strict modular category ensure that this element is invariant under all elementary isotopies of the link.

Definition 16.29 The colored Jones polynomial of an oriented framed link is the element $J_{V_1,\dots,V_{c(L)}}(L)$ corresponding to the morphism represented by the labeled link diagram.

Remark 16.30 This invariant is not a polynomial. The name comes from the fact that the invariant associated to the category of tilting modules is closely related to the classical Jones polynomial [80] (which by the way, is not a polynomial either).

Definition 16.31 Given a framed link (not oriented) L define the invariant

$$F(L) = \sum_{\lambda_1,\dots,\lambda_{c(L)} \in I} J_{\lambda_1,\dots,\lambda_{c(L)}}(L) d_{\lambda_1}, \dots, d_{\lambda_{c(L)}} .$$

Just as any isotopy of an ordinary link can be expressed as a composition of elementary isotopies (Reidemeister moves I, II and III) any isotopy of ribbon tangles with height function can be decomposed into elementary isotopies. According to Turaev [151], Freyd and Yetter [59], and Reshetikhin and Turaev [128], the elementary isotopies for ribbon tangles with height function are Reidemeister moves II and III (see Figure 13.1) together with the moves displayed in Figure 16.11.

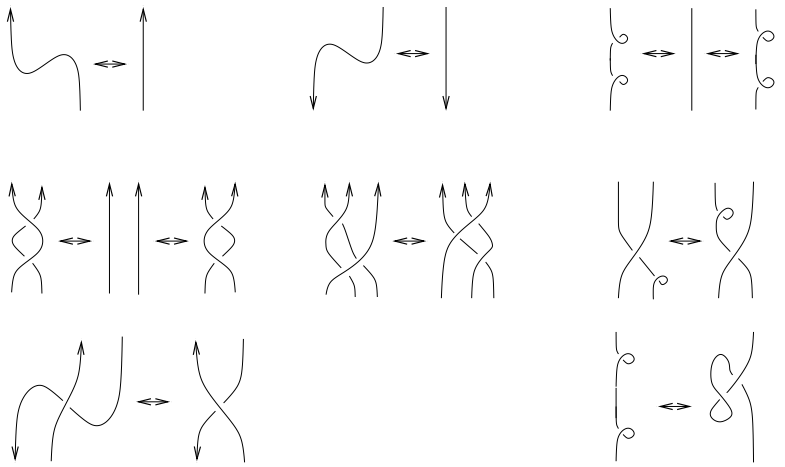


Figure 16.11: Elementary ribbon isotopies

Exercise 16.32 Show that the invariant $F(L)$ is well defined when I is any collection of objects in a ribbon category, ie that it is invariant under elementary isotopies.

More generally given a framed link $L_M \cup L$ where L is colored and oriented define

$$F(L_M, L) = \sum_{\lambda_1,\dots,\lambda_{c(L_M)} \in I} J_{\lambda_1,\dots,\lambda_{c(L_M)},V_1,\dots,V_{c(L)}}(L_M \cup L) .$$

Remark 16.33 This is well defined independently of orientations on L_M because labeling a component with λ is equivalent to changing the orientation on the component and labeling it with λ^*. By axiom 16 the sum is symmetric with respect to taking duals. This is also consistent with the sum over all colorings of unlabeled components convention.

Let M be the manifold obtained by surgery on a framed link L_M. Even though $F(L_M)$ is invariant under the Reidemeister moves as is, it is not invariant under the Kirby moves. Luckily, invariants of non-isotopic links defining the same manifold are related in a very simple manner and we can multiply $F(L_M)$ by normalizing factors that cancel out this dependence.

Given a framed link, L_M, one defines a linking matrix $n_{ij}(L)$ with off-diagonal entries equal to the linking numbers of the corresponding components and with diagonal entries equal to the self-linking or writhe of the corresponding component (see Prasolov and Sossinsky [124] and Rolfsen [130]).

Example 16.34 The linking matrix of the framed left Hopf link from Figure 16.2 is given on the left below and the linking matrix of the framed link from Figure 13.2 is given on the right.

$$\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \qquad \begin{pmatrix} 2 & \pm 1 \\ \pm 1 & 0 \end{pmatrix}$$

Let $\sigma(L_M)$ denote the signature, that is, the number of positive eigenvalues minus the number of negative eigenvalues of the linking matrix. With these notations one can show that the quantity $\tau(M)$ defined below is a topological invariant (see Bakalov and Kirillov [21] and Turaev [152]). This means that it does not depend on the link used to represent M (so any sequence of positive or negative Kirby moves or their inverses may be applied to L_M without changing the answer).

Definition 16.35 The Reshetikhin–Turaev invariant of a 3–manifold is given by

$$\tau(M) := (p^-)^{\sigma(L_M)} \mathcal{D}^{-\sigma(L_M)-c(L_M)-1} F(L_M)\,.$$

It is also possible to define invariants of oriented colored framed links in 3–manifolds. A colored framed link in a 3–manifold can be represented by a link in S^3 with some of the components labeled with objects from a strict modular category. Let L_M denote the sublink consisting of unlabeled components. We assume that L_M is a surgery presentation of M. Each of the labeled components represent a component of the framed link L in M. The labels correspond to the representations labeling Wilson loops in the heuristic description.

The more general invariant is defined by

$$\tau(M, L) := (p^-)^{\sigma(L_M)} \mathcal{D}^{-\sigma(L_M)-c-1} F(L_M, L)\,.$$

We define,

$$Z(M) := \tau(M, \varnothing), \qquad \text{and} \qquad W_{R_1,\dots,R_c}(L) := \tau(S^3, L)/\tau(S^3). \tag{28}$$

to be the mathematical interpretations of the physically motivated invariants discussed earlier.

It may appear that we have a reasonably simple definition of partition functions in 16.35; however it depends on the choice of a strict modular category and we do not have any other than $\mathcal{U}(1)_{2m+1}$ yet. The invariants corresponding to $\mathcal{U}(1)_{2m+1}$ are heuristically the same as the invariants 'defined' via the path integral by integrating holonomies over a space of $\mathrm{U}(1)$ connections. Since $\mathrm{U}(1)$ is Abelian, the cubic terms in the Chern–Simons invariant vanish so the resulting theory is what physicists call a Gaussian theory. This is the case where the path integral would be easiest to formalize mathematically but it leads to fairly weak invariants. In the next subsection we define quantum groups because their representations lead to more complicated strict modular categories that in turn lead to more interesting quantum invariants.

We now compute the invariant of S^3 in three different ways. Using the empty link we see that $\tau(S^3) = \mathcal{D}^{-1}$. The following examples compute the same invariant from different link presentations. Note that there is some algebra involved in establishing that the obtained values are the same.

Example 16.36 Let L be the link (twisted unknot) that defines p^- in Figure 16.10. We compute $F(L)$ and $\tau(M_L)$ in the category $\mathcal{U}(1)_{2m+1}$. Coloring the single component by p and orienting the component clockwise we have

$$J_p(L) = \cap_{p^*} \circ (\theta_p^{-1} \otimes \mathrm{id}_{p^*}) \circ \cup_p = \theta_p^{-1} = e^{-2\pi i p^2/(2m+1)},$$

where p takes values $0, 1, \ldots, 2m$. Since $d_p = 1$ for all p,

$$F(L) = \sum_{p=0}^{2m} d_p J_p(L) = \sum_{p=0}^{2m} e^{-2\pi i p^2/(2m+1)}.$$

Since L only has one component the linking 'matrix' is just the self-linking number which is -1 because of the twist. The signature is also -1. The quantum diameter of $\mathcal{U}(1)_{2m+1}$ is $\mathcal{D} = (2m+1)^{1/2}$ (see Remark 16.24). By the very choice of L we have $p^- = F(L)$. Thus

$$\tau(M_L) = (p^-)^{-1}\mathcal{D}^{1-1-1}F(L) = \mathcal{D}^{-1} = (2m+1)^{-1/2}.$$

Example 16.37 A slightly more difficult example is the left Hopf link from Figure 16.2. In fact, we already computed $F(L) = \sum_{p,q=0}^{2m} e^{-4\pi i pq/(2m+1)} = 2m+1$ back in Example 16.2. Self-linking numbers in this case are both 0 (no twists) and the linking numbers between the components are -1 (orientation!). Hence the linking matrix is

$$\mathrm{lk}(L) = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$$

and $\sigma(L) = 1 - 1 = 0$. Assembling the results we get

$$\tau(M_L) = (p^-)^0 \mathcal{D}^{-3} F(L) = (2m+1)^{-3/2} \sum_{p,q=0}^{2m} e^{-4\pi i pq/(2m+1)} = (2m+1)^{-1/2}.$$

Exercise 16.38 Use the Reidemeister and Kirby moves to show that all three links employed above represent the same 3–manifold, namely the 3–sphere.

To fully describe how the notion of a strict modular category relates to the Witten's Chern–Simons invariants one should describe the associated topological quantum field theory. The general outline is as follows. Diagrams with only closed components may be cut into parts with marked surface boundaries. A Hilbert space will then be associated to each such marked surface and a bounded linear map will be associated to a cobordism. The Hilbert space associated to a genus g surface marked with colors (simple elements of the category) $V_1, \dots, V_n$ is $\mathcal{H} = \mathrm{Hom}(\mathbb{1}, (\bigoplus_{j\in I} \lambda_j \otimes \lambda_j^*)^{\otimes g} \otimes V_1 \otimes \cdots \otimes V_n)$. There are many technical details that must be addressed in order to do this properly. A very careful explanation is given by Turaev [152].

Exercise 16.39 Compute the $\mathcal{U}(1)_{2m+1}$ invariant of the 3–manifold represented by the framed link from Figure 13.2.

Exercise 16.40 Describe how to compute the W invariant of any framed link for this strict modular category. (There is a fairly simple formula in terms of the linking matrix and colors.)

17 Quantum groups and their representations

The ingredients in a strict modular category look like the representations of some algebraic object. This is indeed one of the best methods to construct strict modular categories. In this subsection we consider the representations of deformations of Lie algebras called quantum groups. Deformations are constructed to arrive in a non-commutative setting. Without such a deformation the morphism that would be attached to a right crossing would be the same as the one attached to a left crossing. The invariants constructed from the resulting representations are the original Reshetikhin–Turaev invariants.

17.1 Quantum groups at roots of unity

The axioms of a strict modular category are very complicated, so without some additional motivation it would be difficult to construct an interesting modular category. Luckily the study of symmetry in quantum mechanics led to very similar structures and it was possible to construct interesting modular categories from quantum groups.

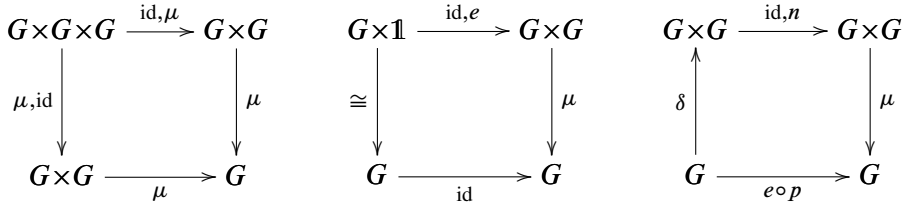
The category that produces the usual quantum or Reshetikhin–Turaev invariants is a category of representations of certain quantum groups. The latter are algebraic objects that replace classical groups in description of symmetries on quantum (or noncommutative) spaces. The term 'quantum group' is used rather loosely and is usually reserved for deformations of algebras associated to classical groups. In particular, they are not groups; instead they generalize classical group algebras and enveloping algebras rather than groups.

We begin this section with a brief description of the ideas that led to the discovery of quantum groups. As one would guess from the name the original motivation for the definition of a quantum group comes from quantum mechanics. However there is also a strong analogy between quantum groups and algebraic groups and much of the theory of quantum groups was first developed for algebraic groups. To quantize a mechanical system via canonical quantization one tries to find an embedding of a deformation of the algebra of observables into the linear operators on a Hilbert space (see Simms and Woodhouse [142]). The failure of operators corresponding to various observables to commute can be interpreted as the uncertainty principle. Often the Hilbert space can be taken to be a space of functions on the configuration space.

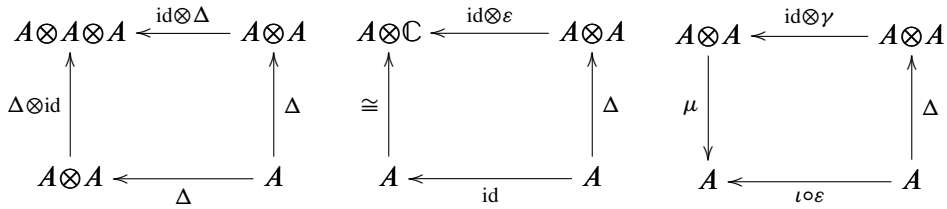
Example 17.1 For a free particle moving in 1–dimensional space the observables are functions of position x and momentum p. A possible Hilbert space to associate to this system is the space of L^2 functions in the variable x. One then associates multiplication by x, denoted by L_x, to the observable x and the operator $L_p = -i\hbar\frac{\partial}{\partial x}$ to the observable p. Notice that one has $[L_x, L_p] = i\hbar$. Thus L_x and L_p do not commute. However in the classical limit $\hbar \to 0$ one recovers the classical algebra of observables. In most everyday situations this is a reasonable approximation because $\hbar = 1.055 \times 10^{-34}$ joule-sec.

For the case of maximal symmetry the configuration space is a group. When the underlying configuration space is a group the space of functions on the group can be given the structure of a Hopf algebra. Thus Hopf algebras appear naturally in quantum mechanics as algebras of functions on groups.

Recall that a group can be described as a tuple $(G, \mu\colon G\times G\to G, e\colon \mathbb{1}\to G, n\colon G\to G)$ such that the following diagrams commute.



Here $\delta\colon G\to G\times G$ is the diagonal map. If A is an algebra of functions on a finite group G, the multiplication μ identity e and inverse n on G will induce a comultiplication $\Delta\colon A\to A\otimes A$, counit $\varepsilon\colon A\to\mathbb{C}$ and antipode $\gamma\colon A\to A$ on A such that the following diagrams commute.



The comultiplication Δ and counit ε are algebra homomorphisms and the antipode is an anti-homomorphism ($\gamma(ab)=\gamma(b)\gamma(a)$.) This is essentially the definition of a Hopf algebra.

Definition 17.2 A Hopf algebra $\mathcal{A}$ over a field $\mathbb{F}$ is an associative algebra with additional algebra homomorphisms (counit, coproduct) $\varepsilon\colon \mathcal{A}\to\mathbb{F}$, $\Delta\colon \mathcal{A}\otimes\mathcal{A}$ and an algebra antihomomorphism (antipode) $\gamma\colon \mathcal{A}\to\mathcal{A}$ satisfying axioms that dualize the usual axioms for the unit, product and inverse in a group (see the examples below, Chari and Pressley [40] or Majid [101] for the complete list).

For quantization one needs to deform the algebra of functions into a non-commutative algebra. Now consider the case of the group $SL_2\mathbb{C}$. The natural action of this group on $\mathbb{C}^2$ will help motivate the correct deformation. The algebra of functions on $\mathbb{C}^2$ is just the ring of polynomials $\mathbb{C}[x,y]$ and the algebra of functions on $SL_2\mathbb{C}$ is $\mathbb{C}[a,b,c,d]/(ad-bc-1)$. Recall that a map between spaces induces a map on the associated function algebras going in the opposite direction. Now the natural action of $SL_2\mathbb{C}$ on $\mathbb{C}^2$ by matrix multiplication induces the corresponding map of function algebras

$$\mathbb{C}[x,y]\to(\mathbb{C}[a,b,c,d]/(ad-bc-1))\otimes\mathbb{C}[x,y].$$

To arrive at a non-commutative deformation of the function algebra of $SL_2\mathbb{C}$ (that is, $\mathbb{C}[a,b,c,d]/(ad-bc-1)$), start with the non-commutative complex plane. The function algebra of the non-commutative plane is $\mathbb{C}\{x,y\}[\![h]\!]/(xy-e^{-h}yx)$. Here $\mathbb{C}\{x,y\}$ is the free algebra on two generators over $\mathbb{C}$ and $R[\![h]\!]$ refers to formal power series in the variable h with coefficients in R. In order to keep track of all of the important algebraic structure one must consider the Hopf algebra structure on the resulting algebra.

There is essentially a unique Hopf algebra that respects the map of function algebras induced by the action of $SL_2\mathbb{C}$ on $\mathbb{C}^2$ with the algebra of $\mathbb{C}^2$ replaced by the non-commutative version and the determinant replaced by $ad-e^{-h}bc$. This algebra is denoted by $SL_2^q\mathbb{C}$. Such algebras or their duals are called quantum groups (sic!). The quantum group $U_q(\mathfrak{sl}_2\mathbb{C})$ with $q=e^h$ is 'dual' to the deformed function algebra $SL_2^q\mathbb{C}$. See Chari and Pressley [40, Chapter 7] for more details.

Because the axioms for operations and co-operations in a Hopf algebra are symmetric one can define a dual Hopf algebra $\mathcal{A}^*$ (in a couple of ways) switching them, that is, the product in the dual comes from the coproduct in the original, etc. We will be mostly interested in the duals of the deformed group algebras such as $SL_2^q\mathbb{C}$. For simply connected Lie groups these duals or distribution algebras can be described as deformations of the universal enveloping algebras $U(\mathfrak{g})$ of the corresponding Lie algebras $\mathfrak{g}$. Our main example comes from deforming $U(\mathfrak{sl}_N\mathbb{C})$.

Many Lie algebras are matrix algebras with bracket given by $[X,Y]=XY-YX$. Any Lie algebra $\mathfrak{g}$ can be embedded in a unital associative algebra so that the Lie bracket takes this form. Recall that the universal enveloping algebra $U(\mathfrak{g})$ is the quotient of the tensor algebra $\oplus_{n\geq 0}\mathfrak{g}^{\otimes n}$ by the ideal generated by $x\otimes y-y\otimes x-[x,y]$ with $x,y\in\mathfrak{g}$. In a form more germane to quantum generalizations $U(\mathfrak{g})$ can be described by generators and relations. Namely, if $x_1,\dots,x_n$ generate $\mathfrak{g}$ with relations given in terms of brackets then $U(\mathfrak{g})$ has a presentation with the same set of generators with relations obtained by replacing all brackets with the corresponding commutators. For example $[x,y]$ is replaced by $xy-yx$. We will usually write brackets even when working in $U(\mathfrak{g})$ interpreting them as commutators.

Before discussing the deformations recall the presentation of the enveloping algebra.

Exercise 17.3 Compute $[e,f]$, $[L,e]$ and $[L,f]$ for the following matrices.

$$e=\begin{pmatrix}0&1\\0&0\end{pmatrix},\qquad f=\begin{pmatrix}0&0\\1&0\end{pmatrix},\qquad L=\begin{pmatrix}1&0\\0&-1\end{pmatrix}.$$

Exercise 17.4 Express $[X,[X,Y]]$ as a product of X and Y factors. Let $E_{i,j}$ be the square matrix with a 1 in the i,j entry and zeros elsewhere. Compute $[E_{i,i+1}, E_{j,j+1}]$, $[E_{i,i+1}, E_{j+1,j}]$ and $[E_{i,i+1}, E_{j,j} - E_{j+1,j+1}]$.

Example 17.5 Recall from Fulton and Harris [62], Humphreys [78] or Appendix D that $\mathfrak{sl}_N\mathbb{C}$ is generated by $e_i, f_i, \alpha_i^\vee$ and $i = 1,\dots,N-1$. Here $\alpha_i^\vee$ represent the simple coroots of $\mathfrak{sl}_N\mathbb{C}$ and e_i, f_i are the corresponding positive and negative root vectors. In other words $e_i = E_{i,i+1}$, $f_i = E_{i+1,i}$ and $\alpha_i^\vee = E_{i,i} - E_{i+1,i+1}$. The relations in $\mathfrak{sl}_N\mathbb{C}$ and $U(\mathfrak{sl}_N\mathbb{C})$ are:

$$\begin{array}{lll} [\alpha_i^\vee, \alpha_j^\vee] = 0 & [e_i, f_j] = \delta_{ij}\alpha_j^\vee, & \\ [\alpha_i^\vee, e_j] = a_{ij}e_j & [e_i, e_j] = [f_i, f_j] = 0, & |i-j| \neq 1 \\ [\alpha_i^\vee, f_j] = -a_{ij}f_j & [e_i,[e_i,e_j]] = [f_i,[f_i,f_j]] = 0, & j = i \pm 1, \end{array} \tag{29}$$

where

$$a_{ij} = \mathrm{Tr}(\alpha_i^\vee \alpha_j^{\vee\dagger}) = \begin{cases} 0 & |i-j|>1 \\ -1 & |i-j|=1 \\ 2 & i=j \end{cases}$$

is the Cartan matrix of $\mathfrak{sl}_N\mathbb{C}$. The last pair of relations are known as the Serre relations. The Serre relations look more familiar if written in the associative form

$$e_i^2 e_j - 2e_i e_j e_i + e_j e_i^2 = f_i^2 f_j - 2f_i f_j f_i + f_j f_i^2 = 0,\ \ j = i \pm 1. \tag{30}$$

$U(\mathfrak{sl}_N\mathbb{C})$ also has a Hopf algebra structure given by

$$\varepsilon(x) = 0, \qquad \gamma(x) = -x, \qquad \Delta(x) = 1\otimes x + x \otimes 1.$$

For $\mathfrak{sl}_2\mathbb{C}$ there is only one generator of each type $e, f, \alpha^\vee$, the Cartan matrix is 1×1, that is, the number 2 and the Serre relations trivialize so (29) reduces to

$$[\alpha^\vee, e] = 2e, \qquad [\alpha^\vee, f] = -2f \qquad [e, f] = \alpha^\vee.$$

Following the motivation above we now explain the quantum deformation $U_q(\mathfrak{sl}_2\mathbb{C})$. According to V Drinfeld [51] the first two relations in (29) should stay intact whereas the last one should become

$$[e, f] = \frac{\sinh(\hbar\alpha^\vee)}{\sinh(\hbar)},$$

where $\hbar$ is a formal parameter (the Planck constant). To make sense of this we would have to consider formal power series in $\hbar$ with coefficients in $U(\mathfrak{sl}_2\mathbb{C})$ which is not

very convenient. Fortunately, there is a bypass due to M Jimbo who suggested setting $q = e^{\hbar}$ and introduced two new generators $q^{\pm\alpha^\vee}$ so that the last relation becomes

$$[e, f] = \frac{q^{\alpha^\vee} - q^{-\alpha^\vee}}{q - q^{-1}}.$$

Now we only have to extend the field of coefficients from $\mathbb{C}$ to $\mathbb{C}(q)$ (rational functions in q). Of course we now need to eliminate $\alpha^\vee$ form the first two relations above which leads to

$$q^{\alpha^\vee} e q^{-\alpha^\vee} = q^2 e, \qquad q^{\alpha^\vee} f q^{-\alpha^\vee} = q^{-2} f.$$

Note that $q^{\pm\alpha^\vee}$ are indeed new generators and by no means a variable q 'taken to the power' $\pm\alpha^\vee$ (some authors denote them $K^{\pm 1}$ to prevent confusion; see Chari and Pressley [40] and Majid [101]). It is interesting that the representations of $U(\mathfrak{sl}_2\mathbb{C})$ and $U(\mathfrak{sl}_2\mathbb{C})[[\hbar]]$ are in bijective correspondence [40]. Moreover, any deformation of $U(\mathfrak{sl}_2\mathbb{C})$ and more generally $U(\mathfrak{g})$ as an associative algebra is trivial, that is it produces an isomorphic algebra [40]. It is in the Hopf algebra structure that the difference between U and U_q becomes essential. For $U_q(\mathfrak{sl}_2\mathbb{C})$ the deformed co-operations are given on the generators by

$$\begin{aligned}
\varepsilon(q^{\pm\alpha^\vee}) &= 1, & \gamma(q^{\pm\alpha^\vee}) &= q^{\mp\alpha^\vee}, & \Delta(q^{\pm\alpha^\vee}) &= q^{\pm\alpha^\vee} \otimes q^{\pm\alpha^\vee},\\
\varepsilon(e) &= 0, & \gamma(e) &= -eq^{-\alpha^\vee}, & \Delta(e) &= e \otimes q^{\alpha^\vee} + 1 \otimes e,\\
\varepsilon(f) &= 0, & \gamma(f) &= -q^{\alpha^\vee} f, & \Delta(f) &= f \otimes 1 + q^{-\alpha^\vee} \otimes f.
\end{aligned}$$

As Hopf algebras $U(\mathfrak{sl}_2\mathbb{C})$ and $U_q(\mathfrak{sl}_2\mathbb{C})$ are not isomorphic.

One would expect that $U_q(\mathfrak{sl}_2\mathbb{C})$ is a deformation of $U(\mathfrak{sl}_2\mathbb{C})$ but this is not quite true. What is true (see Kassel [82]) is

$$U(\mathfrak{sl}_2\mathbb{C}) = U_{q=1}(\mathfrak{sl}_2\mathbb{C})/(q^{\alpha^\vee} - 1).$$

We now give the general definition for $U_q(\mathfrak{sl}_N\mathbb{C})$ following Chari and Pressley [40].

Definition 17.6 The (rational form of the) Drinfeld–Jimbo quantum group $U_q(\mathfrak{sl}_N\mathbb{C})$ is the Hopf algebra generated as an associative algebra over $\mathbb{C}(q)$ by the generators $q^{\pm\alpha^\vee{}_i}, e_i, f_i$ with $i = 1, \dots, N-1$ satisfying the relations

(31)

$$\begin{array}{lll}
q^{\alpha^\vee{}_i} q^{-\alpha^\vee{}_i} = q^{-\alpha^\vee{}_i} q^{\alpha^\vee{}_i} = 1, & [e_i, f_j] = \delta_{ij} \dfrac{q^{\alpha^\vee{}_i} - q^{-\alpha^\vee{}_i}}{q - q^{-1}}, & \\
q^{\alpha^\vee{}_i} e_j q^{-\alpha^\vee{}_i} = q^{a_{ij}} e_j, & [e_i, e_j] = [f_i, f_j] = 0, & |i-j| \neq 1,\\
q^{\alpha^\vee{}_i} f_j q^{-\alpha^\vee{}_i} = q^{-a_{ij}} f_j, & e_i^2 e_j - (q+q^{-1}) e_i e_j e_i + e_j e_i^2 = 0, & j = i \pm 1,\\
 & f_i^2 f_j - (q+q^{-1}) f_i f_j f_i + f_j f_i^2 = 0, & j = i \pm 1.
\end{array}$$

Here $a_{ij} = \mathrm{Tr}(\alpha_i^\vee \alpha_j^{\vee\dagger})$ is the Cartan matrix of $\mathfrak{sl}_N\mathbb{C}$. On the generators the counit, the antipode and the coproduct are given by

$$\begin{aligned}
&\varepsilon(q^{\pm\alpha^\vee{}_i}) = 1, \quad \gamma(q^{\pm\alpha^\vee{}_i}) = q^{\mp\alpha^\vee{}_i}, \qquad \Delta(q^{\pm\alpha^\vee{}_i}) = q^{\pm\alpha^\vee{}_i} \otimes q^{\pm\alpha^\vee{}_i},\\
&\varepsilon(e_i) = 0, \qquad \gamma(e_i) = -e_i q^{-\alpha^\vee{}_i}, \qquad \Delta(e_i) = e_i \otimes q^{\alpha^\vee{}_i} + 1 \otimes e_i,\\
&\varepsilon(f_i) = 0, \qquad \gamma(f_i) = -q^{\alpha^\vee{}_i} f_i, \qquad \Delta(f_i) = f_i \otimes 1 + q^{-\alpha^\vee{}_i} \otimes f_i.
\end{aligned} \tag{32}$$

Remark 17.7 Notice that for $\mathfrak{sl}_N\mathbb{C}$ the coroots and the roots can be identified. The reason for using the notation that we used here is that it generalizes automatically to a quantum group constructed from any semisimple Lie algebra.

The comultiplication is used to construct the tensor product in the strict modular category constructed from representations of quantum groups, the counit leads to the unit, the antipode leads to the duality and the square of the antipode leads to the twist. Constructing the braiding is a bit tricky. We address these issues later in this section.

One other tricky point is making sure that there are only a finite number of simple objects. This is accomplished by specializing to a root of unity; however, this is not as easy as one might guess. It is tempting to consider $U_q(\mathfrak{sl}_N\mathbb{C})$ over $\mathbb{C}$ rather than $\mathbb{C}(q)$ by specializing q to a particular complex number z. When z is not a root of unity one obtains a Hopf algebra with generators and relations given by (31) and (32), and q replaced by z (not in $q^{\pm\alpha_i}$ since those are names of generators). As associative algebras $U_z(\mathfrak{sl}_N\mathbb{C})$ and $U(\mathfrak{sl}_N\mathbb{C})$ are isomorphic and thus have identical representation theories. As nice as this might be it means an infinite number of simple objects (irreducible representations) and no hope of modularity. To understand why the case $z = \epsilon$, $\epsilon^l = 1$ with l an integer is exceptional we need to introduce the notions of q–integers, q–factorials and q–binomials.

Definition 17.8 For any $n \in \mathbb{Z}$ define the q–integers (quantum integers) as

$$[n]_q := \frac{q^n - q^{-n}}{q - q^{-1}} \in \mathbb{C}(q)$$

and for $n \geq 0$ define the q–factorials as

$$[n]_q! := [n]_q[n-1]_q \dots [2]_q[1]_q.$$

For any pair of integers $0 \leq m \leq n$ define the q–binomial coefficient as

$$\begin{bmatrix} n \\ m \end{bmatrix}_q := \frac{[n]_q!}{[n-m]_q![m]_q!}. \tag{33}$$

Exercise 17.9 Verify that $[n]_q = q^m[n-m]_q + q^{-(n-m)}[m]_q$ and use it to derive the Pascal recurrence relation for q–binomials:

$$\begin{bmatrix} n \\ m \end{bmatrix}_q = q^m \begin{bmatrix} n-1 \\ m \end{bmatrix}_q + q^{-(n-m)} \begin{bmatrix} n-1 \\ m-1 \end{bmatrix}.$$

Obviously, in the limit $q \to 1$ the q–numbers turn into the ordinary integers, factorials and binomials. Many formulas in representation theory involve multiplication by $n!$. Under quantum deformation these become multiplications by $[n]_q!$. However, if $\epsilon^l = 1$ and we specialize to $q = \epsilon$ then $[l']_\epsilon! = 0$ for $\epsilon^{l'} = \pm 1$ and some irreducible representations will become reducible. For $\epsilon = \pm 1$ the defining relations in (31) do not even make sense as written due to division by $\epsilon - \epsilon^{-1} = 0$. But even for $l > 2$ simply setting $q = e^{2\pi i/l}$ will not produce the 'right' version of the quantum group. The right version was introduced by G Lusztig [99].

Lusztig notes that $[n]_q = \sum_{k=0}^{n-1} q^{n-1-2k} \in \mathbb{Z}[q, q^{-1}]$, that is, it is a Laurent polynomial in q, and therefore so is $[n]_q!$. Less obviously, the q–binomial $\begin{bmatrix} n \\ m \end{bmatrix}_q$ is also a Laurent polynomial.

Exercise 17.10 Prove the last claim.
Hint: Use the Pascal recurrence and proceed by induction on n and m.

The Laurent polynomials play the same role in $\mathbb{C}(q)$ as the integers play in $\mathbb{C}$. More importantly, for any $\pi \in \mathbb{Z}[q, q^{-1}]$ the value $\pi(z)$ is well-defined for any $z \neq 0$, in particular, q can be specialized even to roots of unity in Laurent polynomials. This suggests defining an integral form of $U_q(\mathfrak{g})$ before specializing to roots of unity. For the classical enveloping algebras such a form is known as the Kostant $\mathbb{Z}$–form (see Humphreys [78]) and it uses the divided powers $\frac{e_i^n}{n!}, \frac{f_i^n}{n!}$ as generators. This motivates the following definition (see Lusztig [99]):

Definition 17.11 The divided powers in $U_q(\mathfrak{sl}_N\mathbb{C})$ are defined by

$$e_i^{(n)} := \frac{e_i^n}{[n]_q!}, \qquad f_i^{(n)} := \frac{f_i^n}{[n]_q!}.$$

The trick now is to rewrite the relations (31) in terms of the divided powers and make sure that their coefficients are Laurent polynomials. This is indeed the case and we give some of the relations below (the full list occupies an entire page in Chari and Pressley

[40]).

$$
\begin{aligned}
q^{\alpha_i} e_j^{(n)} q^{-\alpha_i} &= q^{n a_{ij}} e_j^{(n)}, & e_i^{(m)} e_i^{(n-m)} &= \begin{bmatrix} n \\ m \end{bmatrix}_q e_i^{(n)}, \\
q^{\alpha_i} f_j^{(n)} q^{-\alpha_i} &= q^{-n a_{ij}} f_j^{(n)}, & f_i^{(m)} f_i^{(n-m)} &= \begin{bmatrix} n \\ m \end{bmatrix}_q f_i^{(n)}, \\
e_i^{(n)} f_j^{(m)} &= f_j^{(m)} e_i^{(n)}, \qquad i \neq j, \\
e_i^{(m)} e_i^{(n-m)} &= \sum\nolimits_{j=1}^{\min[n,n-m]} f_i^{(n-m-j)} \begin{bmatrix} q^{\alpha_j};\ 2j-n \\ j \end{bmatrix}_q e_i^{(m-j)}.
\end{aligned}
\tag{34}
$$

In the last formula we used a new notation

$$
\begin{bmatrix} q^{\alpha_i};\ c \\ j \end{bmatrix}_q := \prod\nolimits_{k=1}^{j} \frac{q^{c+1-k} q^{\alpha_i} - q^{-(c+1-k)} q^{-\alpha_i}}{q^k - q^{-k}}
\tag{35}
$$

with $c \in \mathbb{Z}$ and $j \in \mathbb{Z}_{\geq 0}$. These new elements are not generators, they can be expressed via $e_i^{(n)}$, $f_i^{(n)}$ with $\mathbb{Z}[q, q^{-1}]$ coefficients (see Lusztig [99]). We mention them because they will play an important role in the representation theory later. Once we have the new relations we can forget about the origin of the divided powers and treat them as formal symbols that satisfy the relations (34). Indeed, we have to do this since Definition 17.11 makes no sense for $q = \epsilon$ a root of unity.

Definition 17.12 (Quantum groups at roots of unity) The restricted integral form $U^{\mathrm{res}}_{\mathbb{Z}[q,q^{-1}]}(\mathfrak{sl}_N\mathbb{C})$ of the quantum group $U_q(\mathfrak{sl}_N\mathbb{C})$ is the Hopf algebra generated over $\mathbb{Z}[q, q^{-1}]$ by $q^{\pm\alpha_i^\vee}, e_i^{(n)}, f_i^{(n)}$. The corresponding quantum group at a root of unity $U^{\mathrm{res}}_{\epsilon}(\mathfrak{sl}_N\mathbb{C})$ is obtained by specializing q to ϵ and changing the coefficients to $\mathbb{C}$. Formally,

$$
U^{\mathrm{res}}_{\epsilon}(\mathfrak{sl}_N\mathbb{C}) := U^{\mathrm{res}}_{\mathbb{Z}[q,q^{-1}]}(\mathfrak{sl}_N\mathbb{C}) \otimes_{\mathbb{Z}[q,q^{-1}]} \mathbb{C},
$$

where $\mathbb{Z}[q, q^{-1}]$ acts on $\mathbb{C}$ in the obvious way with $\pi \mapsto \pi(\epsilon)$.

Remark 17.13 The algebra $U^{\mathrm{res}}_{\mathbb{Z}[q,q^{-1}]}(\mathfrak{sl}_N\mathbb{C})$ has a presentation with generators $q^{\pm\alpha_i}$, $e_i^{(n)}$, $f_i^{(n)}$ and relations including (31) and co-operations induced from (32).

At this point the whole digression on divided powers may seem superfluous: why not simply take the subalgebra of $U_q(\mathfrak{sl}_N\mathbb{C})$ generated over $\mathbb{Z}[q, q^{-1}]$ by $q^{\pm\alpha_i^\vee}, e_i, f_i$ and then specialize to ϵ? Intuitively, the difference is due to the following. In $U_q(\mathfrak{sl}_N\mathbb{C})$ we have the equality $e_i^n = [n]_q! e_i^{(n)}$. Since this equality only contains a Laurent polynomial it continues to hold in $U^{\mathrm{res}}_{\mathbb{Z}[q,q^{-1}]}(\mathfrak{sl}_N\mathbb{C})$ and therefore in $U^{\mathrm{res}}_{\epsilon}(\mathfrak{sl}_N\mathbb{C})$. However, if $\epsilon^l = 1$ then $[n]_\epsilon! = 0$ for $n \geq 2l$ and the higher powers of e_i, f_i vanish while the divided powers survive! Otherwise there is no way to define divided powers and the

quantum group simply looses part of the structure. On a bright side, for $0 \le n < l/2$ we have $[n]_\epsilon! \neq 0$ and the quantity $e_i^{(n)} = e_i^n/[n]_\epsilon!$ is well-defined even in $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$. This simplifies computations with the divided powers in this range and allows one to use simpler relations (31) instead of (34).

17.2 Representations of $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$ and tilting modules

The category of all finite-dimensional representations of $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$ is still too large to be modular. What we need is the category $\overline{\mathcal{T}ilt}_\epsilon$ of the 'reduced tilting modules'. This is a suitable subquotient of the category of representations of $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$. This means that we will consider only some of the representations (the tilting modules) and construct $\overline{\mathcal{T}ilt}_\epsilon$ by quotients of these by 'negligible' parts (reduced). In this subsection we define the subcategory of tilting modules. Here the word module simply means representation space of $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$. We define the category of reduced tilting modules in the next subsection and defer the modular structure until even later.

Much of the early work on these representations was done by algebraists (see Lusztig [99] and Andersen [9]) interested in the representation theory of algebraic groups in positive characteristic. They introduced the terminology which became standard. We adopt it here despite the fact that it is not customary in the representation theory of Lie groups and algebras that are the closest classical analogs. Since prime roots are the most interesting for algebraic groups the theory was originally developed for odd roots of unity. This makes the algebra somewhat easier (see [9] and Chari and Pressley [40]). On the other hand, for topological applications one has to consider the even roots of unity. In particular, the correspondence with the SU_N Chern–Simons theory at level k requires the order of the root to be $l = 2(k+N)$. More seriously, the category $\overline{\mathcal{T}ilt}_\epsilon$ for odd roots of unity is not modular, the non-degeneracy Axiom 17 fails(see Sawin [135]). This problem can actually be fixed by 'modularization' but this was realized much later and involves additional technicalities (see Bruguières [35]).

To keep track of the difference between the even and odd cases we define

$$l' := \begin{cases} l & l \text{ even} \\ l/2 & l \text{ odd} \end{cases}.$$

Said in a different way l' is the smallest positive integer such that $\epsilon^{l'} = \pm 1$ where ϵ is a primitive lth root of unity.

For a while in the 1990s there existed a well-developed representation theory for odd roots of unity that did not lead to a non-degenerate $\widetilde{s}$–matrix. There was also a non-degeneracy proof for even roots (see Turaev and Wenzl [153]) but no corresponding

representation theory. Thus in papers and monographs written in the 1990s authors either did not treat 3–manifold invariants at all [40] or implicitly assumed that the representation theory transfers from the odd case to the even (see the work of Bakalov, Kirillov, Reshetikhin and Turaev [21; 83; 152]). There is still no single source where all the required algebraic facts are stated and proved in the correct generality. We will state results in a form that works for both cases but a reader interested in connections to Chern–Simons theory may safely assume everywhere that $l = 2(k + N)$.

Representations of the classical enveloping algebra $U(\mathfrak{g})$ are of course the 'same' as those of $\mathfrak{g}$ meaning that every representation of the latter extends to one of the former. The representation theory of $U_q(\mathfrak{g})$ after specializing to $q = \epsilon$ includes some subtleties. To stimulate intuition we begin by recalling how a classical irreducible representation of $\mathfrak{sl}_2\mathbb{C}$ with the highest weight $\lambda \in \Lambda_w^+$ is constructed. Looking at Appendix D first for notation and basic results from classical representation theory may help.

Example 17.14 Let $\mathfrak{g} = \mathfrak{sl}_2\mathbb{C}$ and $\lambda = \omega = \frac{1}{2}\alpha$, where $\alpha = E_{11} - E_{22}$ is the (unique) simple root of $\mathfrak{sl}_2\mathbb{C}$ and ω is the corresponding fundamental weight (here and below we identify the Cartan subalgebra $\mathfrak{h}$ with its dual $\mathfrak{h}^*$ via the Killing form). Recall from Example 17.5 that the generators and relations for $\mathfrak{sl}_2\mathbb{C}$ are e, f, α:

$$[\alpha, e] = 2e, \qquad [\alpha, f] = -2f, \qquad [e, f] = \alpha. \tag{36}$$

Let u_0 be the highest weight vector. By definition

$$\begin{aligned} eu_0 &= 0 \\ \alpha u_0 &= (\omega, \alpha)u_0 = \tfrac{1}{2}(\alpha, \alpha)u_0 = u_0. \end{aligned}$$

Now set $u_1 := fu_0$ and compute using (36)

$$\begin{aligned} eu_1 &= efu_0 = ([e, f] + fe)u_0 = \alpha u_0 + 0 = u_0 \\ \alpha u_1 &= \alpha f u_0 = ([\alpha, f] + f\alpha)u_0 = -2fu_0 + fu_0 = -u_1. \end{aligned}$$

An analogous computation shows that setting $u_2 := fu_1$ leads to $eu_2 = 0$ and $\alpha u_2 = -3u_2$. It follows that we could set $fu_1 = 0$ and obtain a well-defined representation. Furthermore the representation must be simple because the orbit of any non-zero vector under $\mathfrak{sl}_2\mathbb{C}$ generates the whole space. Thus, the representation space V_ω is spanned by u_0, u_1. In the u_0, u_1 basis we have

$$e = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad \alpha = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad f = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

which one easily recognizes as the defining representation of $\mathfrak{sl}_2\mathbb{C}$.

In general, given generators e_i, f_i, α_i of $\mathfrak{sl}_N\mathbb{C}$ and a dominant weight $\lambda \in \Lambda_w^+$ with highest weight vector u_0 one has $e_i u_0 = 0$, $\alpha_i u_0 = (\lambda, \alpha_i)u_0$. This is exactly the classical Verma module $\widehat{V}_\lambda$ where u_0 is the equivalence class of 1. We keep generating new vectors $u_{i_1,\dots,i_m} := f_{i_m} \dots f_{i_1} u_0$ in lexicographic order and keep computing the action of the generators on the new vectors using the commutation relations. When we find the largest proper set of vectors generated by these and the $\mathfrak{sl}_N\mathbb{C}$ action it will be a maximal proper ideal of the Verma module and the quotient will be the maximal abelian quotient. Said differently, we set the action of f_i to 0 on the last vectors produced and obtain an irreducible representation $V_\lambda(\mathfrak{sl}_N\mathbb{C})$. More formally, we take the quotient of the space spanned by all (infinitely many) $u_0, u_{i_1,\dots,i_m}$ by the maximal invariant subspace under the action of e_i, f_i, α_i. In Example 17.14 the maximal proper invariant subspace is spanned by $u_2, u_3, \dots$ and this is why taking the quotient reduced to setting $f u_1 = 0$.

Exercise 17.15 Show that the basis of $V_{\omega_1}(\mathfrak{sl}_3\mathbb{C})$ formed as above is u_0, u_1, u_{12} by computing the action of e_i, f_i, α_i as described above.

Now we wish to apply the same approach to $U_q(\mathfrak{sl}_2\mathbb{C})$.

Example 17.16 The generators now are $q^{\pm\alpha^\vee}, e, f$ and the relations (36) get replaced by

$$q^{\alpha^\vee} e q^{-\alpha^\vee} = q^2 e, \qquad q^{\alpha^\vee} f q^{-\alpha^\vee} = q^{-2} f, \qquad [e, f] = \frac{q^{\alpha^\vee} - q^{-\alpha^\vee}}{q - q^{-1}}. \tag{37}$$

As in Example 17.14 we have

$$\begin{aligned}
q^{\alpha^\vee} u_0 &= q^{(\omega,\alpha^\vee)} u_0 = q u_0 \\
e u_0 &= 0, \\
f u_0 &=: u_1 \\
q^{\alpha^\vee} u_1 &= q^{\alpha^\vee} f q^{-\alpha^\vee} q^{\alpha^\vee} u_0 = q^{-2} f \cdot q u_0 = q u_0, \\
e u_1 &= e f u_0 = ([e, f] + f e) u_0 = \frac{q^{\alpha^\vee} - q^{-\alpha^\vee}}{q - q^{-1}} u_0 + 0 = u_0.
\end{aligned}$$

An analogous computation for $u_2 := f u_1$ shows that $q^{\alpha^\vee} u_2 = q^{-3} u_2$ and $e u_2 = 0$. Hence we should set $f u_1 = 0$ then u_0, u_1 form a basis of $\mathcal{V}_\omega^q(\mathfrak{sl}_2\mathbb{C})$.

It is customary for $\mathfrak{sl}_2\mathbb{C}$ to use the basis of the divided powers $v_i := \frac{1}{i!} u_i = \frac{1}{i!} f^i u_0$ as a canonical one. One can show along the above lines (see Kassel [82]) that in this basis

the action of $\mathfrak{sl}_2\mathbb{C}$ and $U(\mathfrak{sl}_2\mathbb{C})$ in the representation $V_{(m-1)\omega}(\mathfrak{sl}_2\mathbb{C})$ is given by:

$$\begin{aligned}\alpha^\vee v_i &= (m-2i)v_i,\\ e v_i &= (m-i+1)\, v_{i-1}, \quad i = 0,1,\dots,m-1\\ f v_i &= (i+1)\, v_{i+1}.\end{aligned}$$

This generalizes straightforwardly to the quantum case, where ordinary numbers are replaced by q–numbers from Definition 17.8. The corresponding representation $\mathcal{V}^q_{(m-1)\omega}(\mathfrak{sl}_2\mathbb{C})$ of $U_q(\mathfrak{sl}_2\mathbb{C})$ is given by Chari and Pressley [40]

$$\begin{aligned}q^{\pm\alpha^\vee} v_i &= q^{\pm(m-2i)} v_i,\\ e v_i &= [m-i+1]_q\, v_{i-1}, \quad i = 0,1,\dots,m-1\\ f v_i &= [i+1]_q\, v_{i+1}.\end{aligned} \tag{38}$$

Representations of interest to us are constructed from the so-called Weyl modules that are the ϵ analogs of the $\mathcal{V}^q_\lambda$ representations from Example 17.16. Their construction is largely parallel to the construction of the corresponding classical representations but with important caveats.

Definition 17.17 Let Λ_r be the root lattice of $\mathfrak{sl}_N\mathbb{C}$ i.e. the lattice generated by all $\alpha_i = E^*_{ii} - E^*_{i+1i+1}$ and let $\phi\colon \Lambda_r \to \mathbb{Z}_2$ be a homomorphism. A $U_q(\mathfrak{sl}_N\mathbb{C})$–weight is a pair (λ,ϕ) with λ an ordinary $\mathfrak{sl}_N\mathbb{C}$ weight and ϕ a homomorphism. A weight vector in a $U_q(\mathfrak{sl}_N\mathbb{C})$ representation is a vector $v_{(\lambda,\phi)}$ such that $q^{\alpha_i^\vee} v_{(\lambda,\phi)} = \phi(\alpha_i) q^{\lambda(\alpha_i^\vee)} v_{(\lambda,\phi)}$. A type I representation is one with ϕ being the trivial homomorphism.

Remark 17.18 Only the type I representations of $U_q(\mathfrak{sl}_N\mathbb{C})$ have classical analogs. From here forward we only consider type I representations.

Definition 17.19 (Weyl modules) Let λ be a dominant weight and let $\mathcal{I}_\lambda$ be the left ideal of $U_q(\mathfrak{sl}_N\mathbb{C})$ generated by e_i and $q^{\pm\alpha_i^\vee} - q^{\pm\lambda(\alpha_i^\vee)}$. The Verma module is the quotient

$$\widehat{\mathcal{V}}^q_\lambda(\mathfrak{sl}_N\mathbb{C}) := U_q(\mathfrak{sl}_N\mathbb{C})/\mathcal{I}_\lambda\,.$$

Denote by $\mathcal{V}^q_\lambda(\mathfrak{sl}_N\mathbb{C})$ the quotient of $\widehat{\mathcal{V}}^q_\lambda(\mathfrak{sl}_N\mathbb{C})$ by the maximal invariant subspace with the induced action. The restricted integral form of this representation $\mathcal{V}^{q,res}_\lambda(\mathfrak{sl}_N\mathbb{C})$ is the $U^{\mathrm{res}}_{\mathbb{Z}[q,q^{-1}]}(\mathfrak{sl}_N\mathbb{C})$ submodule of $\widehat{\mathcal{V}}^q_\lambda(\mathfrak{sl}_N\mathbb{C})$ generated by 1. The Weyl module $\mathcal{W}^\epsilon_\lambda(\mathfrak{sl}_N\mathbb{C})$ is the vector space over $\mathbb{C}$ generated from $\mathcal{V}^{q,res}_\lambda(\mathfrak{sl}_N\mathbb{C})$ by changing

coefficients from $\mathbb{Z}[q,q^{-1}]$ to $\mathbb{C}$ ($\pi \mapsto \pi(\epsilon)$) with the action of $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$ obtained by specializing q to ϵ. Formally,

$$\mathcal{W}_\lambda^\epsilon(\mathfrak{sl}_N\mathbb{C}) := \mathcal{V}_\lambda^{q,res}(\mathfrak{sl}_N\mathbb{C}) \otimes_{\mathbb{Z}[q,q^{-1}]} \mathbb{C}.$$

This definition seems a bit convoluted but it is in essence parallel to the definition of $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$ itself. We are trying to avoid the trivialization of the powers of e_i, f_i by making sure that the divided powers are present as 'independent' quantities. The space $\widehat{\mathcal{V}}_\lambda^q(\mathfrak{sl}_N\mathbb{C})$ is just the vector space over $\mathbb{C}(q)$ generated by $u_0, u_{i_1,\dots,i_m}$ for $i_k = 1,\dots,N-1$ with the action of $U_q(\mathfrak{sl}_N\mathbb{C})$ determined by

$$e_i u_0 = 0, \quad q^{\pm\alpha_i} u_0 = q^{\pm(\lambda,\alpha)} u_0, \quad f_i u_{i_1,\dots,i_m} = u_{i_1,\dots,i_m,i}\,.$$

To get to the Weyl modules we restrict to an integral form then specialize coeficients to $\mathbb{C}$.

Example 17.20 Recall from Example 17.16 that $\mathcal{V}^q_{(m-1)\omega}(\mathfrak{sl}_2\mathbb{C})$ is spanned by $v_0,\dots,v_{m-1}$. Iterating the action (38) we obtain

$$e^n v_i = [m-i+1]_q \dots [m-i+n]_q\, v_{i-n} = \frac{[m-i+n]_q!}{[m-i]_q!}\, v_{i-n}.$$

and analogously for f^n. Using Definition 17.11 of the divided powers and (33) of q–binomials we get

$$\begin{aligned} e^{(n)} v_i &= \begin{bmatrix} m-i+n \\ n \end{bmatrix}_q v_{i-n}, \\ f^{(n)} v_i &= \begin{bmatrix} i+n \\ n \end{bmatrix}_q v_{i+n}. \end{aligned} \tag{39}$$

Since q–binomials are in $\mathbb{Z}[q,q^{-1}]$ we see that $v_0,\dots,v_{m-1}$ also form a basis of $\mathcal{V}^{\mathrm{q,res}}_{(m-1)\omega}(\mathfrak{sl}_2\mathbb{C})$ with the action given by (39). Thus by Definition 17.19 the Weyl module $\mathcal{W}^\epsilon_{(m-1)\omega}(\mathfrak{sl}_2\mathbb{C})$ is spanned by the same vectors and the action of $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_2\mathbb{C})$ on them is given by replacing q with ϵ in (39). It can be shown (see Chari and Pressley [40]) that $\mathcal{W}^\epsilon_{(m-1)\omega}(\mathfrak{sl}_2\mathbb{C})$ has an invariant subspace $\mathcal{W}_{\mathrm{inv}}$ unless $m \le l'$ or $m \equiv l'-1 \mod l'$; see Exercise 17.27. This means that for other values of m not only is $\mathcal{W}^\epsilon_{(m-1)\omega}(\mathfrak{sl}_2\mathbb{C})$ not irreducible but it is not even a direct sum of irreducibles. Thus, complete reducibility of the classical representations that still holds for $U_q(\mathfrak{sl}_N\mathbb{C})$ is lost for $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$.

This example demonstrates an important method of doing computations in $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$ that avoids using its complicated relations directly (and this is the reason we did not

give a complete list of them in (34)). This idea will be used again and again in the sequel.

> To obtain an equality in $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$ perform all the computations in $U_q(\mathfrak{sl}_N\mathbb{C})$ using (31) and rewrite the end result in terms of the divided powers so that it only contains Laurent polynomials as coefficients. Then specializing q to ϵ gives an equality in $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$.

Before dealing with the loss of complete reducibility we have to address a more basic problem with the definition of a weight space. Recall from Appendix D that classically a vector v has a weight $\lambda \in \Lambda_w$ if $\alpha_i v = (\lambda, \alpha_i)v$ for all simple roots α_i. For q an indeterminate this generalizes straightforwardly to the quantum case by setting $q^{\pm\alpha_i} v = q^{\pm(\lambda,\alpha_i)} v$ instead. This also works when we specialize q to a generic complex number ϵ. If however ϵ is a root of unity and $\beta \in \Lambda_r$ we have $\epsilon^{(\lambda+l\beta,\alpha_i)} = \epsilon^{(\lambda,\alpha_i)}$ and λ would only be defined modulo $l\Lambda_r$.

The underlying reason for the weight ambiguity is that for roots of unity the maximal Abelian subalgebra of $U_\epsilon^{\mathrm{res}}$ is no longer generated by $q^{\pm\alpha_i}$ [40]. The additional generators are the ones we already met in (35) $\left[\begin{smallmatrix} q^{\alpha_i};0 \\ l' \end{smallmatrix}\right]_\epsilon$. Note that substituting ϵ into 'definition' (35) leads to a meaningless expression. To make sense of these elements in $U_\epsilon^{\mathrm{res}}$ one has to reexpress them in terms of the divided powers. Below we incorporate these new elements into the definition of a weight space so that the weight is now well-defined.

Definition 17.21 Let $\mathcal{V}$ be a representation space of $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$. We say that $v \in \mathcal{V}$ is a weight vector with weight $\lambda \in \Lambda_w$ if

$$q^{\pm\alpha_i} v = \epsilon^{\pm(\lambda,\alpha_i)} v, \qquad \begin{bmatrix} q^{\alpha_i};0 \\ l' \end{bmatrix}_\epsilon v = \begin{bmatrix} (\lambda,\alpha_i) \\ l' \end{bmatrix}_\epsilon v,$$

where on the right we have the q–binomial coefficient (33) rewritten as a Laurent polynomial and specialized to ϵ. We denote the subspace of vectors in $\mathcal{V}$ with weight λ by $\mathcal{V}^\lambda$ and call it the λ–weight space. If $\mathcal{V} = \oplus_{\lambda\in\Lambda_w} \mathcal{V}^\lambda$ the righthand side is called the weight space decomposition of $\mathcal{V}$.

Remark 17.22 Notice that we are using V_λ to represent the irreducible representation with heighest weight λ and V^λ to denote the λ–weight space of a representation V. Generally we will use calagraphic fonts to denote representations of quantum groups and roman fonts to denote representations of classical algebras.

By (34) we have

$$q^{\alpha_j}(e_i^{(n)}v) = (q^{\alpha_j}e_i^{(n)}q^{-\alpha_j})q^{\alpha_j}v = q^{na_{ji}}q^{(\lambda,\alpha_i)}v\big|_{q=\epsilon}$$
$$= \epsilon^{n(\alpha_i,\alpha_i)+(\lambda,\alpha_j)} = \epsilon^{(\lambda+n\alpha_i,\alpha_j)}v.$$

As we explained this in itself does not mean that $e_i^{(n)}v$ has the weight $\lambda + n\alpha_i$ but one can show using the full list of relations in Lusztig [99] or Chari and Pressley [40] that indeed

$$e_i^{(n)}(V^\lambda) \subseteq V^{\lambda+n\alpha_i}, \quad f_i^{(n)}(V^\lambda) \subseteq V^{\lambda-n\alpha_i}.$$

With this notation we have a very important result that follows from the definitions by a deformation argument [40]:

> Weight space decompositions of $V_\lambda, \mathcal{V}_\lambda^q, \mathcal{W}_\lambda^\epsilon$ are the same, that is, their weights and the dimensions of their weight spaces are equal.

Since weight space decompositions of classical representations are well-known this observation comes handy when performing computations with representations of quantum groups.

Circumventing the lack of complete reducibility is not as simple. Ultimately, we will have to restrict to the class of admissible representations for which complete reducibility still holds. This will lead to the desired finite number of irreducibles. To proceed in this direction we need to introduce the notions of the dual of a representation and of the tensor product of representations. Recall that given vector spaces $\mathcal{U}, \mathcal{V}$ with a linear action of a Lie group G the actions on $\mathcal{V}^*$ and $\mathcal{U}\otimes\mathcal{V}$ are given by

$$gf(v) := f(g^{-1}v), \quad g(u\otimes v) := gu\otimes gv.$$

Quantum groups $U_q(\mathfrak{sl}_N\mathbb{C}), U_\epsilon^{\text{res}}(\mathfrak{sl}_N\mathbb{C})$ are neither Lie groups nor Lie algebras but Hopf algebras so the inverse and the tensor action are replaced by the antipode and the coproduct respectively.

Definition 17.23 Given representations $\mathcal{A}\to\text{End}(\mathcal{U})$, $\mathcal{A}\to\text{End}(\mathcal{V})$ of a Hopf algebra $\mathcal{A}$ (see Definition 17.2) the dual representation on $\mathcal{V}^*$ is given by $af(v) := f(\gamma(a)v)$ and the tensor representation on $\mathcal{U}\otimes\mathcal{V}$ is given by $a(u\otimes v) := \Delta(a)(u\otimes v)$. The unit representation $\mathbb{1} := \mathcal{A}\to\text{End}(\mathbb{C})$ is given by $az := \varepsilon(a)z$.

This definition works because γ, Δ are an antihomomorphism and a homomorphism respectively. For Lie group algebras and universal enveloping algebras with their usual Hopf structure we recover the standard definitions of the unit (trivial representation)

dual and the tensor product. This provides the unit, dual and tensor product in the modular category that we are defining.

Recall from Appendix D that for classical Lie algebras all irreducible representations are indexed by dominant weights. Since obviously the dual to an irreducible is an irreducible we get a duality involution on the set of dominant weights Λ_w^+. For $\mathfrak{sl}_N\mathbb{C}$ it can be easily described explicitly (see Fulton and Harris [62] or Humphreys [78]). Let $w_0 = \left(\begin{smallmatrix} 1 & 2 & \cdots & N-1 & N \\ N & N-1 & \cdots & 2 & 1 \end{smallmatrix}\right)$ be the order-reversing permutation then $V_\lambda^* \simeq V_{-w_0(\lambda)}$ and this is an involution because $w_0^2 = \mathrm{id}$.

Exercise 17.24 Check that this indeed works for a couple of representations of $\mathfrak{sl}_3\mathbb{C}$.

More importantly for us, this carries over to the quantum groups, in particular,

$$\mathcal{V}_\lambda^{q*}(\mathfrak{sl}_N\mathbb{C}) \simeq \mathcal{V}^q_{-w_0(\lambda)}(\mathfrak{sl}_N\mathbb{C}). \tag{40}$$

The next definition introduces the type of representations that we will use to build our modular category. These 'tilting' modules were originally introduced and studied in the context of algebraic groups. In fact many of the proofs refer to facts that are true by analogy with results from algebraic groups, and we do not know of a reference that addresses the representation theory of quantum groups that we need without assuming familiarity with algebraic groups. For example Chari and Pressley refer to algebraic groups as the 'classical' case [40]. See H H Andersen [9] for more history and some important results related to these modules.

Definition 17.25 (Tilting modules) A representation V of $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$ is said to have a Weyl filtration if there exists an increasing sequence of invariant subspaces $0 = V_0 \subset V_1 \subset \cdots \subset V_r = V$ with $V_i/V_{i-1} \simeq \mathcal{W}^\epsilon_{\lambda_i}(\mathfrak{sl}_N\mathbb{C})$ for some dominant weights $\lambda_i \in \Lambda_w^+$. A representation V is called a tilting module if both it and its dual V^* have Weyl filtrations. The category of all tilting modules is denoted by $\mathcal{T}ilt_\epsilon(\mathfrak{sl}_N\mathbb{C})$

Note that if we had complete reducibility, the existence of the Weyl filtration is equivalent to V being a direct sum of $\mathcal{W}^\epsilon_{\lambda_i}$–s. However, such direct decomposition does not hold in general, that is, the filtration 'tilts'. In particular, the tilting modules still are not completely reducible and we need a condition weaker than irreducibility to describe the 'elementary' tilting modules.

Definition 17.26 A representation V is called indecomposable if it does not split into a direct sum of two proper invariant subspaces $V \neq V_1 \oplus V_2$.

It follows from Example 17.20 that in the $\mathfrak{sl}_2\mathbb{C}$ case every Weyl module is indecomposable. Also every Weyl module obviously has a trivial Weyl filtration. However, not every Weyl module is tilting. The problem is that its dual does not necessarily have a Weyl filtration. This is closely related to the fact that some Weyl modules are reducible. The following exercise provides a good example to think about when studying these issues.

Exercise 17.27 Take $N=2$ and $k=3$ giving $l=10$ and $\epsilon=e^{\pi i/5}$. Show that the subspace of $\mathcal{W}^\epsilon_{7\omega_1}(\mathfrak{sl}_2\mathbb{C})$ generated by $f^{(3)}u_0$ and $f^{(4)}u_0$ is an invariant subspace ($[5]_\epsilon=0$) so that $\mathcal{W}^\epsilon_{7\omega_1}(\mathfrak{sl}_2\mathbb{C})$ is reducible. Use this invariant subspace to explicitly construct a non-split extension

$$0\to\mathcal{W}^\epsilon_{7\omega_1}(\mathfrak{sl}_2\mathbb{C})\to\mathcal{Q}^\epsilon_{7\omega_1}\to\mathcal{W}^\epsilon_{\omega_1}(\mathfrak{sl}_2\mathbb{C})\to 0\,.$$

The module $\mathcal{Q}^\epsilon_{7\omega_1}$ is the unique indecomposable tilting module of weight $7\omega_1$.

Nonetheless, we have the following major theorem originally due to H H Andersen [9] for l odd. The case of even l should follow the general arguments of [9] but it is not stated explicitly there.

Theorem 17.28 *Direct sums and summands, duals and tensor products of tilting modules are again tilting modules. For every $\lambda\in\Lambda^+_w$ there is a unique indecomposable tilting module $\mathcal{Q}^\epsilon_\lambda$ with the highest weight λ and one-dimensional highest weight space. Moreover, every tilting module V admits a decomposition*

$$V=\oplus_{\lambda\in\Lambda^+_w}\left(\mathcal{Q}^\epsilon_\lambda\right)^{\oplus m_\lambda(V)}$$

with multiplicities $m_\lambda(V)$ canonically determined by V and only finitely many of them non-zero. The dual is given by $\mathcal{Q}^{\epsilon *}_\lambda\cong\mathcal{Q}^\epsilon_{-w_0(\lambda)}$.

Thus, tilting modules form a subcategory $\mathcal{T}ilt_\epsilon(\mathfrak{sl}_N\mathbb{C})$ of the category of representations of $U^{\mathrm{res}}_\epsilon(\mathfrak{sl}_N\mathbb{C})$ with morphisms being equivariant maps $f(av)=af(v)$, which is closed under duality and tensor products and is dominated by indecomposable objects $\mathcal{Q}^\epsilon_\lambda$. In a way, the tilting modules resemble the classical representations much more than general representations of $U^{\mathrm{res}}_\epsilon(\mathfrak{sl}_N\mathbb{C})$ do. However, there are two difficulties that prevent $\mathcal{T}ilt_\epsilon(\mathfrak{sl}_N\mathbb{C})$ from being modular as is: $\mathcal{Q}^\epsilon_\lambda$ are not exactly simple objects being indecomposable but not irreducible and there are still infinitely many of them. Based on ideas from physics or more likely ideas from algebraic groups Andersen was able to resolve both problems by discarding tilting modules of quantum dimension 0. This is explained in the next subsection.

17.3 Quantum dimensions and the Weyl alcove

We do not have a ribbon structure on $\mathcal{T}ilt_\epsilon(\mathfrak{sl}_N\mathbb{C})$ yet, so defining quantum traces and dimensions as it was done in Section 16.4 is not possible. Recall however that quantum traces were reinterpreted in Section 16.6 in terms of double duals. Namely, given a double dual isomorphism $V \xrightarrow{\delta_V} V^{**}$ one can set $\mathrm{Tr}_q(f) := \cap_{V^*} \circ (\delta_V \circ f \otimes \mathrm{id}_{V^*}) \circ \cup_V$ and $\dim_q(V) := \mathrm{Tr}_q(\mathrm{id}_V)$, where $\cap_{V^*}, \cup_V$ are the standard pairing and copairing for vector spaces. Any candidate for δ_V must of course be equivariant under the $U_\epsilon^{\mathrm{res}}$ action. The standard identification of V and V^{**} in the category of finite-dimensional vector spaces $v \mapsto [v]$ with $[v](\varphi) := \varphi(v)$ is not equivariant. Indeed, we have

$$(a[v])(\varphi) = [v](\gamma(a)\varphi) = (\gamma(a)\varphi)(v) = \varphi(\gamma^2(a)v) = (\gamma^2(a)[v])(\varphi)$$

and one easily sees from (32) that $\gamma^2 \neq \mathrm{id}$. The idea is to 'fix' $[\cdot]$ to make it equivariant.

Proposition 17.29 *Let $\rho := \frac{1}{2}\sum_{\alpha\in\Delta^+}\alpha$ be the Weyl weight (see Appendix D). Then for a in $U_q(\mathfrak{sl}_N\mathbb{C})$ or in $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$ we have*

$$\gamma^2(a) = q^{2\rho} a q^{-2\rho} \tag{41}$$

and the double dual map $\delta_V\colon V \to V^{**}$ given by $v \mapsto [q^{2\rho}v]$ is an equivariant isomorphism. Consequently,

$$\mathrm{Tr}_q(f) = \mathrm{Tr}(q^{2\rho} f), \tag{42}$$

where Tr *is the usual trace.*

Proof Since both sides of (41) are algebra homomorphisms it suffices to check the equality on the generators $q^{\pm\alpha^\vee{}_i}, e_i, f_i$. We have $\gamma^2(q^{\pm\alpha^\vee{}_i}) = q^{\mp\alpha^\vee{}_i}$ by (32) and the equality is obvious. For e_k we have from (31) and (32):

$$\gamma^2(e_k) = \gamma(-e_k q^{-\alpha^\vee{}_k}) = -\gamma(q^{-\alpha^\vee{}_k})\gamma(e_k) = q^{\alpha^\vee{}_k} e_k q^{-\alpha^\vee{}_k} = q^2 e_k.$$

On the other hand $q^{2\rho}e_k q^{-2\rho} = q^{(\alpha_k,2\rho)}e_k = q^2 e_k = \gamma^2(e_k)$ as claimed. (Recall from Appendix D that $\rho = \sum \omega_i$.) In $U_q(\mathfrak{sl}_N\mathbb{C})$ the equality for the divided powers $e_k^{(n)}$ follows by dividing both sides by $[n]_q!$. The case of $f_k^{(n)}$ is analogous. Since (41) does not involve any rational functions of q it remains valid in $U^{\mathrm{res}}_{\mathbb{Z}[q,q^{-1}]}(\mathfrak{sl}_N\mathbb{C})$ and therefore specializes to $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$. The equivariance is now straightforward

$$a\delta_V(v) = a[q^{2\rho}v] = [\gamma^2(a)q^{2\rho}v] = [q^{2\rho}av] = \delta_V(av).$$

Finally, for the quantum trace we get

$$\mathrm{Tr}_q(f) := \cap_{V^*} \circ (\delta_V \circ f \otimes \mathrm{id}_{V^*}) \Big(\sum_k v_k \otimes v^k \Big) = \cap_{V^*} \Big(\sum_k [q^{2\rho} f(v_k)] \otimes v^k \Big)$$
$$= \sum_k v^k \Big(q^{2\rho} f(v_k) \Big) = \mathrm{Tr}(q^{2\rho} f).$$

This completes the proof. □

Remark 17.30 Something interesting is happening here. We are starting with a representation of a deformation of the universal enveloping algebra of $\mathfrak{sl}_N$. In order to define the appropriate deformation we introduced elements such as q^β in place of the coroots. Furthermore, because we are exponentiating ($q = e^{2\pi i/l}$ after specialization giving $q^\beta = e^{2\pi i\beta/l}$) these are actually elements of the group $\mathrm{SU}(N)$ so the characteristic numbers that we compute here can be expressed in terms of characters of the group $\mathrm{SU}(N)$. The Weyl character formula can then be used to compute the resulting characters.

Recall from the previous article that the Weyl modules $\mathcal{W}^\epsilon_\lambda(\mathfrak{sl}_N\mathbb{C})$ have the same weight space decompositions as the classical representation spaces V_λ. The quantum dimension is the quantum trace of the identity, and it is given by the Weyl character formula (see Appendix D).

Corollary 17.31 *The quantum dimensions of the space $\mathcal{W}^\epsilon_\lambda$ is given by*

$$\dim_q \mathcal{W}^\epsilon_\lambda = \prod_{\alpha\in\Delta^+} \frac{\epsilon^{(\lambda+\rho,\alpha)} - \epsilon^{-(\lambda+\rho,\alpha)}}{\epsilon^{(\rho,\alpha)} - \epsilon^{-(\rho,\alpha)}}. \tag{43}$$

Exercise 17.32 Since $\epsilon^{-1} = \overline{\epsilon}$ it is obvious that $\dim_q \mathcal{W}^\epsilon_\lambda$ is real. Prove that in fact $\dim_q \mathcal{W}^\epsilon_\lambda \geq 0$ if ϵ is a primitive root of unity.

It is obvious from (43) that $\dim_q \mathcal{W}^\epsilon_\lambda = 0$ if and only if $(\lambda+\rho,\alpha)$ is divisible by l' for some $\alpha \in \Delta^+$. Therefore, for λ in the range $0 < (\lambda+\rho,\alpha) < l'$ for all positive roots α the quantum dimension of $\mathcal{W}^\epsilon_\lambda$ is non-zero.

Example 17.33 For the Weyl modules $\mathcal{W}^\epsilon_{(m-1)\omega}(\mathfrak{sl}_2\mathbb{C})$ from Example 17.20 the Weyl weight $\rho := \alpha/2$, where α is the only positive (and also simple) root of $\mathfrak{sl}_2\mathbb{C}$. Therefore

$$\big((m-1)\omega + \tfrac{\alpha}{2}, \alpha\big) = m - 1 + \tfrac{1}{2}(\alpha,\alpha) = m$$

and $\dim_q \mathcal{W}^{\epsilon}_{(m-1)\omega}(\mathfrak{sl}_2\mathbb{C}) = 0$ if and only if l' divides m. The condition $0 < (\lambda + \rho, \alpha) < l'$ is in turn equivalent to $0 < m < l'$. Note that by what we mentioned in Example 17.20 when $m = l'$ the Weyl module is still irreducible even though its quantum dimension is already 0.

It turns out that the Weyl modules with the highest weights in the range $0 < (\lambda + \rho, \alpha) < l'$ are the most important ones for our purposes.

Definition 17.34 (Weyl alcove) The (open) Weyl alcove is a subset of the Cartan subalgebra $\mathfrak{h}$ of $\mathfrak{sl}_N\mathbb{C}$ defined by

$$C^l := \{x \in \mathfrak{h} \mid 0 < (x + \rho, \alpha) < l', \text{ for all } \alpha \in \Delta^+\},$$

where $l' = l$ for l odd and $= l/2$ for l even. We also denote the set of dominant weights in the Weyl alcove by $\Lambda^l_w := C^l \cap \Lambda_w$ and call this the Weyl alcove when there is no confusion. The closure $\overline{C}^l$ is called the closed Weyl alcove. We will often abuse notation by saying that a Weyl module is or is not in the Weyl alcove according to the location of its highest weight.

We already saw that the modules in the alcove have non-zero quantum dimensions. In fact, much more is true (see Andersen [9], Andersen and Paradowski [10], Chari and Pressley [40], and Sawin [135]):

Theorem 17.35 *The Weyl alcove Λ^l_w contains precisely all the (highest weights of) Weyl modules $\mathcal{W}^{\epsilon}_{\lambda}$ that are both irreducible and have non-zero quantum dimensions. An indecomposable module $\mathcal{Q}^{\epsilon}_{\lambda}$ (see Theorem 17.28) has non-zero quantum dimension if and only if $\lambda \in \Lambda^l_w$ in which case $\mathcal{Q}^{\epsilon}_{\lambda} = \mathcal{W}^{\epsilon}_{\lambda}$. If $\lambda \in \Lambda^l_w$ then $\lambda^* := -w_0(\lambda) \in \Lambda^l_w$.*

Recall that $\mathcal{Q}^{\epsilon *}_{\lambda} \simeq \mathcal{Q}^{\epsilon}_{\lambda^*}$ so the Weyl alcove is closed under duality. In practice for $\mathfrak{sl}_N\mathbb{C}$, it is more convenient to write the defining condition of the alcove as follows

$$\begin{cases} 0 < (x + \rho, \alpha_i), & \text{for all simple roots } \alpha_i \\ (x + \rho, \theta) < l', & \text{for the highest positive root } \theta. \end{cases} \tag{44}$$

We proceed with some sample computations that make (44) more explicit.

Example 17.36 Recall from Appendix D that $L_i := E_{ii} - \frac{1}{N} I$ $i = 1, \dots, N-1$ form a basis in $\mathfrak{h} = \mathfrak{h}^*$ of $\mathfrak{sl}_N\mathbb{C}$. Given a weight it is convenient to write $\lambda := \sum_{i=1}^N \lambda_i L_i$ by setting $\lambda_N := 0$. The highest root is $\theta = E_{11} - E_{NN} = \sum_{i=1}^{N-1} \alpha_i$ with $\alpha_i =$

$E_{ii}-E_{i+1\,i+1}=L_i-L_{i+1}$. One easily checks that $(L_j,\alpha_i)=\delta_{ij}-\delta_{i+1\,j}$ and since $\rho=\sum_{i=1}^{N-1}\omega_i$ we have

$$\begin{aligned}(\lambda+\rho,\theta)&=\sum_{i=1}^{N-1}\sum_{j=1}^{N}\lambda_j(L_j,\alpha_i)+\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}(\omega_j,\alpha_i)\\&=\sum_{i=1}^{N-1}\sum_{j=1}^{N}\lambda_j(\delta_{ij}-\delta_{i+1\,j})+\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\delta_{ij}\\&=\sum_{i=1}^{N-1}(\lambda_i-\lambda_{i+1}+1)\\&=\lambda_1+N-1.\end{aligned}$$

Hence the second condition in (44) is equivalent to $\lambda_1\le l'-N$. By analogous computation the first condition reduces to $\lambda_i\ge\lambda_{i+1}$ for all i. This means that λ_i is a partition and the weights λ from the alcove are always dominant: $\Lambda_w^l\subset\Lambda_w^+$. Thus (44) reduces to λ being dominant and $0\le\lambda_i\le l'-N$ for all i. Note that in the case of Chern–Simons theory where $l=2(k+N)$ this leads to an N–independent condition $0\le\lambda_i\le k$. Another direct computation shows that $\lambda^*=\sum_{i=1}^{N}(\lambda_1-\lambda_{N-i+1})L_i$.

Exercise 17.37 Show that in the case of $\mathfrak{sl}_2\mathbb{C}$ we have $\lambda^*=\lambda$, that is, all Weyl modules are self-dual.

Exercise 17.38 The symmetric group $W=\mathfrak{S}_N$ acts on E_{ii} by $\sigma E_{ii}=E_{\sigma(i)\sigma(i)}$. Compute the induced action on L_i:

$$\sigma L_i=\begin{cases}L_{\sigma(i)}, & \sigma(i)<N\\ -\sum_{j=1}^{N-1}L_j, & \sigma(i)=N\end{cases}$$

and use it to derive a formula for $\lambda^*=-w_0(\lambda)$.

Exercise 17.39 Recall from Appendix D that one can also express any weight as a sum of fundamental weights $\lambda=\sum_{i=1}^{N-1}n_i\omega_i$ and $n_i\ge0$ for the dominant weights. Show that (44) is equivalent to $n_i\ge0$ and $\sum_{i=1}^{N-1}n_i\le l'-N$. Also in terms of n_is we have $\lambda^*=\sum_{i=1}^{N-1}n_{N-i}\omega_i$

Exercise 17.40 In the Cartan subalgebra $\mathfrak{h}$ of $\mathfrak{sl}_N\mathbb{C}$ neither α_i nor ω_i form an orthonormal basis. For $\mathfrak{sl}_3\mathbb{C}$ orthonormalize α_1,α_2 into u_1,u_2 and show that

$$\alpha_1=\sqrt{2}u_1,\qquad \alpha_2=\sqrt{2}\big(-\tfrac{1}{2}u_1+\tfrac{\sqrt{3}}{2}u_2\big),\qquad \theta=\alpha_1+\alpha_2=\sqrt{2}\big(\tfrac{1}{2}u_1+\tfrac{\sqrt{3}}{2}u_2\big).$$

Given $x = x_1u_1 + x_2u_2$ describe the condition $x \in C^l$ in terms of x_1, x_2 and draw a picture of C^l (cf. Figure 18.3 that shows $\rho + C^l$).

The properties of the Weyl alcove indicate how to proceed with the construction of a modular category. First of all, we finally have a finite set Λ^l_w of simple objects and they are indeed simple because alcove Weyl modules are irreducible. To make them dominate our category it suffices to consider only tilting modules that decompose into direct sums of alcove Weyl modules. Since dual alcove modules are still in the alcove the duals to such tilting modules will again decompose into direct sums of the alcove Weyl modules. This idyllic picture is spoiled by the behavior of the tensor product: the tensor product of two alcove modules may have non-alcove modules in its decomposition. This means that if we want a category with 'tensor products' the usual tensor product has to be redefined. The idea is to discard the submodule of the tensor product that has quantum dimension zero. Since $\mathcal{Q}^\epsilon_\lambda$ has positive dimension if and only if it is an alcove Weyl module keeping only the latter should do the trick.

Definition 17.41 Let $V = \oplus_{\lambda\in\Lambda^+_w} \left(\mathcal{Q}^\epsilon_\lambda\right)^{\oplus m_\lambda(V)}$ be a decomposition of a tilting module V into the indecomposables as given by Theorem 17.28. Then its reduction is defined by keeping only summands with highest weight in the Weyl alcove, that is,

$$\overline{V} := \oplus_{\lambda\in\Lambda^l_w} \left(\mathcal{Q}^\epsilon_\lambda\right)^{\oplus m_\lambda(V)}.$$

A tilting module is said to be reduced if $\overline{V} = V$ and negligible if $\overline{V} = 0$. The reduced tensor product of two reduced modules U, V is defined by

$$U \overline{\otimes} V := \overline{U \otimes V}.$$

Remark 17.42 Actually, we only defined $\overline{V}$ up to isomorphism. For a strict category one needs a canonical construction of it which goes as follows (see Chari and Pressley [40]). Let $\widetilde{V}$ be the maximal reduced submodule of V and V' its maximal negligible submodule then $\overline{V} = \widetilde{V}/(V' \cap \widetilde{V})$ is the canonical representative of the reduction. In physics literature reduced tensor product is often called fusion tensor product and rules for computing it are called fusion rules.

The following theorem also due to Andersen [9] (see also [40]) indicates that the new tensor product behaves 'properly'.

Theorem 17.43 *If V is any tilting module and U is negligible then so are $U \otimes V$ and $V \otimes U$. Thus, if U, V, W are reduced we have canonical isomorphisms $U \overline{\otimes} V \simeq V \overline{\otimes} U$ and $(U \overline{\otimes} V)\overline{\otimes} W \simeq U \overline{\otimes}(V \overline{\otimes} W)$*

Exercise 17.44 A morphism $U \xrightarrow{f} V$ of tilting modules is called neglible if it factors through a neglible tilting module. An equivalence class of equivariant morphisms is called a non-negligible morphism. Prove that f is neglible if and only if for any morphism $V \xrightarrow{g} U$ we have $\mathrm{Tr}_q(fg) = \mathrm{Tr}_q(gf) = 0$.

Remark 17.45 Note that we do not lose any information about link invariants by discarding tilting modules of quantum dimension zero. It follows from Definition 16.29 that the colored invariant $J_{V_1,\dots,V_k}(L)$ is 0 if any one of V_i is negligible. Indeed, we can always elongate the strand colored by V_i so that it cups under and caps over the rest of the link. The entire link then reduces to a morphism $V_i \xrightarrow{f_L} V_i$ with $J_{V_1,\dots,V_k}(L) = \mathrm{Tr}_q(f_L)$. But f_L is negligible along with V_i and so its quantum trace is zero. This serves as topological justification for discarding the negligible modules.

Definition 17.46 We denote the category of reduced tilting modules with equivariant linear maps by $\overline{\mathcal{T}ilt}_\epsilon(\mathfrak{sl}_N\mathbb{C})$.

Remark 17.47 There is an alternative construction of this category used eg in Bakalov–Kirillov [21]. One keeps the ordinary tensor product but takes morphisms to be equivalence classes of equivariant linear maps modulo negligible ones (so called non-negligible morphisms). The category so defined turns out to be isomorphic to $\overline{\mathcal{T}ilt}_\epsilon(\mathfrak{sl}_N\mathbb{C})$, see Chari–Pressley [40].

The category $\overline{\mathcal{T}ilt}_\epsilon(\mathfrak{sl}_N\mathbb{C})$ together with the duality and reduced tensor products is exactly what we were looking for. However, two key ingredients are still missing: the braiding and the twist. The latter can be restored from the former using dual cups and caps as in Section 16.6. But to get a braiding we need to introduce a new structure on $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$ known as quasitriangular structure or the R–matrix. After defining the braiding and twist we will see that for even roots of unity $\overline{\mathcal{T}ilt}_\epsilon(\mathfrak{sl}_N\mathbb{C})$ will turn out to be modular.

17.4 R–matrices and braiding

We are making good progress defining a concrete nontrivial modular tensor category. By considering appropriate representations of $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$ as representations of an associative algebra, we were able to construct the objects and morphisms of an abelian category $\overline{\mathcal{T}ilt}_\epsilon$ dominated by a finite collection of simple objects. The Hopf algebra structure of $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$, in particular the counit, the antipode and the coproduct were responsible for providing additional structures in the category of representations,

namely the unit object, the duality and the tensor product. In this article we define the appropriate braiding in our category.

The braiding requires something beyond the Hopf algebra structure. As motivation, recall that the (trivial) braiding on REP_G from Example 16.6 is just the flip $\times_{U,V}(u \otimes v) = v \otimes u$. Of course, given any Hopf algebra $\mathcal{A}$ we can define a braiding σ on $\mathcal{A} \otimes \mathcal{A}$ by $a \otimes b \mapsto b \otimes a$ and this will induce a trivial 'braiding' on any pair of representations. The catch is to find a necessarily nontrivial braiding compatible with the rest of the structure – in particular leading to the same definitions of quantum traces and dimensions as in the previous article. Drinfeld showed that for a finite-dimensional $\mathcal{A}$ the following notion works [51] (see also Chari–Pressley [40] and Majid [101]).

Definition 17.48 A Hopf algebra $\mathcal{A}$ is called quasitriangular if there exists an invertible element R (called the universal R–matrix) living in a certain completion $\mathcal{A}\widehat{\otimes}\mathcal{A}$ of the tensor square $\mathcal{A} \otimes \mathcal{A}$ that satisfies the following axioms

$$\begin{aligned} \sigma \circ \Delta &= R\Delta R^{-1}, \\ (\Delta \otimes \mathrm{id})(R) &= (\sigma \otimes \mathrm{id})(1 \otimes R) \cdot (1 \otimes R) \\ (\mathrm{id} \otimes \Delta)(R) &= (\sigma \otimes \mathrm{id})(1 \otimes R) \cdot (R \otimes 1). \end{aligned} \tag{45}$$

Here $\sigma(x \otimes y) := y \otimes x$. If $\mathcal{A}$ is finite-dimensional one can take $\mathcal{A}\widehat{\otimes}\mathcal{A} = \mathcal{A} \otimes \mathcal{A}$.

When a universal R–matrix exists it is essentially unique. Given a pair of representations $\rho_U, \rho_V \colon \mathcal{A} \to \mathrm{End}(U), \mathrm{End}(V)$ the universal R–matrix induces specializations

$$R_{U,V} \colon U \otimes V \to U \otimes V; \quad R_{U,V}(x \otimes y) := (\rho_U \otimes \rho_V)(x \otimes y)\,.$$

and the braiding will be composition with the flip $\times_{U,V} := \sigma \circ R_{U,V}$.

Remark 17.49 We have not defined the completed tensor product $\mathcal{A}\widehat{\otimes}\mathcal{A}$ because we will not use it. We will express the universal R–matrix as an infinite formal sum and it will not matter what space it lives in. The point is that when it is applied to any product of reduced tilting modules it will reduce to a finite sum. Infinite-dimensional Hopf algebras such as $U_q(\mathfrak{sl}_N\mathbb{C})$, $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$ rarely contain a universal R–matrix in $\mathcal{A} \otimes \mathcal{A}$. For our examples one really has to go out of one's way to construct a big enough extension $\mathcal{A}\widehat{\otimes}\mathcal{A}$ to contain R. In other words we are treating the universal R–matrix as the Cheshire cat whose smile lingers on after the cat is gone. After all, tilting modules are finite-dimensional and all we need are the induced R–matrices on finite-dimensional representations (the smile).

Aside 17.50 There is a standard way to construct a matrix that satisfies the conditions from the definition of a quasitriangular Hopf algebra. The opposite algebra of a Hopf algebra A is the algebra A_{op} with multiplication given by $a \cdot_{\mathrm{op}} b = b \cdot a$. One first constructs an object living in $A_{\mathrm{op}} \otimes A^*$ satisfying the correct relations and then takes the image of this object in the correct completed tensor square. If A is any Hopf algebra let $\{a_i\}$ be a basis, $\{a^i\}$ be the dual basis and set $R = \sum a_k \otimes a^k$. We will see that this satisfies the required property in $A_{\mathrm{op}} \otimes A^*$. Write the comultiplication in A in index notation as $\Delta(a_k) = \sum \gamma_k^{ij} a_i \otimes a_j$. This determines the product in A^* by

$$(a^i a^j)(a_k) = (a^i \otimes a^j)(\Delta(a_k)) = \gamma_k^{ij},$$

so $a^i a^j = \gamma_k^{ij} a^k$. The product in A_{op} is given by $a \cdot_{\mathrm{op}} b = ba$ and the comultiplication is the same Δ. Now compute

$$\begin{aligned}(\Delta \otimes \mathrm{id})(R) &= \sum \gamma_k^{ij} a_i \otimes a_j \otimes a^k \\ &= \sum \delta_{\ell m}^{ij} a_\ell \otimes a_m \otimes \gamma_k^{ij} a^k \\ &= \sum \delta_{\ell m}^{ij} a_\ell \otimes a_m \otimes a^i a^j \\ &= \sum a_i \otimes a_j \otimes a^i a^j = (\sigma \otimes \mathrm{id})(1 \otimes R) \cdot (1 \otimes R).\end{aligned}$$

The proof that the R–matrix that we are about to define satisfies the same equation is that there is a homomorphism taking the R–matrix constructed here to the complicated expression given below. This idea is the heart of the quantum double construction. Working out the details takes from page 128 to 273 of Chari and Pressley [40]. Drinfeld won a Fields medal in part for the work formalizing the idea of the quantum double. The quantized enveloping algebras $U_q(\mathfrak{sl}_N\mathbb{C})$, $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$ are not quantum doubles themselves but they are quotients of such and the R–matrices can be computed by 'projection' [40], see also Rosso [131]. Roughly speaking, if $U_q^+(U_q^-)$ denote the subalgebras generated by $q^{\pm\alpha_i}, e_i(q^{\pm\alpha_i}, f_i)$ then $U_q^- \simeq U_q^{+*}$ and U_q is a quotient of $D(U_q^+) \simeq U_q^+ \otimes U_q^-$. Note that $U_q^+(U_q^-)$ are quantum deformations of algebras of the upper (lower) triangular matrices hence the name quasitriangular.

For $U_q(\mathfrak{sl}_2\mathbb{C})$ a formal computation following the previous aside yields [40; 101]:

$$R = q^{\alpha\otimes\omega} \sum_{n=0}^{\infty} q^{n(n-1)/2} \frac{(q-q^{-1})^n}{[n]_q!} e^n \otimes f^n, \tag{46}$$

where the second term acts on $x \otimes y$ in the obvious way and

$$q^{\alpha\otimes\omega}(x \otimes y) := q^{(\alpha,\lambda)(\omega,\mu)} x \otimes y,$$

when x, y are weight vectors with weights λ, μ respectively. As before α is the simple root of $\mathfrak{sl}_2\mathbb{C}$ and $\omega = \alpha/2$ is the corresponding fundamental weight. This formula is to be understood as follows: on any finite-dimensional representation

e^n, f^n act nilpotently and only finite number of terms in the sum (46) are non-zero. Thus the induced matrices $R_{U,V}$ are well-defined for any pair of finite-dimensional representations even though the 'universal R' itself is just a formal expression.

Example 17.51 Recall from Example 17.20 that the action of $U_q(\mathfrak{sl}_2\mathbb{C})$ on the representation $\mathcal{V}^q_{(m-1)\omega}(\mathfrak{sl}_2\mathbb{C})$ spanned by $v_1,\dots,v_{m-1}$ is given by

$$(47)\quad q^{\pm\alpha}v_i = q^{\pm(m-2i)}v_i,\quad e^n v_i = \frac{[m-i+n]_q!}{[m-i]_q!}v_{i-n},\quad f^n v_i = \frac{[m-i+n]_q!}{[i]_q!}v_{i+n}.$$

Obviously e^n acts as 0 on v_i for $n > i$ and f^n acts as 0 on v_i for $n > m-1-i$. Therefore the sum in (46) applied to $v_i \otimes v_j$ truncates at $\min\{i, m-1-j\}$. By (47) v_i is a weight vector with the weight $(m-2i)\omega$ and therefore

$$\begin{aligned} q^{\alpha\otimes\omega}(v_{i-n}\otimes v_{j+n}) &= q^{(\alpha,(m-2i+2n)\omega)(\omega,(m-2j-2n)\omega)}v_{i-n}\otimes v_{j+n}\\ &= q^{\frac{1}{2}(m-2i+2n)(m-2j-2n)}v_{i-n}\otimes v_{j+n}. \end{aligned}$$

Substituting this into (46) we explicitly get

$$(48)\quad R(v_i\otimes v_j) = \sum_{n=0}^{\min\{i,m-1-j\}} q^{\frac{1}{2}((m-2i+2n)(m-2j-2n)+n(n-1))}\frac{(q-q^{-1})^n}{[n]_q!}\frac{[m-i+n]_q![j+n]_q!}{[m-i]_q![j]_q!}v_{i-n}\otimes v_{j+n}.$$

To make this more transparent we will compute this expression for $m = 3$ and some pairs i, j.

$$\begin{aligned} R(v_0\otimes v_0) &= q^{\frac{9}{2}}v_0\otimes v_0\\ R(v_1\otimes v_1) &= q^{\frac{1}{2}(3-2)(3-2)}v_1\otimes v_1 + q^{\frac{1}{2}(3-2+2)(3-2-2)}(q-q^{-1})[3]_q[2]_q v_0\otimes v_2\\ &= q^{\frac{1}{2}}v_1\otimes v_1 + q^{-\frac{3}{2}}(q-q^{-1})\frac{(q^3-q^{-3})(q^2-q^{-2})}{(q-q^{-1})^2}v_0\otimes v_2\\ &= q^{\frac{1}{2}}v_1\otimes v_1 + q^{-\frac{3}{2}}(q^3-q^{-3})(q+q^{-1})v_0\otimes v_2\\ R(v_0\otimes v_2) &= q^{\frac{1}{2}3(3-4)}v_0\otimes v_2 = q^{-\frac{3}{2}}v_0\otimes v_2. \end{aligned}$$

Exercise 17.52 Compute $R(v_i\otimes v_j)$ for $m = 2$, that is, for $V = \mathcal{V}^q_\omega(\mathfrak{sl}_2\mathbb{C})$. Show that in the lexicographic basis $v_0\otimes v_0, v_0\otimes v_1, v_1\otimes v_0, v_1\otimes v_1$ of $V\otimes V$ the $R_{V,V}$

is given by the matrix

$$\begin{pmatrix} q^2 & 0 & 0 & 0 \\ 0 & 1 & q^2-q^{-2} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & q^2 \end{pmatrix}.$$

Remark 17.53 A reader may notice that expressions in Example 17.51 involve fractional powers of q that do not belong to $\mathbb{C}(q)$ that is technically our field of coefficients. In fact this is a direct consequence of having the term $q^{\alpha\otimes\omega}$ in (46) that produces (ω,μ)th power of q when acting on $u\otimes v$ with v having weight μ. Since $\omega=\frac{\alpha}{2}$ this number is potentially a half-integer. In general, the analogous term produces $\frac{1}{\det(a_{ij})}$ powers of q, where a_{ij} is the Cartan matrix. For $\mathfrak{sl}_N\mathbb{C}$ an elementary computation shows that $\det(a_{ij})=N$ and for the R–matrix formula to make sense we have to extend the coefficients to $\mathbb{C}(q^{1/N})$. Formally, we now have to work in $U_{q^{1/N}}(\mathfrak{sl}_N\mathbb{C}) := U_q(\mathfrak{sl}_N\mathbb{C})\otimes_{\mathbb{C}(q)}\mathbb{C}(q^{1/N})$ instead of $U_q(\mathfrak{sl}_N\mathbb{C})$. If we apply the same process to $U_{q^{1/N}}(\mathfrak{sl}_N\mathbb{C})$ that was used in Definition 17.12 to obtain $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$ from $U_q(\mathfrak{sl}_N\mathbb{C})$ we will get the same algebra assuming that $q^{1/N}$ is specialized to the principal value of $\epsilon^{\frac{1}{N}} := e^{\frac{\pi i}{N(k+N)}}$. This allows us to continue doing all computations with the indeterminate q as explained in Section 17.2, rewrite the results in terms of the divided powers and Laurent polynomials and only then specialize to ϵ.

The formula (46) can not be specialized to $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_2\mathbb{C})$ directly because of the $[n]_q!$ in the denominator. This is easy to fix following our general ideology by noticing that

$$\frac{e^n\otimes f^n}{[n]_q!} = [n]_q!\, e^{(n)}\otimes f^{(n)}.$$

Thus for $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_2\mathbb{C})$ formula (46) is replaced by

$$R = q^{\alpha\otimes\omega}\sum_{n=0}^{\infty}\epsilon^{n(n-1)/2}(\epsilon-\epsilon^{-1})^n[n]_\epsilon!\, e^{(n)}\otimes f^{(n)}. \tag{49}$$

The inconvenience of computing the action of the divided powers for this formula can be bypassed as follows. Recall that we are interested in the Weyl modules $\mathcal{W}^\epsilon_{(m-1)\omega}(\mathfrak{sl}_2\mathbb{C})$ where m lies in the range $1\le m<(k+N)$ (the Weyl alcove). In this range $\frac{e^n\otimes f^n}{[n]_\epsilon!}$ still makes sense and (46) with $q=\epsilon$ can be used equivalently.

The appearance of fractional powers is not the only nuisance we have to deal with when moving on to $U_{q^{1/N}}(\mathfrak{sl}_N\mathbb{C})$. To write the R–matrix for $N>2$ we need analogs of the root vectors e_i, f_i for non-simple positive roots α. Unlike the classical case there is no canonical way to introduce such. A correct generalization comes from the

following classical observation [40]. Let $W(=\mathfrak{G}_N)$ be the Weyl group of $\mathfrak{sl}_N\mathbb{C}$ (=the symmetric group) generated by reflections $s_i = s_{\alpha_i}$ in the hyperplanes orthogonal to the simple roots α_i. In the basis E_{kk} these act as the transpositions $s_k = (k\ k+1)$. It is clear from this description that the Weyl group acts transitively on the root vectors $e_{ij} := E_{ii} - E_{jj}$.

More explicitly, each element in the Weyl group admits a presentation $\sigma = s_{i_1} \dots s_{i_\nu}$ as a 'word' in generators. A word representing an element is called reduced if it has the shortest possible length ν (such a word may not be unique). Let $w_0 \in W$ be the order-reversing permutation that we already met in connection with duality and $w_0 = s_{i_1} \dots s_{i_\nu}$ be a reduced word for it. Then each positive root occurs exactly once in the following sequence

$$\beta_1 := \alpha_{i_1},\ \beta_2 := s_{i_1}(\alpha_{i_2}), \dots, \beta_\nu := s_{i_1} \dots s_{i_{\nu-1}}(\alpha_{i_\nu}). \tag{50}$$

This gives a natural enumeration of the set of all positive roots. The standard choice of a reduced word for $\mathfrak{sl}_N\mathbb{C}$ is

$$w_0 = s_1(s_2s_1)(s_3s_2s_1)\dots(s_{N-1}\dots s_2s_1) \tag{51}$$

and it gives the anti-lexicographic (read from right to left) enumeration for the roots, namely

$$\begin{aligned} &\beta_1 = E_{11} - E_{22}, && \beta_2 = E_{11} - E_{33}, && \beta_3 = E_{22} - E_{33},\\ &\beta_4 = E_{11} - E_{44}, \dots, && \beta_6 = E_{33} - E_{44}, \dots, &&\\ &\beta_{\nu-N+2} = E_{11} - E_{NN}, \dots, && \beta_\nu = E_{N-1\,N-1} - E_{NN}. && \end{aligned} \tag{52}$$

Exercise 17.54 Verify the last claim.

We now consider how to define standard root vectors in the quantum setting. For this we need analogs of reflections s_i for quantum groups. They are given by the Lusztig automorphisms [40; 99].

Definition 17.55 Define algebra automorphisms $T_i\colon U_{q^{1/N}}(\mathfrak{sl}_N\mathbb{C}) \to U_{q^{1/N}}(\mathfrak{sl}_N\mathbb{C})$ by the following action on the generators

$$\begin{array}{rrr} i = j & |i-j| > 2 & j = i \pm 1 \\ \hline T_i q^{\pm\alpha_i} = q^{\mp\alpha_i} & T_i q^{\pm\alpha_j} = q^{\pm\alpha_j}, & T_i q^{\alpha_j} = q^{\alpha_j} q^{\alpha_i}, \\ T_i e_i = -f_i q^{\alpha_i} & T_i e_j = e_j, & T_i e_j = q^{-1} e_j e_i - e_i e_j, \\ T_i f_i = -q^{\alpha_i} e_i & T_i f_j = f_j, & T_i f_j = q f_i f_j - f_j f_i. \end{array} \tag{53}$$

For $\sigma \in W = \mathfrak{G}_N$ represented by the reduced word $\sigma = s_{i_1} \dots s_{i_\nu}$ set $T_\sigma := T_{i_1} \dots T_{i_\nu}$. The operators $T_i = T_{s_i}$, T_σ are called the Lusztig automorphisms.

Note that T_i and T_σ are only algebra, not Hopf algebra automorphisms (they do not preserve the coproduct). For general permutations T_σ are well-defined due to the following result of G Lusztig [99].

Theorem 17.56 *The action of T_σ depends only on σ and not on the choice of a reduced expression for it. All T_σ are invertible and $U^{\mathrm{res}}_{q^{1/N}}(\mathfrak{sl}_N\mathbb{C})$ is invariant under all of them. Furthermore, the following relations hold*

$$\begin{aligned} T_iT_j &= T_jT_i, \qquad |i-j|>1 \\ T_iT_{i+1}T_i &= T_{i+1}T_iT_{i+1}. \end{aligned} \tag{54}$$

Relations (54) are nothing other than the defining relations for the braid group $\mathcal{B}_N$ on N strands. In this group multiplication is concatenation of braids and inverse is the mirror image. The t_i generator of the braid group corresponds to a simple crossing between the i th and $(i+1)$st strands [40; 101]. Theorem 17.56 implies in particular that $t_i \mapsto T_i$ defines a representation of the braid group in $U_{q^{1/N}}(\mathfrak{sl}_N\mathbb{C})$. Also note that the forgetful map $\mathcal{B}_N \to \mathfrak{S}_N$ that only keeps track of the permutation on the strands is an infinite cover of the Weyl group of $\mathfrak{sl}_N\mathbb{C}$.

We now introduce the quantized versions of the root vectors corresponding to non-simple roots.

Definition 17.57 Let $\alpha \in \Delta^+$ be a positive root and $\alpha = \beta_k$ from (50). Then set

$$e^{(n)}_\alpha := T_{i_1}\dots T_{i_{k-1}}(e^{(n)}_{i_k}), \quad f^{(n)}_\alpha := T_{i_1}\dots T_{i_{k-1}}(f^{(n)}_{i_k}), \tag{55}$$

where $i_1,\dots,i_k$ are as in the reduced expression (51) for w_0. Naturally, we denote $e_\alpha := e^{(1)}_\alpha$, $f_\alpha := f^{(1)}_\alpha$.

Exercise 17.58 Show that if $\alpha = \alpha_i$ is simple then $e^{(n)}_\alpha = e^{(n)}_i$, and $f^{(n)}_\alpha = f^{(n)}_i$ as expected.

With this notation we can now write down formulas for the R–matrices for $U_{q^{1/N}}(\mathfrak{sl}_N\mathbb{C})$ and $U^{\mathrm{res}}_{q^{1/N}}(\mathfrak{sl}_N\mathbb{C})$ that generalize (46) and (49):

$$\begin{aligned} R &= q^{\sum_{i=1}^{N-1}\alpha_i\otimes\omega_i}\sum_{n_1,\dots,n_\nu=0}^{\infty}\ \prod_{\alpha\in\Delta^+}^{\leftarrow} q^{\frac{n_k(n_k-1)}{2}}\frac{(q-q^{-1})^{n_k}}{[n_k]_q!}\,e_\alpha^{n_k}\otimes f_\alpha^{n_k}, \\ &= q^{\sum_{i=1}^{N-1}\alpha_i\otimes\omega_i}\sum_{n_1,\dots,n_\nu=0}^{\infty}\ \prod_{\alpha\in\Delta^+}^{\leftarrow} q^{\frac{n_k(n_k-1)}{2}}(q-q^{-1})^{n_k}[n_k]_q!\,e_\alpha^{(n_k)}\otimes f_\alpha^{(n_k)} \end{aligned} \tag{56}$$

The product in (56) is not commutative and should be computed 'in reverse' to (52), that is, $\beta_\nu, \dots, \beta_1$ so that β_1 term is applied to the tensor product first (hence the $\leftarrow$). The R–matrix for $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$ is obtained from the second line in (56) by replacing q with ϵ [40].

Example 17.59 In $\mathfrak{sl}_3\mathbb{C}$ the simple roots are $\alpha_1 = E^*_{11} - E^*_{22}$ and $\alpha_2 = E^*_{22} - E^*_{33}$ and the only other positive root is $\alpha_1+\alpha_2 = E^*_{11} - E^*_{33}$. In the ordering of (52) we have $\beta_1 = \alpha_1, \beta_2 = \alpha_1+\alpha_2, \beta_3 = \alpha_2$. Hence $e^{(n)}_{\beta_1} = e^{(n)}_1, e^{(n)}_{\beta_3} = e^{(n)}_2$ and $e^{(n)}_{\beta_2} = T_1(e^{(n)}_2)$. For $n = 1$ we get from (53)

$$T_1(e_2) = q^{-1}e_2e_1 - e_1e_2 = -e_1e_2 + q^{-1}e_2e_1$$

and because T_1 is an automorphism

$$T_1(e_2^2) = T_1(e_2)^2 = (e_1e_2e_1e_2 + q^{-2}e_2e_1e_2e_1) - q^{-1}(e_1e_2^2e_1 + e_2e_1^2e_2). \tag{57}$$

We now want to rewrite this in terms of the divided powers and Laurent polynomials in accordance with the general strategy from Remark 17.53. Taking into account that $q + q^{-1} = [2]_q$ we have by the quantum Serre relations from (31) $e_2e_1e_2 = \frac{1}{[2]_q}(e_2^2e_1 + e_1e_2^2)$ and therefore

$$\begin{aligned} e_1(e_2e_1e_2) + q^{-2}(e_2e_1e_2)e_1 &= \frac{e_1}{[2]_q}(e_2^2e_1 + e_1e_2^2) + q^{-2}(e_2^2e_1 + e_1e_2^2)\frac{e_1}{[2]_q} \\ &= \frac{e_1^2e_2^2}{[2]_q} + q^{-2}\frac{e_2^2e_1^2}{[2]_q} + (1+q^{-2})\frac{e_1e_2^2e_1}{[2]_q} \\ &= \frac{e_1^2e_2^2}{[2]_q} + q^{-2}\frac{e_2^2e_1^2}{[2]_q} + q^{-1}e_1e_2^2e_1. \end{aligned}$$

Substituting into (57) yields

$$T_1(e_2^2) = \frac{e_1^2e_2^2}{[2]_q} + q^{-2}\frac{e_2^2e_1^2}{[2]_q} - q^{-1}e_2e_1^2e_2,$$

and dividing both sides by $[2]_q$

$$e^{(2)}_{\beta_2} = T_1(e^{(2)}_2) = e^{(2)}_1e^{(2)}_2 - q^{-1}e_2e^{(2)}_1e_2 + q^{-2}e^{(2)}_2e^{(2)}_1. \tag{58}$$

Analogously,

$$f^{(2)}_{\beta_2} = T_1(f^{(2)}_2) = q^2e^{(2)}_1e^{(2)}_2 - qf_2f^{(2)}_1f_2 + f^{(2)}_2f^{(2)}_1. \tag{59}$$

Formulas for $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$ can now be obtained by setting $q = \epsilon$.

Exercise 17.60 Prove by induction that

$$e_{\beta_2}^{(n)} = \sum_{k=0}^{n} q^{-k} e_2^{(k)} e_1^{(n)} e_2^{(n-k)}$$

and derive an analogous formula for $f_{\beta_2}^{(n)}$, see [99].

Example 17.61 We will now compute the action of the R–matrix of the fundamental representation in the $\mathfrak{sl}_3\mathbb{C}$ case. Recall from Exercise 17.15 that the fundamental representation $\mathcal{V}_{\omega_1}^q(\mathfrak{sl}_3\mathbb{C})$ has a basis of weight vectors $v_1 := u_0$, $v_2 := f_1 u_0$, $v_3 := f_2 v_2$ having weights ω_1, $\omega_1 - \alpha_1$ and $\omega_1 - \alpha_1 - \alpha_2$ respectively. This specifies the action of the q^{α_i}. The remainder of the action is specified by $f_i v_j = \delta_{ij} v_{j+1}$ and $e_i v_j = \delta_{i+1\,j} v_{j-1}$. Continuing with the computation from the previous example we have

$$\begin{aligned} e_{\beta_1} &= e_1, & e_{\beta_2} &= -e_1 e_2 + q^{-1} e_2 e_1, & e_{\beta_3} &= e_2, \\ f_{\beta_1} &= f_1, & f_{\beta_2} &= -f_2 f_1 + q f_1 f_2, & f_{\beta_3} &= f_2. \end{aligned}$$

Since v_1 is annihilated by e_1 and e_2, only the $q^{\sum \alpha_i \otimes \omega_i}$ factor of the R matrix acts on vectors of the form $v_1 \otimes x$ (see equation (56)). Since the weight of the fundamental representation is $\omega_1 = \square = \frac{2}{3}E_{11}^* - \frac{1}{3}E_{22}^* - \frac{1}{3}E_{33}^*$, it follows that

$$\begin{aligned} R(v_1 \otimes v_1) &= q^{\langle \alpha_1, \omega_1 \rangle \langle \omega_1, \omega_1 \rangle} v_1 \otimes v_1 = q^{2/3} v_1 \otimes v_1, \\ R(v_1 \otimes v_2) &= q^{\langle \alpha_1, \omega_1 \rangle \langle \omega_1, \omega_1 - \alpha_1 \rangle} v_1 \otimes v_2 = q^{-1/3} v_1 \otimes v_2, \\ R(v_1 \otimes v_3) &= q^{\langle \alpha_1, \omega_1 \rangle \langle \omega_1, \omega_1 - \alpha_1 \rangle} v_1 \otimes v_3 = q^{-1/3} v_1 \otimes v_3. \end{aligned}$$

Now

$$e_{\beta_1} v_2 = v_1, \qquad e_{\beta_2} v_2 = 0, \qquad e_{\beta_3} v_2 = 0,$$

so the only terms that contribute to the R–matrix evaluated on vectors of the form $v_2 \otimes x$ will correspond to the root β_1. We compute

$$f_{\beta_1} v_1 = v_2, \qquad f_{\beta_1} v_2 = 0, \qquad e_{\beta_1} v_3 = 0.$$

It follows that

$$\begin{aligned}
R(v_2\otimes v_1) &= q^{\sum \alpha_i\otimes\omega_i}(v_2\otimes v_1+(q-q^{-1})v_1\otimes v_2)\\
&= q^{\langle\alpha_1,\omega_1-\alpha_1\rangle\langle\omega_1,\omega_1\rangle+\langle\alpha_2,\omega_1-\alpha_1\rangle\langle\omega_2,\omega_1\rangle}v_2\otimes v_1\\
&\quad +(q-q^{-1})q^{\langle\alpha_1,\omega_1\rangle\langle\omega_1,\omega_1-\alpha_1\rangle+\langle\alpha_2,\omega_1\rangle\langle\omega_2,\omega_1-\alpha_1\rangle}v_1\otimes v_2,\\
&= q^{-1/3}v_2\otimes v_1+q^{-1/3}(q-q^{-1})v_1\otimes v_2,\\
R(v_2\otimes v_2) &= q^{\langle\alpha_1,\omega_1-\alpha_1\rangle\langle\omega_1,\omega_1-\alpha_1\rangle+\langle\alpha_2,\omega_1-\alpha_1\rangle\langle\omega_2,\omega_1-\alpha_1\rangle}v_2\otimes v_2\\
&= q^{2/3}v_2\otimes v_2,\\
R(v_2\otimes v_3) &= q^{\langle\alpha_1,\omega_1-\alpha_1\rangle\langle\omega_1,\omega_1-\alpha_1-\alpha_2\rangle+\langle\alpha_2,\omega_1-\alpha_1\rangle\langle\omega_2,\omega_1-\alpha_1-\alpha_2\rangle}v_2\otimes v_3\\
&= q^{-1/3}v_2\otimes v_3.
\end{aligned}$$

Now

$$\begin{aligned}
&e_{\beta_1}v_3=0, && e_{\beta_2}v_3=-v_1, && e_{\beta_3}v_3=v_2,\\
&f_{\beta_2}v_1=-v_3, && f_{\beta_2}v_2=0, && f_{\beta_2}v_3=0,\\
&f_{\beta_3}v_1=0, && f_{\beta_3}v_2=v_3, && f_{\beta_3}v_3=0.
\end{aligned}$$

It follows that

$$\begin{aligned}
R(v_3\otimes v_1) &= q^{\sum \alpha_i\otimes\omega_i}(v_3\otimes v_1+(q-q^{-1})v_1\otimes v_3)\\
&= q^{\langle\alpha_1,\omega_1-\alpha_1-\alpha_2\rangle\langle\omega_1,\omega_1\rangle+\langle\alpha_2,\omega_1-\alpha_1-\alpha_2\rangle\langle\omega_2,\omega_1\rangle}v_3\otimes v_1\\
&\quad +(q-q^{-1})q^{\langle\alpha_1,\omega_1\rangle\langle\omega_1,\omega_1-\alpha_1-\alpha_2\rangle}v_1\otimes v_3,\\
&= q^{-1/3}v_3\otimes v_1+q^{-1/3}(q-q^{-1})v_1\otimes v_3,\\
R(v_3\otimes v_2) &= q^{\sum \alpha_i\otimes\omega_i}(v_3\otimes v_2+(q-q^{-1})v_2\otimes v_3)\\
&= q^{\langle\alpha_1,\omega_1-\alpha_1-\alpha_2\rangle\langle\omega_1,\omega_1-\alpha_1\rangle+\langle\alpha_2,\omega_1-\alpha_1-\alpha_2\rangle\langle\omega_2,\omega_1-\alpha_1\rangle}v_3\otimes v_2\\
&\quad +(q-q^{-1})q^{\langle\alpha_1,\omega_1-\alpha_1\rangle\langle\omega_1,\omega_1-\alpha_1-\alpha_2\rangle+\langle\alpha_2,\omega_1-\alpha_1\rangle\langle\omega_2,\omega_1-\alpha_1-\alpha_2\rangle}v_2\otimes v_3,\\
&= q^{-1/3}v_3\otimes v_2+q^{-1/3}(q-q^{-1})v_2\otimes v_3,\\
R(v_3\otimes v_3) &= q^{\langle\alpha_1,\omega_1-\alpha_1-\alpha_2\rangle\langle\omega_1,\omega_1-\alpha_1-\alpha_2\rangle+\langle\alpha_2,\omega_1-\alpha_1-\alpha_2\rangle\langle\omega_2,\omega_1-\alpha_1-\alpha_2\rangle}v_3\otimes v_3\\
&= q^{2/3}v_3\otimes v_3.
\end{aligned}$$

Ordering the basis for $\mathcal{V}^q_{\omega_1}(\mathfrak{sl}_3\mathbb{C}) \otimes \mathcal{V}^q_{\omega_1}(\mathfrak{sl}_3\mathbb{C})$ lexicographically we may write the matrix for R as

$$R = q^{-1/3}\begin{bmatrix} q & & & & & & & & \\ & 1 & (q-q^{-1}) & & & & & & \\ & & 1 & & & (q-q^{-1}) & & & \\ & & & 1 & & & & & \\ & & & & q & & & & \\ & & & & & 1 & & (q-q^{-1}) & \\ & & & & & & 1 & & \\ & & & & & & & 1 & \\ & & & & & & & & q \end{bmatrix}.$$

This last example can be generalized to the fundamental representation $\mathcal{V}^q_{\omega_1}(\mathfrak{sl}_N\mathbb{C})$ for all N.

Exercise 17.62 In general the fundamental representation has a basis with highest weight vector v_1 and $v_{j+1} := f_j v_j$. Show that in this basis we have, $\omega_1(e_j) = E_{j\,j+1}$, $\omega_1(f_j) = E_{j+1\,j}$, $\omega_1(q^{\alpha_i^\vee}) = (q-1)E_{ii} + (q^{-1}-1)E_{i+1\,i+1} + I$, and

$$R = q^{-1/N}\left(q\sum E_{ii}\otimes E_{ii} + \sum_{i\neq j} E_{ii}\otimes E_{jj} + (q-q^{-1})\sum_{i<j} E_{ij}\otimes E_{ji}\right).$$

The answer to this exercise can be found on page 277 of [40]. It is important because it provides the link between the quantum invariants as we are defining them, and the THOMFLYP polynomial and skein theory as sketched in Appendix C.

It follows from Exercise D.7 that the representation

$$\mathcal{V}^q_{\ydiagram{3,1}}(\mathfrak{sl}_4\mathbb{C})$$

for $U_q(\mathfrak{sl}_4\mathbb{C})$ has 126 nontrivial weight spaces, we know that the dimension of

$$\mathcal{V}^q_{\ydiagram{3,3}}(\mathfrak{sl}_4\mathbb{C})$$

is 126. It follows that the R–matrix is a 15876×15876 matrix. As the reader can see, computing the R–matrices explicitly is no mean feat even for simple representations. The next exercise is one of the few remaining cases that can be worked out reasonably by hand.

Exercise 17.63 Compute a basis for the representation

$$\mathcal{V}^q_{\square}(\mathfrak{sl}_3\mathbb{C}).$$

Compute the action of q^{α_i}, $e_i^{(n)}$, $f_i^{(n)}$, and the R–matrix.

Recall that when $U_q(\mathfrak{sl}_N\mathbb{C})$ acts on a vector space V, we can define an action on the dual V^* by

$$(ag)(v) := g(\gamma(a)v).$$

To see how this works, consider the action of $U_q(\mathfrak{sl}_2\mathbb{C})$ on the dual of the representation

$$\mathcal{V}^q_{\square}(\mathfrak{sl}_2\mathbb{C}).$$

Example 17.64 Recall that the representation

$$\mathcal{V}^q_{\square}(\mathfrak{sl}_2\mathbb{C})$$

has a basis v_1, v_2. We denote the dual basis by v^1, v^2. Since $(ev^1)(v_1)=v^1(\gamma(e)v_1)=0$ and $(ev^1)(v_2) = v^1(\gamma(e)v_2) = -q$, we have $ev^1 = -qv^2$. Similarly, $ev^2 = 0$, $fv^1 = 0$ and $fv^2 = -q^{-1}v^1$. We can now see how the R–matrix acts on terms that include functionals such as $\omega_1 \otimes \omega_1^*$ in the $\mathfrak{sl}_2$ case. We have

$$\begin{aligned}
R(v_1 \otimes v^1) &= q^{-1/2} v_1 \otimes v^1 \\
R(v_1 \otimes v^2) &= q^{1/2} v_1 \otimes v^2 \\
R(v_2 \otimes v^1) &= q^{1/2} v_2 \otimes v^1 \\
R(v_2 \otimes v^2) &= q^{-1/2} v_2 \otimes v^2 - q^{-3/2}(q-q^{-1}) v_1 \otimes v^1.
\end{aligned}$$

Recall that we can define link and framed link invariants from a ribbon category. The R–matrix is the main tool to build a ribbon category from the representations of a quantum group. In fact the category of type I representations of $U_{q^{1/N}}(\mathfrak{sl}_N\mathbb{C})$ becomes a ribbon category if we use the braiding given by $\times_{u,V} = \sigma \circ R_{U,V}$. We denote the (framed) link invariants resulting from this category by $W_\Lambda^{\mathfrak{sl}_N}$ where Λ is a collection of representations (one for each component of the link). We next compute the invariant $W^{\mathfrak{sl}_2}_{\square,\square}$(left Hopf link) from the definition. First compute the composition of the top three morphisms from Figure 16.2. We have

$$\begin{aligned}
&v_1 \otimes v^1 \mapsto q^{3/2} v_1 \otimes v^1 \mapsto q v^1 \otimes v_1 \mapsto q \\
&v_2 \otimes v^2 \mapsto q^{3/2} v_2 \otimes v^2 \mapsto q^{3/2}[q^{-1/2} v^2 \otimes v_2 - q^{-3/2}(q-q^{-1}) v^1 \otimes v_1] \mapsto q^{-1}.
\end{aligned}$$

We also have $v_1 \otimes v^2 \mapsto 0$ and $v_2 \otimes v^1 \mapsto 0$. Notice that this recreates the double dual isomorphism. To continue the computation of the invariant for the left Hopf link it is helpful to write out the matrix for the braiding in the lexicographic basis for $\mathcal{V}^q_{\square}(\mathfrak{sl}_2\mathbb{C}) \otimes \mathcal{V}^q_{\square}(\mathfrak{sl}_2\mathbb{C})$ from Exercise 17.62. This is our first sample computation of a braiding.

Example 17.65 We have

$$\times_{\square,\square} = q^{-1/2}\begin{bmatrix} q & & & \\ & 0 & 1 & \\ & 1 & (q-q^{-1}) & \\ & & & q \end{bmatrix}, \quad \times^{-2}_{\square,\square} = q\begin{bmatrix} q^{-2} & & & \\ & (q^{-1}-q)^2+1 & (q^{-1}-q) & \\ & (q^{-1}-q) & 1 & \\ & & & q^{-2} \end{bmatrix}.$$

We can combine our computations of the twist, double dual pairing and braiding to compute the invariant of the Hopf link. We use v_i^k to denote $v_i \otimes v^k$ and $v_{ij}^{k\ell}$ to denote $v_i \otimes v_j \otimes v^k \otimes v^\ell$. The computation is contained in the following example.

Example 17.66 Following the element 1 up through the morphisms as depicted in Figure 16.2 gives

$$\begin{aligned} 1 &\mapsto v_1^1 + v_2^2 \\ &\mapsto v_{11}^{11} + v_{12}^{21} + v_{21}^{12} + v_{22}^{22} \\ &\mapsto q^{-1}v_{11}^{11} + (q(q-q^{-1})^2+q)v_{12}^{21} + (1-q^2)v_{21}^{21} + (1-q^2)v_{12}^{12} + qv_{21}^{12} + q^{-1}v_{22}^{22} \\ &\mapsto v_1^1 + ((q-q^{-1})^2+1)v_1^1 + q^2v_2^2 + q^{-2}v_2^2 \\ &\mapsto q + q(q-q^{-1})^2 + q + q + q^{-3} = q^3 + q + q^{-1} + q^{-3}. \end{aligned}$$

We conclude that

$$W^{\mathfrak{sl}_2}_{\square,\square}(\text{left Hopf link}) = q^3 + q + q^{-1} + q^{-3}.$$

We can use this type of computation to derive an interesting recurrence relation for the framed invariants. Notice that the twist in the fundamental representation is given by

$$\begin{aligned} \times_{\square,\square}(v_{ii}) &= q^{-1/N}qv_{ii} \\ \times_{\square,\square}(v_{ij}) &= q^{-1/N}v_{ji}, \quad i<j \\ \times_{\square,\square}(v_{k\ell}) &= q^{-1/N}(v_{\ell k} + (q-q^{-1})v_{k\ell}), \quad k>\ell. \end{aligned}$$

It follows that

$$\begin{aligned}
\times^{-1}_{\square,\square}(v_{ii}) &= q^{1/N}q^{-1}v_{ii}\\
\times^{-1}_{\square,\square}(v_{ji}) &= q^{1/N}v_{ij}, \quad i<j\\
\times^{-1}_{\square,\square}(v_{\ell k}) &= q^{1/N}(v_{k\ell}-(q-q^{-1})v_{\ell k}), \quad k>\ell.
\end{aligned}$$

Exercise 17.67 Use the previous computations to show that

$$q^{1/N}\times_{\square,\square}-q^{-1/N}\times^{-1}_{\square,\square}=(q-q^{-1})\mathrm{id}_{\square,\square}\,.$$

Notice that this implies that the framed link invariant $W^{\mathfrak{sl}_N}$ (all representations taken to by fundamental) satisfies a skein relation. Also notice that the invariant defined via quantum groups is provably invariant under isotopies. A framed link invariant is very close to being a link invariant; it is invariant under Reidemeister moves two and three from Figure 13.1. It follows from Exercise 18.7 below that

$$W^{\mathfrak{sl}_N}(\mathrm{closure}(\theta_|\circ f))=q^{N-1/N}W^{\mathfrak{sl}_N}(\mathrm{closure}(f))\,.$$

Here closure is just the quantum trace in the category of framed tangles; see Figure 16.6. This motivates the following definition of the THOMFLYP polynomial.

Definition 17.68 The THOMFLYP polynomial is defined by

$$P(L):=q^{(1/N-N)\sum_{i=1}^{c(L)}\sum_{j=1}^{c(L)}n_{ij}(L)}W^{\mathfrak{sl}_N}(L)\,.$$

Here $c(L)$ is the number of components of L and $n_{ij}(L)$ is the linking matrix of L.

Remark 17.69 To define 3–manifold invariants we have to go to the category or reduced tilting modules $\overline{Tilt}_\epsilon(\mathfrak{sl}_N\mathbb{C})$. Of course this is also a ribbon category that generates link invariants. The resulting invariants are just the evaluation of $P(L)$ at $q^{1/N}=e^{\pi i/(N(k+N))}$ In this way $P(L)$ is a function of N and a primitive even root of unity $\epsilon=e^{\pi i/(k+N)}$. One can show [40] that for each L there is a unique rational function in variables $\lambda^{1/2}$, $\mathfrak{q}^{1/2}$ denoted by $\mathcal{P}(L)$ such that $P(L)=\mathcal{P}(L)(\lambda=\epsilon^{2N},\mathfrak{q}=\epsilon^2)$.

Exercise 17.70 From Exercise 18.4 we know that

$$\mathcal{P}(\text{unknot})=\frac{\lambda^{1/2}-\lambda^{-1/2}}{\mathfrak{q}^{1/2}-\mathfrak{q}^{-1/2}}$$

Prove that the THOMFLYP polynomial satisfies the following skein relation

$$\lambda^{1/2}\mathcal{P}(\mathrm{closure}(\times_{|,|}\circ f))-\lambda^{-1/2}\mathcal{P}(\mathrm{closure}(\times^{-1}_{|,|}\circ f))=(\mathfrak{q}^{1/2}-\mathfrak{q}^{-1/2})\mathcal{P}(\mathrm{closure}(f)).$$

This (skein) recurrence relation together with the value of the unknot specifies the link invariant uniquely. In fact one can turn the entire process around and use the skein relation and normalization as the definition of the link invariants. It is not obvious that such a definition is well-formed. One must show that different ways of applying the skein relation lead to the same answers and one must show the resulting quantity is invariant under the Reidemeister moves. In addition to the approach to the definition that we used via quantum groups there are several different approaches to do this. The first are described in the original paper [60].

Much more is true – the full Reshetikhin–Turaev invariants of links in general 3–manifolds can be recovered from the skein theory. This is described in Appendix C.

Exercise 17.71 Compute $\mathcal{P}$(Right Hopf link) from the definition and via the skein relation and reconcile the two answers.

18 Reshetikhin–Turaev invariants from quantum groups

We are now in a position to combine all of the facts we derived about representations of quantum groups into a definition of a nontrivial strict modular category. This is the category of reduced tilting modules at even roots of unity based on the Lie algebra $\mathfrak{sl}_N\mathbb{C}$. Historically, Reshetikhin and Turaev constructed their invariants directly [128; 129] and only later Turaev [152] streamlined their construction into a notion of modular category. In this section we will finally define these invariants. First we review the main features of $\overline{Tilt}_\epsilon(\mathfrak{sl}_N\mathbb{C})$, then compute its characteristic numbers, the $\widetilde{s}$–matrix and finally the Reshetikhin–Turaev invariant of S^3, aka the Chern–Simons partition function.

18.1 Modular category of reduced tilting modules at even roots of unity

In this article we summarize the modular category structure on $\overline{Tilt}_\epsilon(\mathfrak{sl}_N\mathbb{C})$ when ϵ is a primitive *even* root of unity. This is the modular category that produces the Reshetikhin–Turaev (quantum) invariants of 3–manifolds along the lines of Section 16.7. After discussing the main ingredients we compute some of the characteristic numbers of this category (see Definition 16.23). It is assumed throughout that $\epsilon = e^{\frac{2\pi i}{l}}$, where l is *even*. To draw a connection with the Chern–Simons theory set $k := l/2 - N$ for the level of the theory. Then $l = 2(k+N)$, $l' = k+N$ and $\epsilon = e^{\frac{\pi i}{k+N}}$. One should keep in mind that there is no mathematical definition of Witten's path integral, hence no rigorous way to compare Witten's invariants to the Reshetikhin–Turaev ones and the above connection is merely conjectural.

Objects: The objects are reduced tilting modules V, that is, representations of $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$ (Definition 17.12) that are finite direct sums of Weyl modules (Definition 17.19) with the highest weights in the Weyl alcove Λ_w^l (see Definition 17.34 or **Index set** below): $V = \oplus_{\lambda\in\Lambda_w^l}\left(\mathcal{W}_\lambda^\epsilon\right)^{\oplus m_\lambda(V)}$.

Morphisms: The morphisms are equivariant linear maps $f(av) = a(f(v))$.

Unit object: The unit object is $\mathbb{1} := \mathbb{C}$ with the trivial representation structure $az := \varepsilon(a)z$, where ε is the counit (32) of $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$.

Dual objects: The dual objects are the vector space duals V^* of tilting modules with the action given by the antipode (32) $(a\varphi)(v) := \varphi(\gamma(a)z)$.

Pairing and copairing: The pairing and copairing are the usual ones from the category of finite-dimensional vector spaces

$$\cap_V(\varphi\otimes v) := \varphi(v), \qquad \cup_V(1) := \sum_k v_k\otimes v^k,$$

where v_k, v^k are dual bases in V, V^* respectively.

Tensor product: The tensor product is the reduced tensor product $\overline{\otimes}$ from Definition 17.23, that is, the maximal invariant subspace $U\overline{\otimes}V$ of the ordinary tensor product $U\otimes V$ that is a direct sum of Weyl modules with highest weights in the Weyl alcove. The action restricts from the action on the usual tensor product $a(u\otimes v) := \Delta(a)(u\otimes v)$ on $U\otimes V$.

Braiding: The braiding is $\times_{U,V} := \sigma\circ R_{U,V}$, where $\sigma(u\otimes v) := v\otimes u$ and $R_{U,V}$ is the restriction of the 'universal R–matrix' given by (56) to the ordinary tensor product of U, V.

Twist: The twist is obtained from the braiding, duality and double duality as in Figure 16.8, explicitly

$$\theta_V := (\mathrm{id}_V\otimes\cap_V^*)\circ(\times_{V,V}\otimes\mathrm{id}_{V^*})\circ(\mathrm{id}_V\otimes\cup_V),$$

where $\cap_V^*(v\otimes\varphi) := \varphi(q^{2\rho}v)$.

Index set: The index set of simple objects I is the set of dominant weights (see Appendix D) in the open Weyl alcove

$$I := \Lambda_w^l = \{\lambda\in\Lambda_w^+ \mid (\lambda+\rho,\alpha) < l/2, \text{ for all } \alpha\in\Delta^+\}.$$

The unit object is indexed by $\lambda = 0$.

Index involution: The index involution is given by $\lambda^* := -w_0(\lambda)$ (see Example 17.36), where w_0 is the order-reversing permutation and the action of the symmetric group $W = \mathfrak{S}_N$ on weights is described in Appendix D. In

particular, $0^* = 0$. The dual weight is the weight of the dual simple object: $\mathcal{W}^\epsilon_\lambda(\mathfrak{sl}_N\mathbb{C})^* = \mathcal{W}^\epsilon_{\lambda^*}(\mathfrak{sl}_N\mathbb{C})$.

Simple objects: The simple objects are the Weyl modules $\mathcal{W}^\epsilon_\lambda(\mathfrak{sl}_N\mathbb{C})$ with highest weights in the Weyl alcove $\lambda \in \Lambda^l_w$ (alcove Weyl modules).

For convenience we list the salient points of the definition in Table 18.1.

Category	$\overline{Tilt}_\epsilon(\mathfrak{sl}_N\mathbb{C})$, $\epsilon := e^{2\pi i/l}$, l even
Objects	Reduced tilting modules $V = \oplus_{\lambda\in\Lambda^l_w} \left(\mathcal{W}^\epsilon_\lambda\right)^{\oplus m_\lambda(V)}$
Morphisms	Equivariant linear maps $f(au) = af(u)$
Unit object	$\mathbb{C}$ with $az := \varepsilon(a)z$
Dual object	Dual vector space with $(a\varphi)(v) := \varphi(\gamma(a)z)$
Tensor product	Reduced product with $a(u\otimes v) := \Delta(a)(u\otimes v)$
Pairing,copairing	$\cap_V(\varphi\otimes v) := \varphi(v), \cup_V(1) := \sum_k v_k\otimes v^k$
Braiding	$\times_{U,V}(u\otimes v) := \sigma(R(u\otimes v))$
Twist	$\theta_V := (\mathrm{id}_V\otimes\cap^*_V)\circ(\times_{V,V}\otimes\mathrm{id}_{V^*})\circ(\mathrm{id}_V\otimes\cup_V)$
Index set	Weyl alcove $I = \Lambda^l_w$
Index involution	$\lambda^* := -w_0(\lambda)$
Simple objects	Alcove Weyl modules $\{\mathcal{W}^\epsilon_\lambda\}_{\lambda\in\Lambda^l_w}$

Table 18.1: Modular category of reduced tilting modules at even roots of unity

As we described in Section 16.6 one can define a dual pairing and copairing in any modular category. In $\overline{Tilt}_\epsilon(\mathfrak{sl}_N\mathbb{C})$ we have, $\cap^*_V(v\otimes\varphi) := \varphi(q^{2\rho}v)$ and $\cup^*_V(1) := \sum_k v^k\otimes q^{-2\rho}v_k$, where ρ is the Weyl weight of $\mathfrak{sl}_N\mathbb{C}$ (see Appendix D).

This is not yet a strict category because equality signs in the axioms are only canonical isomorphisms. For example, V is not equal to $V\otimes\mathbb{C}$. In truth, we should be talking about isomorphism classes of tilting modules rather than tilting modules themselves, or selecting canonical representatives of those classes. One faces the same nuance in the category of representations of classical Lie groups. Formally, some tweaking in the notions of morphisms and ribbon operations is required. It can be done in a standard way by the Mac Lane coherence theorem [100] but in practice one can safely ignore the difference.

To see that the ingredients do indeed form a strict modular category, we will sketch proofs of some of the axioms 16.5. The tensor axioms (Axioms 1–4) essentially follow from Theorem 17.43. The braiding axioms (Axioms 5–7) follow from the defining

properties (45) of the R–matrix. There is a subtlety here since (56) is only a formal expression but (45) holds if R is restricted to a pair of finite-dimensional representations [40].

Exercise 18.1 If $R = \sum a \otimes b$ write $R_{13} = \sum a \otimes 1 \otimes b$ and $R_{23} = \sum 1 \otimes a \otimes b$. Compute the left hand side and the right hand side of Axiom 6. Conclude that Axiom 6 follows from

$$(\Delta \otimes \mathrm{id})(R) = R_{13} R_{23}\,.$$

To verify Axiom 7 write $R = \sum a \otimes b$ and compute

$$\begin{aligned}(g \otimes f) \circ \times_{V,W}(s \otimes t) &= (g \otimes f)\Big(\sum (bt) \otimes (as)\Big) \\ &= \sum (bg(t)) \otimes (af(s)) = \times_{V',W'} \circ (f \otimes g)(s \otimes t).\end{aligned}$$

The twist axioms and the last duality axiom (Axioms 8, 9 and 12) can be verified by manipulating tangle diagrams. One uses the graphical definition of the twist from Figure 16.8. For example Axiom 8 is verified in Figure 18.1. The remaining duality axioms (Axioms 10 and 11) are trivial computations. Axioms 13–16 follow from Andersen's Theorem 17.35 and the construction of $\overline{\mathcal{T}ilt}_\epsilon(\mathfrak{sl}_N\mathbb{C})$. See Andersen [9], Andersen and Paradowski [10], Chari and Pressley [40] and Sawin [135] for more details. The only remaining axiom is 17, the non-degeneracy of the $\widetilde{s}$–matrix. It will be proved as a byproduct of computing the Chern–Simons partition function of S^3 in Section 18.2.

As we already mentioned the link between the Reshetikhin–Turaev and Witten–Chern–Simons invariants is only conjectural. In the absence of a formal definition for the latter we can simply identify them with the former for the category $\overline{\mathcal{T}ilt}_\epsilon(\mathfrak{sl}_N\mathbb{C})$.

Definition 18.2 The Reshetikhin–Turaev invariant corresponding to $U_q(\mathfrak{sl}_N\mathbb{C})$ at level k is defined to be the invariant τ from definition 16.35 arising from the category $\overline{\mathcal{T}ilt}_\epsilon(\mathfrak{sl}_N\mathbb{C})$. It is denoted by $\tau^{\mathfrak{sl}_N\mathbb{C}}(M)$. The Chern–Simons partition function is a different name for the same thing. It is denoted by $Z^{CS}(M) := \tau^{\mathfrak{sl}_N\mathbb{C}}(M)$.

As a warm-up to the computation of $Z^{CS}(S^3)$ we compute the characteristic numbers (Definition 16.23) of $\overline{\mathcal{T}ilt}_\epsilon(\mathfrak{sl}_N\mathbb{C})$ here. We already computed the number d_λ in (43), but we would like to simplify it. Recall how this went. The morphism associated to an unknot labeled with λ sends 1 to d_λ. Let $\{v_i\}$ and $\{v^i\}$ be bases for the representation $\mathcal{V}_\lambda^q$ and its dual respectively. The morphism is a composition of two morphisms and we see that

$$1 \mapsto \sum v_i \otimes v^i \mapsto \sum v^i(\lambda(q^{2\rho})v_i) = \mathrm{Tr}(\lambda(q^{2\rho})) = \chi_\lambda(q^{2\rho}).$$

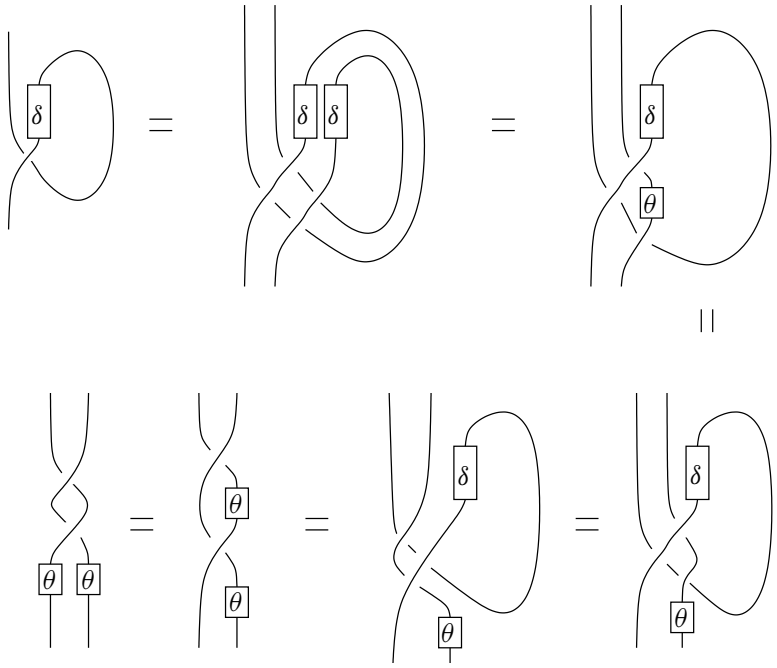


Figure 18.1: Proof of axiom 8

It follows that $d_\lambda = \chi_\lambda(q^{2\rho})$. The Weyl character formula states that the characters are just Schur polynomials, which in turn are ratios of determinants. In this case the determinants are Vandermonde determinants. The appearance of group characters was explained in Remark 17.30.

Since the simple objects are the Weyl modules and the Weyl modules are specializations of $\mathcal{V}_\lambda^q$ we have,

$$d_\lambda = \dim_q \mathcal{W}_\lambda^\epsilon = \prod_{\alpha\in\Delta^+} \frac{\epsilon^{(\lambda+\rho,\alpha)} - \epsilon^{-(\lambda+\rho,\alpha)}}{\epsilon^{(\rho,\alpha)} - \epsilon^{-(\rho,\alpha)}}.$$

Substituting $\epsilon = e^{\frac{\pi i}{k+N}}$ yields

$$\epsilon^{(\mu,\alpha)} - \epsilon^{-(\mu,\alpha)} = 2i\ \frac{e^{\frac{\pi i}{k+N}(\mu,\alpha)} - e^{-\frac{\pi i}{k+N}(\mu,\alpha)}}{2i} = 2i\ \sin\left(\frac{\pi(\mu,\alpha)}{k+N}\right)$$

for any weight μ. For $\mathfrak{sl}_N\mathbb{C}$ the positive roots are $\alpha_{ij} = E_{ii} - E_{jj}$ with $i < j$; therefore, we explicitly get

$$d_\lambda = \prod_{i<j} \frac{\sin\left(\frac{\pi(\lambda+\rho,\alpha_{ij})}{k+N}\right)}{\sin\left(\frac{\pi(\rho,\alpha_{ij})}{k+N}\right)} \tag{60}$$

Exercise 18.3 Note that $\alpha_{ij} = \sum_{k=i}^{j-1} \alpha_k$ and $\rho = \sum_{i=1}^{N-1} \omega_i$. Use the fact that α_i, ω_j are biorthogonal to derive that $(\rho, \alpha_{ij}) = j - i$.

Exercise 18.4 Show that $d_{\omega_1} = (\epsilon^N - \epsilon^{-N})/(\epsilon - \epsilon^{-1})$. This is the invariant of the unknot in the fundamental representation.

In principal to compute other invariants it appears that we have to compute some braiding morphisms. Notice that when the dimension of a representation is n, the R–matrix will be a n^2 by n^2 matrix. Clearly it is not practical to write out many R–matrices. Fortunately, sometimes one can get away with knowledge of its action only on elements of a special form, namely $u_\lambda \otimes v \in U \otimes V$ with u_λ being the highest weight vector in U with the weight $\lambda \in \Lambda_w^+$. This observation is due to V Turaev and H Wenzl [153] and will come handy for the computations of p_λ^+ and $\widetilde{s}_{\lambda\mu}$.

Proposition 18.5 *Let u_λ (v_λ) be the highest weight vector of U (V) and let v (u) be arbitrary. Then*

$$R_{U,V}(u_\lambda \otimes v) = u_\lambda \otimes q^\lambda v, \tag{61}$$

$$R_{U,V}(u \otimes v_\lambda) = q^\lambda u \otimes v_\lambda + z_{<\lambda}, \tag{62}$$

where $z_{<\lambda}$ is an element of $U \otimes V_{<\lambda}$ and $V_{<\lambda}$ is the sum of weight subspaces of V with weights $< \lambda$.

Proof By definition of the highest weight we have $e_{\alpha_i}^{(n)} u_\lambda = 0$ for all i. We also get $e_\alpha^{(n)} u_\lambda = 0$ for all positive roots from (55). Hence the only term that survives in the sum of products in (56) is the one that corresponds to $\nu = 0$, $n_1, \dots, n_\nu = 0$ and the whole sum reduces to $1 \otimes 1$. Therefore, we only have to compute $q^{\sum_{i=1}^{N-1} \alpha_i \otimes \omega_i}(u_\lambda \otimes v)$. Since both sides of (61) are linear in v it suffices to prove it for the case when $v = v_\mu$ is a weight vector with weight μ. We have

$$\begin{aligned} q^{\sum_{i=1}^{N-1} \alpha_i \otimes \omega_i}(u_\lambda \otimes v_\mu) &= q^{\sum_{i=1}^{N-1} (\lambda,\alpha_i)(\mu,\omega_i)} u_\lambda \otimes v_\mu \\ &= q^{\sum_{i,j=1}^{N-1} (\lambda,\alpha_i)(\mu,\omega_j)(\omega_i,\alpha_j)} u_\lambda \otimes v_\mu \\ &= q^{(\sum_{i=1}^{N-1} (\lambda,\alpha_i)\omega_i, \sum_{i=1}^{N-1} (\lambda,\omega_j)\alpha_j)} u_\lambda \otimes v_\mu = q^{(\lambda,\mu)} u_\lambda \otimes v_\mu \end{aligned}$$

since ω_i, α_j are biorthogonal (see Appendix D). But by the definition of the weight $q^{(\lambda,\mu)}v_\mu = q^\lambda v_\mu$ and we are done with the first formula. The proof of the second formula proceeds analogously but v_λ is acted upon by $f_\alpha^{(n)}$ and $f_\alpha^{(n)}v_\lambda$ is no longer zero. Instead, it has a weight lower than λ except in the case when $n = 0$. Thus, up to lower weight vectors in V the sum of products in (56) again reduces to $1 \otimes 1$ and $q^{\sum_{i=1}^{N-1} \alpha_i \otimes \omega_i}$ reduces as above to q^λ acting on u. □

Exercise 18.6 Derive formulas analogous to (61) and (62) for the case when u_λ, v_λ are the lowest weight vectors.

Now we compute $p_\lambda^\pm := \mathrm{Tr}_q(\theta_\lambda^{\pm 1})$. Since the twist commutes with all morphisms (by Axiom 9) and $\mathcal{W}_\lambda^\epsilon$ is irreducible for λ in the Weyl alcove (Theorem 17.35) by the Schur lemma $\theta_\lambda^{\pm 1}$ must be a scalar operator and it suffices to compute it on a single vector in $\mathcal{W}_\lambda^\epsilon$. In view of the simple form of the R–matrix on pairs containing the highest weight vector (Proposition 18.5) we choose the highest weight vector $v_\lambda \in \mathcal{W}_\lambda^\epsilon$:

$$\begin{aligned}
\theta_\lambda v_\lambda :=& (\mathrm{id}_\lambda \otimes \cap_\lambda^*) \circ (\times_{\lambda,\lambda} \otimes \mathrm{id}_{\lambda^*}) \Big(\sum_k v_\lambda \otimes v_k \otimes v^k \Big) \\
=& (\mathrm{id}_\lambda \otimes \cap_\lambda^*) \Big(\sum_k \times_{\lambda,\lambda}(v_\lambda \otimes v_k) \otimes v^k \Big) \\
=& (\mathrm{id}_\lambda \otimes \cap_\lambda^*) \Big(\sum_k (q^\lambda v_k \otimes v_\lambda) \otimes v^k \Big), \quad \text{by (61)} \\
=& \sum_k v^k(q^{2\rho}v_\lambda) q^\lambda v_k = \sum_k \epsilon^{(2\rho,\lambda)} v^k(v_\lambda)\, q^\lambda v_k \\
=& \epsilon^{(2\rho,\lambda)} q^\lambda \Big(\sum_k v^k(v_\lambda) v_k \Big) = \epsilon^{(2\rho,\lambda)} q^\lambda v_\lambda = \epsilon^{(2\rho,\lambda)} \epsilon^{(\lambda,\lambda)} v_\lambda = \epsilon^{(2\rho+\lambda,\lambda)} v_\lambda .
\end{aligned}$$

The case of θ_λ^{-1} is analogous so

$$\theta_\lambda^{\pm 1} = \epsilon^{\pm(2\rho+\lambda,\lambda)}\, \mathrm{id}_\lambda . \tag{63}$$

Taking quantum traces on both sides yields.

$$p_\lambda^\pm = \epsilon^{\pm(2\rho+\lambda,\lambda)}\, \mathrm{Tr}_q(\mathrm{id}_\lambda) = e^{\pm \frac{\pi i}{k+N}(2\rho+\lambda,\lambda)} d_\lambda . \tag{64}$$

Exercise 18.7 We have seen that θ_{ω_1} is just multiplication by a scalar. This scalar is called the framing anomaly. Show that the framing anomaly is $\epsilon^{N-1/N}$.

18.2 $\widetilde{s}$–Matrix and Chern–Simons partition function for S^3

In this subsection we compute the entries $\widetilde{s}_{\lambda\mu}$ of the structure matrix from Axiom 17 and we compute the partition function of S^3. Following Turaev and Wenzl [153] to compute $\widetilde{s}_{\lambda\mu}$ we introduce the meridian morphism $\Gamma_{\lambda\mu}\colon \mathcal{W}^\epsilon_\lambda \to \mathcal{W}^\epsilon_\lambda$, see Figure 18.2. One can see by inspection that $\widetilde{s}_{\lambda\mu} = \mathrm{Tr}_q(\Gamma_{\lambda\mu})$. Again, $\Gamma_{\lambda\mu}$ commutes with

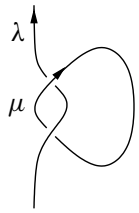


Figure 18.2: The meridian morphism $\Gamma_{\lambda\mu}$

any morphism since we can slide the latter up and down the λ strand, and $\Gamma_{\lambda\mu}$ acts as a scalar since it is an endomorphism of an irreducible representation. It is therefore sufficient to compute it on a highest weight vector v_λ. Let $\{u_i\}$ and $\{u^i\}$ be dual bases for V_μ and V_μ^*. Reading the ribbon expression off Figure 18.2 we have

$$\begin{aligned} v_\lambda &\mapsto \sum_i v_\lambda \otimes u_i \otimes u^i \mapsto \sum_i (q^\lambda u_i) \otimes v_\lambda \otimes u^i \\ &\mapsto \sum_i v_\lambda (q^{2\lambda} u_i + \text{lower weight}) \otimes u^i \mapsto \sum_i v_\lambda u^i (q^{2(\rho+\lambda)} u_i) \\ &= \big(\mathrm{Tr}\,\mu(q^{2(\rho+\lambda)})\big) v_\lambda = \chi_\mu(q^{2(\rho+\lambda)}) v_\lambda. \end{aligned}$$

Here we abuse notation denoting by μ the irreducible representation of $\mathfrak{sl}_N\mathbb{C}$ indexed by the dominant weight μ so that $\mathrm{Tr}\,\mu(\cdot)$ is the ordinary trace in this representation. Taking the quantum trace of this $\Gamma_{\lambda\mu}$ morphism implies that

$$\widetilde{s}_{\lambda\mu} = \mathrm{Tr}_q(\lambda(\Gamma_{\lambda\mu})) = \chi_\lambda(q^{2\rho})\chi_\mu(q^{2(\rho+\lambda)}) = \chi_\mu(q^{2(\rho+\lambda)}) d_\lambda. \tag{65}$$

In particular, $\widetilde{s}_{00} = 1$.

Since S^3 can be obtained from itself by a surgery on the empty link $\varnothing$, we compute

$$Z^{CS}(S^3) := \tau^{\mathfrak{sl}_N\mathbb{C}}(S^3) := (p^-)^{\sigma(\varnothing)} \mathcal{D}^{-\sigma(\varnothing)-c(\varnothing)-1} F(\varnothing) = \mathcal{D}^{-1},$$

where recall $\mathcal{D}^2 := \sum_{\lambda\in I} d_\lambda^2$ and $p_- = \sum_{\lambda\in I} \theta_\lambda^{-1} d_\lambda$. Note that there are many other links that produce S^3 and using expressions for them produces universal algebraic relations among the characteristic numbers of any modular category. Since we already

computed d_λ in (60) it appears that we just have to attach a sum and a square root to the formula. The problem is that we are ultimately interested in the Chern–Simons free energy and this is the logarithm of the partition function. Hence a multiplicative, not an additive, expression is desirable. This can be treated of course as a problem in special functions theory, but we prefer a more conceptual approach based on an understanding of the geometry of the Weyl alcove. The trick is to represent the Weyl alcove as a fundamental domain of a group action; see Kirillov [85], Kac [81], Samelson [134] and Sawin [135]. In addition this will allow us to finally verify the non-degeneracy of the $\widetilde{s}$–matrix.

Definition 18.8 Given an integer $l \in \mathbb{Z}$ define the affine Weyl group $\widetilde{W}^l$ of $\mathfrak{sl}_N\mathbb{C}$ as the group of isometries of its Cartan subalgebra $\mathfrak{h}$ generated by reflections about the hyperplanes $(x, \alpha_i) = kl'$ for every integer $k \in \mathbb{Z}$ and every simple root α_i (as before l' is l for l odd and $l/2$ for l even). We also define the translated action of $\widetilde{W}^l$ as $\widetilde{w} \cdot x := \widetilde{w}(x + \rho) - \rho$, where ρ is the Weyl weight.

Remark 18.9 Notice that we use both actions of elements of the Weyl group. In the Weyl character formula the usual action given by reflections perpendicular to simple roots is used. In the proof of Theorem 18.13 we use the translated action. We will always use the notation $\sigma(\lambda)$ for the usual action and $\sigma \cdot (\lambda)$ for the translated action.

Setting $k = 0$ in the definition we get the reflections s_i that generate the (ordinary) Weyl group $W \subset \widetilde{W}^l$. Another type of elements of $\widetilde{W}^l$ is obtained by performing the reflection about $(x, \alpha_i) = 0$ followed by the reflection about

$$(x, \alpha_i) = l' = l' \frac{(\alpha_i, \alpha_i)}{2} = \Big(\frac{l'\alpha_i}{2}, \alpha_i\Big).$$

This is easily seen to be the translation by $l'\alpha_i$ and we identified the subgroup of translations $l'\Lambda_r \subset \widetilde{W}^l$. Every other reflection can be performed by translating the corresponding hyperplane to the origin, reflecting and then translating back. Thus we arrive at the presentation

$$\widetilde{W}^l = W \ltimes l'\Lambda_r, \tag{66}$$

where in the semi-direct product W acts on $l'\Lambda_r$ by conjugation. In particular, $l'\Lambda_r$ is a normal subgroup of $\widetilde{W}^l$.

Recall that a closed set D is a fundamental domain of a continuous group action if every orbit intersects it and no orbit intersects its interior more than once. The subgroup

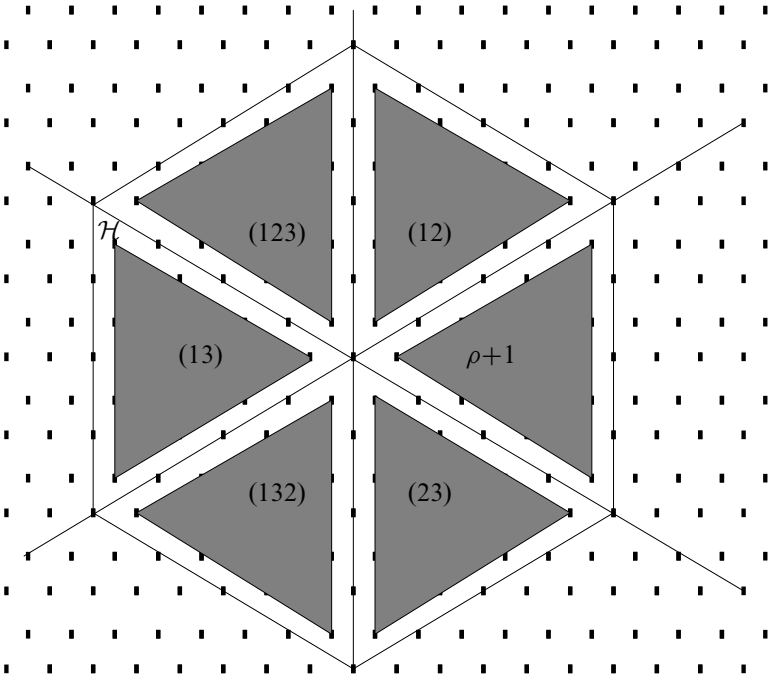


Figure 18.3: Affine Weyl group and $\overline{C}^l$

of translations comprises integer multiples of $l'\alpha$ for $\alpha \in \Delta^+$ and therefore one of its fundamental domains is the polygon

$$\mathcal{H} := \{x \in \mathfrak{h} \mid (x, \pm\alpha) \leq l', \text{ for all } \alpha \in \Delta^+\}.$$

Of course, any translate of $\mathcal{H}$ is also a fundamental domain of $l'\Lambda_r$. Note that by definition the translated action of $l'\Lambda_r$ is the same as the usual one while $\mathcal{H} - \rho$ is invariant under the translated action of W. The fundamental domain for the ordinary action of W is well known to be the Weyl chamber (see Humphreys [78] or Fulton and Harris [62])

$$\Lambda^+ := \{x \in \mathfrak{h} \mid (x, \alpha) \geq 0, \text{ for all } \alpha \in \Delta^+\}$$

and hence $\Lambda^+ - \rho$ serves as a fundamental domain for the translated action.

Exercise 18.10 Let $G = A \ltimes B$ with B being a normal subgroup. Suppose D_A, D_B are fundamental domains of A and B respectively, and D_B is invariant under A. Prove that $D := D_A \cap D_B$ is a fundamental domain of G.

Intersecting $\mathcal{H} - \rho$ with $\Lambda^+ - \rho$ (see Figure 18.3) we establish

Corollary 18.11 *The closed Weyl alcove*

$$\overline{C}^l = \mathcal{H} \cap \Lambda^+ - \rho = \{x \in \mathfrak{h} \mid 0 \le (x+\rho, \alpha) \le l', \text{ for all } \alpha \in \Delta^+\},$$

is a fundamental domain for the translated action of the affine Weyl group $\widetilde{W}^l$ on $\mathfrak{h}$.

After these geometric preliminaries let us move on to a simplification of the quantum diameter $\mathcal{D}^2 := \sum_{\lambda \in I} d_\lambda^2$. First note that

$$\widetilde{s}_{0\lambda} = \widetilde{s}_{\lambda 0} = \mathrm{Tr}_q(\times_{0,\lambda} \circ \times_{\lambda,0}) = \mathrm{Tr}_q(\mathrm{id}_\lambda) = d_\lambda.$$

This suggests considering the square of the $\widetilde{s}$–matrix, indeed

$$(\widetilde{s}^2)_{00} = \sum_{\nu \in \Lambda_w^l} \widetilde{s}_{0\nu}\widetilde{s}_{\nu 0} = \sum_{\nu \in \Lambda_w^l} d_\nu^2 = \mathcal{D}^2. \tag{67}$$

Non-degeneracy of the $\widetilde{s}$–matrix was originally established by Turaev and Wenzl in a roundabout way based on results of Kac and Petersen from the theory of affine Lie algebras [153]. The direct proof based on an explicit computation of $\widetilde{s}^2$ that we present here is due to A Kirillov [21; 85]. As a bonus, it gives a nice formula for the quantum diameter. This proof requires the orthogonality relation for characters of finite groups sketched in the next exercise. Another, purely algebraic proof is given by M Müger [111].

Exercise 18.12 Define $\langle \phi, \psi \rangle := |G|^{-1} \sum_{g \in G} \phi(g)\psi(g)^*$ for functions defined on an arbitrary finite group. Prove that $\langle \chi_\lambda, \chi_\mu \rangle = \delta_{\lambda\mu}$. This is similar to Exercise D.36 from Appendix D.

Theorem 18.13 *Let l be an even integer. Then $(\widetilde{s}^2)_{\lambda\mu} = \mathcal{D}^2 \delta_{\lambda\mu^*}$ and*

$$\mathcal{D}^2 = (-1)^{w_0} \frac{|\Lambda_w / l'\Lambda_r|}{\delta_0(q^{2\rho})^2}, \tag{68}$$

where w_0 is the order-reversing permutation and δ_0 is the Weyl denominator (92)*. In particular, $\mathcal{D} \neq 0$ and the $\widetilde{s}$–matrix is non-degenerate.*

Proof We first transform expression (65) for $\widetilde{s}_{\lambda\nu}$ using the Weyl character formula (93)

$$\begin{aligned} \widetilde{s}_{\lambda\nu} &= \chi_\nu(q^{2(\rho+\lambda)}) d_\lambda = \chi_\nu(q^{2(\rho+\lambda)}) \chi_\lambda(q^{2\rho}) \\ &= \frac{\delta_\nu(q^{2(\rho+\lambda)})}{\delta_0(q^{2(\rho+\lambda)})} \frac{\delta_\lambda(q^{2\rho})}{\delta_0(q^{2\rho})} = \frac{\delta_\nu(q^{2(\rho+\lambda)})\, \delta_\lambda(q^{2\rho})}{\delta_\lambda(q^{2\rho})\, \delta_0(q^{2\rho})}, \quad \text{by (95)} \\ &= \frac{1}{\delta_0(q^{2\rho})} \sum_{\sigma \in W = \mathfrak{S}_N} (-1)^\sigma \epsilon^{2(\sigma(\lambda+\rho),\, \nu+\rho)}. \end{aligned} \tag{69}$$

Notice that while $\widetilde{s}_{\lambda\nu}$ is originally only defined for weights λ, ν in the Weyl alcove Λ^l_w, the right side of equation (69) makes sense for all weights. We extend the definition of $\widetilde{s}_{\lambda\nu}$ to all weights by equation (69). We can now replace the summation over Λ^l_w in (67) by summation over a more convenient set, the quotient group $\Lambda_w/l'\Lambda_r$. This is done in two steps. First, we know from Theorem 17.35 that the Weyl modules with the highest weights in $\overline{C}^l\backslash C^l$ have quantum dimensions 0 and Exercise 17.44 implies $\widetilde{s}_{\lambda\nu} = \mathrm{Tr}_q(\times_{\lambda,\nu}\circ\times_{\nu,\lambda}) = 0$ for $\nu\in\overline{C}^l\backslash C^l$ so

$$(\widetilde{s}^2)_{\lambda\mu} = \sum_{\nu\in\Lambda^l_w}\widetilde{s}_{\lambda\nu}\widetilde{s}_{\nu\mu} = \sum_{\nu\in\overline{\Lambda}^l_w}\widetilde{s}_{\lambda\nu}\widetilde{s}_{\nu\mu},$$

where $\overline{\Lambda}^l_w := \Lambda_w\cap\overline{C}^l$. For the second step notice that the value of $\widetilde{s}_{\lambda\nu}$ is only changed by the sign of the orientation of the transformation when λ or ν are acted on by elements of the affine Weyl group via the translated action. Indeed applying a translation $l'\alpha_i$ from the affine Weyl group to ν does not change the value of $\widetilde{s}_{\lambda\nu}$ since

$$\epsilon^{2(\sigma(\lambda+\rho),\,\nu+l'\alpha_i+\rho)} = \epsilon^{2(\sigma(\lambda+\rho),\,\nu+\rho)}\epsilon^{2(\sigma(\lambda+\rho),\,l'\alpha_i)} = \epsilon^{2(\sigma(\lambda+\rho),\,\nu+\rho)},$$

and α_i pairs with any weight to give an integer and $\epsilon^{2l'}=1$. Applying the translated action of an element of the Weyl group to λ gives

$$\begin{aligned}\widetilde{s}_{\tau\cdot\lambda\nu} &= \frac{1}{\delta_0(q^{2\rho})}\sum_{\sigma\in W=\mathfrak{S}_N}(-1)^\sigma\,\epsilon^{2(\sigma(\tau\cdot\lambda+\rho),\,\nu+\rho)}\\ &= \frac{1}{\delta_0(q^{2\rho})}\sum_{\sigma\in W=\mathfrak{S}_N}(-1)^\sigma\,\epsilon^{2(\sigma(\tau(\lambda+\rho)),\,\nu+\rho)}\\ &= \frac{(-1)^\tau}{\delta_0(q^{2\rho})}\sum_{\sigma\in W=\mathfrak{S}_N}(-1)^\sigma\,\epsilon^{2(\sigma(\lambda+\rho),\nu+\rho)}.\end{aligned}\tag{70}$$

By Corollary 18.11 and (70) summation over $\overline{\Lambda}^l_w$ can be replaced by summation over equivalence classes in $\Lambda_w/\widetilde{W}^l$ and furthermore in view of (66)

$$\sum_{\nu\in\overline{\Lambda}^l_w}\widetilde{s}_{\lambda\nu}\widetilde{s}_{\nu\mu} = \sum_{\nu\in\Lambda_w/\widetilde{W}^l}\widetilde{s}_{\lambda\nu}\widetilde{s}_{\nu\mu} = \frac{1}{|W|}\sum_{\nu\in\Lambda_w/l'\Lambda_r}\widetilde{s}_{\lambda\nu}\widetilde{s}_{\nu\mu}.\tag{71}$$

This gives

$$(\widetilde{s}^2)_{\lambda\mu} = \frac{1}{|W|\,\delta_0(q^{2\rho})^2}\sum_{\nu\in\Lambda_w/l'\Lambda_r}\sum_{\sigma_1,\sigma_2\in W}(-1)^{\sigma_1\sigma_2}\,\epsilon^{2(\sigma_1(\lambda+\rho)+\sigma_2(\mu+\rho),\,\nu+\rho)}.\tag{72}$$

To simplify the last double sum we change the order of summation and look closer at the maps $\epsilon^{2(\kappa,\cdot)}$, where $\kappa\in\Lambda_w$. Since $\epsilon^{2l'}=\epsilon^l=1$ they are well-defined as maps

$\Lambda_w/l'\Lambda_r \to \mathbb{C}^*$ and in fact are characters of the Abelian group $\Lambda_w/l'\Lambda_r$. Since ϵ is a primitive lth root of unity $\epsilon^{2(\kappa,\cdot)}$ defines the trivial character if and only if (κ,ν) is divisible by $l/2$ for any $\nu\in\Lambda_w$. Since l is even and $l'=l/2$ this is equivalent to $\kappa\in l'\Lambda_r$. The orthogonality relation for characters now yields

$$(\epsilon^{2(\kappa,\cdot)},1)=\sum_{\nu\in\Lambda_w/l'\Lambda_r}\epsilon^{2(\kappa,\nu)}=\sum_{\nu\in\Lambda_w/l'\Lambda_r}\epsilon^{2(\kappa,\nu+\rho)}=\begin{cases}0, & \kappa\notin l'\Lambda_r\\ |\Lambda_w/l'\Lambda_r|, & \kappa\in l'\Lambda_r.\end{cases}$$

This implies that the terms in the double sum (72) are 0 unless $\sigma_1(\lambda+\rho)+\sigma_2(\mu+\rho)\in l'\Lambda_r$. But by Theorem 17.35: $\mu+\rho=-w_0(\mu^*+\rho^*)=-w_0(\mu^*+\rho)$ since $\rho^*=\rho$. Therefore, the following are equivalent.

$$\begin{aligned}&\sigma_1(\lambda+\rho)+\sigma_2(\mu+\rho)\in l'\Lambda_r\\ &\sigma_1(\lambda+\rho)-\sigma_2 w_0(\mu^*+\rho)\in l'\Lambda_r\\ &\lambda+\rho\in\sigma_1^{-1}\sigma_2 w_0(\mu^*+\rho)+l'\Lambda_r=:\sigma(\mu^*+\rho)+l'\Lambda_r\\ &\lambda\in\sigma(\mu^*+\rho)-\rho+l'\Lambda_r=\sigma\cdot\mu^*+l'\Lambda_r\\ &\lambda=\widetilde{w}\cdot\mu^*,\text{ for }\widetilde{w}\in\widetilde{W}^l.\end{aligned}$$

But $\lambda,\mu^*\in\Lambda_w^l\subset C^l$ are within a fundamental domain of the translated action and the last condition can only be satisfied when $\lambda=\mu^*$ and $\widetilde{w}=1$ or equivalently $\sigma_1=\sigma_2 w_0$. In particular, $(\widetilde{s}^2)_{\lambda\mu}=0$ for $\lambda\neq\mu^*$. When $\lambda=\mu^*$ expression (72) reduces to

$$\begin{aligned}(\widetilde{s}^2)_{\lambda\mu}&=\frac{1}{|W|\,\delta_0(q^{2\rho})^2}\sum_{w_1\in W}(-1)^{2w_1+w_0}\,|\Lambda_w/l'\Lambda_r|\\ &=\frac{1}{|W|\,\delta_0(q^{2\rho})^2}(-1)^{w_0}|W||\Lambda_w/l'\Lambda_r|=(-1)^{w_0}\frac{|\Lambda_w/l'\Lambda_r|}{\delta_0(q^{2\rho})^2}.\end{aligned}$$

The formula for the quantum diameter now follows directly from (67). □

Using the formula for the Weyl denominator, we obtain the following formula for the quantum diameter that generalizes to any semisimple Lie algebra.

$$\mathcal{D}=i^{w_0}|\Lambda_w/l'\Lambda_r|^{1/2}\prod_{\alpha\in\Delta^+}(\epsilon^{(\rho,\alpha)}-\epsilon^{-(\rho,\alpha)})^{-1}.$$

We can make this formula more explicit for $\mathfrak{sl}_N\mathbb{C}$. Recall that the level is $k := l' - N$. By (94) and Exercise 18.3

$$\delta_0(q^{2\rho}) = \prod_{\alpha\in\Delta^+} (\epsilon^{(\rho,\alpha)} - \epsilon^{-(\rho,\alpha)}) = \prod_{i<j} 2i\,\frac{e^{\frac{\pi i(\rho,\alpha_{ij})}{k+N}} - e^{-\frac{\pi i(\rho,\alpha_{ij})}{k+N}}}{2i}$$

$$= i^{\frac{N(N-1)}{2}} \prod_{i<j} 2\,\sin\left(\frac{\pi(j-i)}{k+N}\right) = i^{\frac{N(N-1)}{2}} \prod_{j=1}^{N-1} 2\,\sin\left(\frac{\pi j}{k+N}\right)^{N-j}.$$

Exercise 18.14 Prove that $|\Lambda_w/(k+N)\Lambda_r| = N(k+N)^{N-1}$, and show that $(-1)^{w_0} = (-1)^{N(N-1)/2}$.

We arrive at the following.

Corollary 18.15 *The Chern–Simons partition function for the three-sphere is*

$$Z^{CS}(S^3) = \mathcal{D}^{-1} = N^{-1/2}(k+N)^{(1-N)/2} \prod_{j=1}^{N-1} \left(2\sin\left(\frac{\pi j}{k+N}\right)\right)^{N-j}. \tag{73}$$

Remark 18.16 When l is odd the level $k = l/2 - N$ is only a half-integer and Witten's heuristic argument for the invariance of Z^{CS} breaks down. Thus, there is no 'physical' reason to expect that $U_\epsilon^{\mathrm{res}}(\mathfrak{sl}_N\mathbb{C})$ produces topological invariants for ϵ a primitive odd root of unity. The above proof of non-degeneracy of the $\widetilde{s}$–matrix also fails for l odd. Indeed, it is sufficient that $\kappa \in \frac{l}{2}\Lambda_r$ as opposed to $\kappa \in l\Lambda_r = l'\Lambda_r$ for $\epsilon^{2(\kappa,\cdot)}$ to define the trivial character and we can not reduce the double sum completely. As a matter of fact, the $\widetilde{s}$–matrix *can be degenerate* in this case. This is due to appearance of nontrivial transparent [35] (also called degenerate [135] or central [111]) simple objects in $\overline{Tilt}_\epsilon(\mathfrak{sl}_N\mathbb{C})$. These are the Weyl modules $\mathcal{W}_\tau^\epsilon$ such that $R_{\lambda,\tau} = R_{\tau,\lambda}^{-1}$ for all alcove weights λ. Obviously, the trivial object is transparent in any ribbon category. The name comes from the fact that $\times_{\lambda,\tau} = \times_{\tau,\lambda}$ and any strand can be pushed through τ–colored ones without changing the invariant. As a result, transparent objects create a row $\widetilde{s}_{\tau\lambda}$ in the $\widetilde{s}$–matrix that is proportional to the row of the trivial object $\widetilde{s}_{0\lambda}$ making the matrix degenerate. When l is odd and N is even the fundamental weight $\omega_{\frac{N}{2}}$ indexes a nontrivial transparent object in $\overline{Tilt}_\epsilon(\mathfrak{sl}_N\mathbb{C})$. In particular, for $\mathfrak{sl}_2\mathbb{C}$ reduced tilting modules at odd roots do not form a modular category (see Sawin [135]). Moreover, under mild technical assumptions one can prove that a category satisfying Axioms 1–16 of 16.5 (such categories are called premodular) with $\mathcal{D} \neq 0$ is modular if and only if it does not have nontrivial transparent simple objects (see Bruguiéres [35] and Müger

[111]). Surprisingly, $\overline{Tilt_\epsilon}(\mathfrak{sl}_N\mathbb{C})$ can be 'modularized' by a quotient construction and made to produce nontrivial invariants even when l is odd except when $N = 2 \mod 4$ [135]. At present, a geometric or physical explanation for this phenomenon is lacking.

We conclude with some remarks about computing the Reshetikhin–Turaev invariants for more general 3–manifolds. There are two different ways that may be used to organize many such computations. The first method is via the shadow invariants of V Turaev [152]. The second method is to expand the theory into a full topological quantum field theory. Most computations in the physical literature use the topological quantum field theory viewpoint.

Recall the twist matrix $\widetilde{t}_{\lambda\mu} := \epsilon^{(2\rho+\lambda,\lambda)}\delta_{\lambda\mu}$ from (63). For many 3–manifolds such as circle bundles over Riemann surfaces one can avoid using the R–matrices directly by utilizing the fact that renormalized $\widetilde{s}$, $\widetilde{t}$ matrices form a projective representation of the modular group $\mathrm{SL}_2\mathbb{Z}$ that appears in many surgeries. Namely, set $s := \mathcal{D}^{-1}\widetilde{s}$ and $t := \zeta^{-1}\widetilde{t}$ with $\zeta := (p^+/p^-)^{\frac{1}{6}}$ then s, t can be taken as the images of the standard generators S, T of $\mathrm{SL}_2\mathbb{Z}$

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \qquad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Exercise 18.17 Show that s, t satisfy the standard relations of $\mathrm{SL}_2\mathbb{Z}$: $(st)^3 = s^2$, $s^2t = ts^2$, $s^4 = 1$, see Bakalov and Kirillov [21; 85].

In terms of the s–matrix we have

$$d_\lambda = \mathcal{D}s_{0\lambda} = \frac{s_{0\lambda}}{s_{00}}$$
$$Z^{CS}(S^3) = \mathcal{D}^{-1} = s_{00}$$

These identities are frequently used in practical computations.

¿We combine equations (69) and (68) in this subsection into the following corollary.

Corollary 18.18 *In any semisimple Lie algebra we have the following explicit formula for the s–matrix and $\widetilde{t}$–matrix.*

$$s_{\lambda\nu} = i^{-|\Delta_+|}|\Lambda_w/l'\Lambda_r^\vee|^{-1/2}\sum_{\sigma\in W}(-1)^\sigma\epsilon^{2\langle\!\langle\sigma(\lambda+\rho),\nu+\rho\rangle\!\rangle}.$$
$$\widetilde{t}_{\lambda\mu} := \epsilon^{\langle\!\langle 2\rho+\lambda,\lambda\rangle\!\rangle}\delta_{\lambda\mu}.$$

Here $\Lambda_r^\vee$ is the coroot lattice and $\langle\!\langle\cdot,\cdot\rangle\!\rangle$ is the Killing form normalized so that $\langle\!\langle\alpha,\alpha\rangle\!\rangle = 2$ for short roots.

Specializing to $\mathfrak{sl}_N\mathbb{C}$ gives

$$s_{\lambda\nu} = N^{-1/2}(k+N)^{(1-N)/2}(i)^{-N(N-1)/2}\sum_{\sigma\in\mathfrak{S}_N}(-1)^{\sigma}\epsilon^{2\langle\rho+\nu,\sigma(\rho+\lambda)\rangle}.$$

Remark 18.19 The sign in the exponent $(i)^{-N(N-1)/2}$ differs from that of [21] since we are using the right-handed Hopf link to define the s–matrix (not the left-handed one.) The sign used here and the use of the right-handed Hopf link are correct.

In the next section we compute the free energy of the 3–sphere and perform the final comparison between the Gromov–Witten free energy and Chern–Simons free energies.

Part III Comparisons and recent developments

19 Comparison of the free energies

In this section we reach the final goal of this paper by comparing the Gromov–Witten and Chern–Simons free energies for X_{S^3} and S^3. We start by recalling some necessary facts and formulas from the theory of special functions that are used in transforming the expressions derived earlier in the paper. We apply these formulas to reduce both energies to a similar form and perform the comparison. Even though the energies do not match exactly the difference does not contain any positive powers of $x := 2\pi/(k+N)$ (called the string coupling constant and denoted g_s in the physical literature) which means that we have an exact match for the counting invariants.

19.1 Bernoulli numbers and special functions

After generating functions for the invariants are computed the claim of the Gopakumar–Vafa Large N Duality reduces to a problem in the theory of special functions. Aside from Bernoulli numbers we need the Eisenstein functions, the Riemann zeta-function, polylogarithms and the Barnes function. Rather than simply referring the reader to various sources for necessary formulas we briefly review the definitions and relationship between them in this subsection.

Since free energies are essentially the natural logarithms of partition functions it helps to present all terms in partition functions as products. In particular, the Chern–Simons partition function contains sines and we start by deriving the Euler product formula for the sine function. There are different ways to do this but we choose the one using Eisenstein functions since we need them later, see Weil [157].

Definition 19.1 The Eisenstein functions are defined by

$$E_k(z) := \text{v.p.} \sum_{n=-\infty}^{\infty} (z+n)^{-k},$$

where we use the Eisenstein convention:

$$\text{v.p.} \sum_{n=-\infty}^{\infty} f(n) := \lim_{N\to\infty} \sum_{n=-N}^{N} f(n)$$

The main properties of the Eisenstein functions are collected in the following exercises:

Exercise 19.2 Show that $E_1(z)-\pi\cot(\pi z)$ is a bounded entire function that evaluates to zero at $z=\frac{1}{2}$ and use Liouville's theorem to conclude that it is identically zero.

Exercise 19.3 From the definition, it follows that $E_k'(z)=-kE_{k+1}(z)$, so all of these functions may be found by differentiating the cotangent function. Using this show that $E_2(z)=\pi^2\csc^2(\pi z)$ and $E_3(z)=E_1(z)E_2(z)$.

Exercise 19.4 Set $s(z)=z\prod_{n=1}^{\infty}\left(1-\frac{z^2}{n^2}\right)$ and prove that $s'(z)/s(z)=E_1(z)$.

Exercise 19.5 Express $\frac{d}{dz}(s(z)^2E_2(z))$ using just $s(z)$, $E_1(z)$ and $E_2(z)$. Conclude that it is zero.

It is not hard to see that $s(0)=0$ and $s'(0)=1$. It follows that

$$\lim_{z\to 0}\frac{\pi s(z)}{\sin\pi z}=1\,.$$

However we know from Exercise 19.5 that $\left(\frac{\pi s(z)}{\sin\pi z}\right)^2$ is constant. Thus

$$\sin(\pi z)=\pi z\prod_{n=1}^{\infty}\left(1-\frac{z^2}{n^2}\right).$$

To transform multiple sums that appear in the Chern–Simons free energy we need closed formulas for power sums of natural numbers. These formulas involve the Bernoulli numbers.

Definition 19.6 The Bernoulli numbers B_k are defined by their generating function:

$$\frac{z}{e^z-1}=\sum_{k=0}^{\infty}B_k\frac{z^k}{k!}\,.$$

Again we collect the main properties we need in an exercise, see Andrews–Askey–Roy [11].

Exercise 19.7 Show that $\frac{z}{2}\coth\left(\frac{z}{2}\right)=\frac{z}{2}+\frac{z}{e^z-1}$ and conclude that $B_1=-\frac{1}{2}$ and the rest of the odd Bernoulli numbers are zero. Also compute B_0, B_2, B_4 and B_6.

Exercise 19.8 Expand $\frac{e^{Nz}-1}{z}\frac{z}{e^z-1}$ as a sum of exponentials, and then expand the exponentials into power series to obtain $N+\sum_{n=1}^{\infty}\sum_{j=1}^{N-1}j^p\frac{z^n}{n!}$.

Exercise 19.9 Expand each factor of $\frac{e^{Nz}-1}{z}\frac{z}{e^z-1}$ in a power series and multiply the two resulting power series to obtain

$$\sum_{n=0}^{\infty}\sum_{j=0}^{n}\binom{n}{j}B_j\frac{N^{n-k+1}}{n-k+1}\frac{z^n}{n!}\,.$$

Exercise 19.10 Compare the expansions from the two previous exercises to obtain a formula for the sum of powers of the first $N-1$ natural numbers. Use the fact that all of the odd Bernoulli numbers other than B_1 vanish together with the definition of the binomial coefficients to conclude

$$\sum_{j=1}^{N-1} j^{2p+1} = -\frac{N^{2p+1}}{2} + \sum_{k=0}^{p}\binom{2p+1}{2k}\frac{B_{2k}N^{2p-2k+2}}{2p-2k+2}\,,$$
$$\sum_{j=1}^{N-1} j^{2p} = -\frac{N^{2p}}{2} + \sum_{k=0}^{p}\binom{2p+1}{2k}\frac{B_{2k}N^{2p-2k+1}}{2p+1}\,.$$

Bernoulli numbers are closely related to values of the celebrated Riemann zeta function at even integers.

Definition 19.11 The Riemann zeta function is defined by (Andrews–Askey–Roy [11]):

$$\zeta(z) := \frac{1}{\Gamma(z)}\int_0^{\infty}\frac{u^{z-1}}{e^u-1}\,du = \sum_{n=1}^{\infty}n^{-z}\,,$$

where $\Gamma(z)$ is the usual gamma function of Euler

$$\Gamma(z) := \int_0^{\infty}e^{-t}t^{z-1}\,dt\,.$$

Exercise 19.12 Use integration by parts to derive the usual relation between the gamma function and the factorial. Expand $(e^u-1)^{-1}$ in powers of e^{-u} and substitute into the definition of the zeta function to obtain the formula $\zeta(z)=\sum_{n=1}^{\infty}n^{-z}$. Use the generating function definition of the Bernoulli numbers in the definition of the zeta function to obtain for $n\geq 1$,

$$\zeta(2n) = (-1)^{n+1}\frac{(2\pi)^{2n}B_{2n}}{2(2n)!}\,. \tag{74}$$

The gamma function can be meromorphically extended to the entire complex plane via $\Gamma(z+1)=z\Gamma(z)$ and the zeta function can be meromorphically extended to the

entire complex plane via the functional equation discovered for the zeta function by Riemann himself (see Hardy and Wright [72]):

$$\zeta(1-k) = 2(2\pi)^{-k}\cos(\pi k/2)\Gamma(k)\zeta(k)\,.$$

The Chern–Simons partition function contains a factor that can be identified with the volume of $\mathrm{SU}(N)$. This volume can be expressed in terms of the Barnes function.

Definition 19.13 The Barnes function is defined by

$$G_2(z+1) = (2\pi)^{z/2}e^{-(z(z+1)+\gamma z^2)/2}\prod_{k=1}^{\infty}\left((1+z/k)^k e^{(z^2-z)/2k}\right),$$

where γ is the Euler constant given by $\gamma := \lim_{n\to 0}\left(\sum_{k=1}^{n} k^{-1} - \ln n\right)$.

Remark 19.14 For integers greater than 1 one can show that $G_2(N) := \prod_{j=1}^{N-2} j!$.

The following asymptotic expansion for G_2 in terms of Bernoulli numbers is given by Adamchik [2]:

$$\ln(G_2(N+1)) = \tfrac{1}{2}N^2\ln N - \tfrac{3}{4}N^2 - \tfrac{1}{12}\ln N - N\zeta'(0) + \zeta'(-1) + \sum_{g=2}^{\infty}\frac{B_{2g}}{2g(2g-2)}N^{2-2g} \tag{75}$$

Remark 19.15 The formula in [2] has a negative sign in front of the sum. Careful inspection shows that a sign was lost when the expression following line (29) in this paper was substituted into equation (20) from this paper.

Finally, manipulations with the Gromov–Witten free energy require the use of polylogarithms.

Definition 19.16 The polylogarithm functions are defined by (see eg [42])

$$\mathrm{Li}_p(z) := \sum_{n=1}^{\infty} n^{-p}z^n\,. \tag{76}$$

One can see by inspection that $\mathrm{Li}_1(z) = -\ln(1-z)$ and $z\frac{d}{dz}\mathrm{Li}_{p+1}(z) = \mathrm{Li}_p(z)$. Therefore, polylogarithms indexed by positive integers have a logarithmic branch point at $z=1$. On the other hand, polylogarithms with negative integers are meromorphic in the complex plane and relate to the Eisenstein functions by a change of variables and renormalization.

Exercise 19.17 Expand $\pi\cot(\pi z)$ as a power series in the exponential $e^{-2\pi i z}$ (see [11]) and differentiate the expression obtained by combining this with Exercise 19.2 and the definition of the polylogarithm to obtain for $q \geq 1$

$$\mathrm{Li}_{-q}(e^{-2\pi i z}) = \frac{q!}{(2\pi i)^{q+1}} \sum_{n=-\infty}^{\infty} (n+z)^{-q-1} = \frac{q!}{(2\pi i)^{q+1}} E_{q+1}(z)\,.$$

We will need power series expansions for $\mathrm{Li}_{3-2g}(e^{-t})$ at 0. Note also that these functions are manifestly periodic with the period $2\pi i$. For $g \geq 2$ combining the last exercise with the negative power binomial theorem gives

$$\begin{aligned}
\mathrm{Li}_{3-2g}(e^{-t}) &= (2g-3)!t^{2-2g} + \frac{(2g-3)!}{(2\pi i)^{2g-2}} \sum_{\substack{n=-\infty \\ n\neq 0}}^{\infty} (1+(t/2\pi i n))^{2-2g} n^{2-2g} \\
&= (2g-3)!t^{2-2g} \\
&\quad + (2g-3)! \sum_{\substack{n=-\infty \\ n\neq 0}}^{\infty} \sum_{h=0}^{\infty} \binom{2g+h-3}{h} (-t)^h (2\pi i)^{2-2g-h} n^{2-2g-h} \qquad (77) \\
&= (2g-3)!t^{2-2g} \\
&\quad + (2g-3)! \sum_{\substack{h \text{ even} \\ h\geq 0}} 2\binom{2g+h-3}{h} (-t)^h (2\pi i)^{2-2g-h} \zeta(2g+h-2)\,.
\end{aligned}$$

As mentioned above the functions $\mathrm{Li}_1(e^{-t})$ and $\mathrm{Li}_3(e^{-t})$ have a branch point at $t=0$ and can not be expanded into a Taylor series. However, this is easy to fix by adding logarithmic 'counterterms' that render them holomorphic in a neighborhood of 0. For instance, $\mathrm{Li}_1(e^{-t}) = -\ln(1-e^{-t})$ behaves like $-\ln t$ near 0 so the sum $\mathrm{Li}_1(e^{-t})+\ln t$ is holomorphic. Of course, by writing $\ln t$ we are implicitly fixing a branch of the logarithm and thus the polylogarithm as well. The corresponding expansions are derived in the next lemma.

Lemma 19.18 *The polylogarithms admit the following series expansions:*

$$Li_1(e^{-t}) + \ln t = t/2 + \sum_{\substack{m \text{ even} \\ m\geq 2}} \frac{2}{m} (2\pi)^{-m} \zeta(m) (it)^m, \qquad (78)$$

$$\begin{aligned}
Li_3(e^{-t}) + \frac{t^2}{2}\ln t &= \zeta(3) - \zeta(2)t + 3t^2/4 + t^3/12 \\
&\quad - \sum_{\substack{m \text{ even} \\ m\geq 4}} \frac{2}{m(m-1)(m-2)} (2\pi)^{2-m} \zeta(m-2) (it)^m. \qquad (79)
\end{aligned}$$

Proof The function $f(t) = \ln(t) + \mathrm{Li}_1(e^{-t}) = \ln\left(t(1-e^{-t})^{-1}\right)$ is clearly holomorphic at zero. Now compute

$$\begin{aligned} f'(t) &= t^{-1}\left(1 - \frac{t}{e^t - 1}\right) \\ &= \frac{1}{2} - \sum_{n=1}^{\infty} \frac{B_{2n}}{(2n)!} t^{2n-1}. \end{aligned}$$

Since $f(0) = 0$ we can integrate and use formula (74) relating the Bernoulli numbers to the zeta function see that

$$\begin{aligned} f(t) &= \tfrac{1}{2}t - \sum_{n=1}^{\infty} \frac{B_{2n}}{(2n)!(2n)} t^{2n} \\ &= \tfrac{1}{2}t + \sum_{n=1}^{\infty} \frac{(2\pi)^{-2n}}{n} \zeta(2n)(it)^{2n}. \end{aligned}$$

For the second equality set $g(t) = \mathrm{Li}_3(e^{-t}) + \frac{1}{2}t^2 \ln t - \frac{3}{4}t^2$. One easily computes $g(0) = \zeta(3)$,

$$g'(t) = -\mathrm{Li}_2(e^{-t}) + t \ln t - t,$$

giving $g'(0) = -\zeta(2)$ and

$$g''(t) = \mathrm{Li}_1(e^{-t}) + \ln t = \tfrac{1}{2}t + \sum_{\substack{m \text{ even} \\ m \geq 2}} \frac{2}{m} (2\pi)^{-m} \zeta(m)(it)^m.$$

Integrating twice as above with the computed constants of integration, reindexing the sum with $h = m + 2$ and using the definition of $g(t)$ gives

$$\begin{aligned} \mathrm{Li}_3(e^{-t}) + \tfrac{1}{2}t^2 \ln t = \zeta(3) - \zeta(2)t + \tfrac{3}{4}t^2 + \tfrac{1}{12}t^3 \\ - \sum_{\substack{h \text{ even} \\ h \geq 4}} \frac{2}{h(h-1)(h-2)} (2\pi)^{2-h} \zeta(h-2)(it)^h. \end{aligned}$$

This completes the proof. □

19.2 The Chern–Simons free energy

Recall from equation (73) that the $SU(N)$–Witten–Reshetikhin–Turaev invariant of S^3 is given by

$$Z(S^3) = \tau_k^{\mathfrak{sl}_N\mathbb{C}}(S^3) = N^{-1/2}(k+N)^{(1-N)/2} \prod_{j=1}^{N-1} \left[2\sin\left(\frac{\pi j}{k+N}\right)\right]^{N-j}.$$

Recall that the Chern–Simons free energy is given by

$$F_M = \ln(Z^{\mathrm{U}(N)}(M)/Z_0^{\mathrm{U}(N)}(M)).$$

The formal mathematical definition of the $U(N)$ partition function has not yet been agreed upon. In addition there is no mathematical definition of $Z_0^{\mathrm{U}(N)}(M)$. To check Large N Duality we define the unnormalized Chern–Simons free energy to be the logarithm of the $\mathrm{SU}(N)$–Witten–Reshetikhin–Turaev invariant.

Definition 19.19 The unnormalized Chern–Simons free energy is

$$F_M^{\mathrm{CS}} = \ln(\tau_k^{\mathfrak{sl}_N\mathbb{C}}(M)).$$

It follows immediately that

$$\begin{aligned} F_{S^3}^{\mathrm{CS}} &= \frac{1-N}{2}\ln(k+N) - \tfrac{1}{2}\ln N + \sum_{j=1}^{N-1}(N-j)\ln\Big[2\sin\Big(\frac{\pi j}{k+N}\Big)\Big] \\ &= \frac{1-N}{2}\ln(k+N) - \tfrac{1}{2}\ln N + \sum_{j=1}^{N-1}(N-j)\ln\Big(\frac{2\pi j}{k+N}\Big) \\ &\quad + \sum_{j=1}^{N-1}(N-j)\Big(\sum_{n=1}^{\infty}\ln\Big(1-\frac{j^2}{n^2(k+N)^2}\Big)\Big). \end{aligned} \tag{80}$$

Here we have used the product expansion of the sine function derived just below Exercises 19.2–19.5. We will now concentrate on the last sum in this expression. We call it the perturbative Chern–Simons free energy, and denote it by F^{pert}. Using the coupling constant $x = 2\pi/(k+N)$, replace $k+N$ in F^{pert}, expand the logarithm in a series ($\ln(1-z) = -z - \frac{1}{2}z^2 - \frac{1}{3}z^3 - \cdots$) and use $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$ to obtain

$$\begin{aligned} F^{\mathrm{pert}} &= \sum_{j=1}^{N-1}(N-j)\Big(\sum_{n=1}^{\infty}\ln\Big(1-\frac{x^2j^2}{4\pi^2n^2}\Big)\Big) \\ &= -\sum_{j=1}^{N-1}(N-j)\Big(\sum_{n=1}^{\infty}\sum_{p=1}^{\infty}\frac{x^{2p}j^{2p}}{4^p\pi^{2p}pn^{2p}}\Big) = -\sum_{j=1}^{N-1}(N-j)\Big(\sum_{p=1}^{\infty}\frac{x^{2p}j^{2p}}{4^p\pi^{2p}p}\zeta(2p)\Big). \end{aligned}$$

Recall from Section 15 on perturbative Chern–Simons theory that we expect based on intuition from the path integral that the free energy should have an interesting 't Hooft expression of the form $F = \sum_g \sum_h x^{2g-2+h}N^h F_{g,h}$. This motivates our next

manipulations. Applying the two formulas derived in Exercise 19.10 gives

$$\begin{aligned}F^{\text{pert}} &= \sum_{p=1}^{\infty}\sum_{k=0}^{p}\binom{2p+1}{2k}\frac{B_{2k}}{p(2p-2k+2)}(2\pi)^{-2p}\zeta(2p)N^{2p-2k+2}x^{2p}\\ &\quad-\sum_{p=1}^{\infty}\frac{1}{2p}(2\pi)^{-2p}\zeta(2p)N^{2p+1}x^{2p}\\ &\quad-\sum_{p=1}^{\infty}\sum_{k=0}^{p}\binom{2p+1}{2k}\frac{B_{2k}}{p(2p+1)}(2\pi)^{-2p}\zeta(2p)N^{2p-2k+2}x^{2p}\\ &\quad+\sum_{p=1}^{\infty}\frac{1}{2p}(2\pi)^{-2p}\zeta(2p)N^{2p+1}x^{2p}\\ &=\sum_{p=1}^{\infty}\sum_{k=0}^{p}\binom{2p+1}{2k}\frac{(2k-1)B_{2k}}{p(2p+1)(2p-2k+2)}(2\pi)^{-2p}\zeta(2p)N^{2p-2k+2}x^{2p}.\end{aligned}$$

We now re-index the double sum setting $g=k$ and $h=2p-2k+2$. Notice that

$$\begin{aligned}\binom{2p+1}{2k}\frac{2k-1}{p(2p+1)(2p-2k+2)} &= \binom{2p-1}{2p-2k+2}\frac{1}{k(2k-2)}\\ &= \binom{2g+h-3}{h}\frac{1}{g(2g-2)},\end{aligned}$$

for $g>1$. We conclude that

$$\begin{aligned}F^{\text{pert}} &= \sum_{g=2}^{\infty}\sum_{\substack{h \text{ even}\\ h\geq 2}}\binom{2g+h-3}{h}\frac{B_{2g}}{g(2g-2)}(2\pi)^{2-2g-h}\zeta(2g-2+h)N^h x^{2g-2+h}\\ &\quad+\sum_{\substack{h \text{ even}\\ h\geq 2}}\frac{1}{6h}(2\pi)^{-h}\zeta(h)N^h x^h-\sum_{\substack{h \text{ even}\\ h\geq 4}}\frac{2}{h(h-1)(h-2)}(2\pi)^{2-h}\zeta(h-2)N^h x^{h-2}.\end{aligned}\tag{81}$$

The second sum in the above expression is the $g=1$ term and the third sum is the $g=0$ term. In the original paper of Gopakumar and Vafa the last two sums were rewritten using polylogarithm identities (79) and (78) (see Gopakumar and Vafa [65, (3.7), (3.10)]). We prefer to keep them as they are and simplify the polylogarithms on the Gromov–Witten side instead.

The term

$$\sum_{j=1}^{N-1}(N-j)\ln\left(\frac{2\pi j}{k+N}\right)=\frac{N(N-1)}{2}\ln x+\sum_{j=1}^{N-1}(N-j)\ln j,$$

from the Chern–Simons free energy is called the Barnes term [65]. This is because one has

$$\sum_{j=1}^{N-1}(N-j)\ln j = \ln\Bigg(\prod_{j=1}^{N-1}(j!)\Bigg) = \ln(G_2(N+1)).$$

We use the asymptotic expansion (75) for the Barnes function to analyze this term in the final comparison.

19.3 The Gromov–Witten free energy

Recall from Section 12 that the restricted Gromov–Witten free energy is

$$\widehat{F}^{GW}_{X_{S^3}} = \sum_{g=0}^{\infty}\sum_{d=1}^{\infty} N_{g,d}\, y^{2g-2} e^{-td} = \sum_{d=1}^{\infty}\frac{1}{d}\Big(2\sin\frac{dy}{2}\Big)^{-2} e^{-td} = \sum_{d=1}^{\infty}\frac{1}{4d}\csc^2\Big(\frac{dy}{2}\Big)e^{-td}, \tag{82}$$

where t is assumed to be in the right half-plane for the series to converge.

Remark 19.20 The formula derived in Exercise 19.12 allows one to compute the signs of the Bernoulli numbers. Using these signs one can check that the following formulas that we give for the Gromov–Witten invariants agree with the formulas from Faber and Pandharipande [55].

We can rewrite equation (82) using the polylogarithms 19.16 and Exercise 19.7,

$$\begin{aligned}
\widehat{F}^{GW}_{X_{S^3}} &= \sum_{d=1}^{\infty} \frac{1}{4d} \csc^2(\frac{dy}{2}) e^{-td} \\
&= -\sum_{d=1}^{\infty} \frac{\partial}{\partial y} \left(\frac{1}{2d^2} \cot(\frac{dy}{2}) \right) e^{-td} \\
&= -\sum_{d=1}^{\infty} \frac{\partial}{\partial y} \left(\frac{i}{2d^2} \coth(\frac{i\,dy}{2}) \right) e^{-td} \\
&= -\sum_{d=1}^{\infty} \frac{1}{d^3} \frac{\partial}{\partial y} \left(\frac{i\,dy}{2} \coth(\frac{i\,dy}{2})/y \right) e^{-td} \\
&= -\sum_{d=1}^{\infty} \frac{1}{d^3} \frac{\partial}{\partial y} \left(\sum_{g=0}^{\infty} (-1)^g \frac{B_{2g}}{(2g)!} d^{2g} y^{2g-1} \right) e^{-td} \\
&= \sum_{g=0}^{\infty} \sum_{d=1}^{\infty} \frac{1}{d^{3-2g}} (-1)^{g-1} \frac{(2g-1)B_{2g}}{(2g)!} y^{2g-2} e^{-td} \qquad (83) \\
&= \sum_{g=0}^{\infty} \left((-1)^{g-1} (2g-1) B_{2g} \mathrm{Li}_{3-2g}(e^{-t})/(2g)! \right) y^{2g-2} \qquad (84) \\
&= \sum_{g=0}^{\infty} \widehat{F}^{X_{S^3}}_g y^{2g-2} . \qquad (85)
\end{aligned}$$

Comparing (83) with the definition of the Gromov–Witten free energy (82) we see that

$$N_{g,d}(X_{S^3}) = d^{2g-3}(-1)^{g-1}(2g-1)\frac{B_{2g}}{(2g)!} .$$

The last line in the above equation serves to define the genus g contribution to the restricted Gromov–Witten energy. In general the expansion of the free energy can be expressed as a sum of polylogarithms. The result is described in the following exercise.

Exercise 19.21 Expand the $p = 0$ term in the following expression as we did above, then express the sines in the remaining terms using exponentials and simplify with the binomial formula

$$\widehat{F}^{GW}_X = \sum_{p=0}^{\infty} \sum_{\beta} \sum_{d=1}^{\infty} n^p_\beta \frac{1}{d} \left(2 \sin \frac{dy}{2} \right)^{2p-2} e^{-d\langle t,\beta \rangle} .$$

The answer you should get is

$$\widehat{F}^{GW}(X) = \sum_{g=0}^{\infty}\Bigg(\sum_{\beta}\Big(n_\beta^0(-1)^g\frac{(2g-1)B_{2g}}{(2g)!}$$
$$+\sum_{p=1}^{\infty}\sum_{j=0}^{2p-2} n_\beta^p(-1)^{p+g}\binom{2p-2}{j}\frac{(1-p+j)^{2g-2}2g(2g-1)}{(2g)!}\Big)\mathrm{Li}_{3-2g}(e^{-\langle t,\beta\rangle})\Bigg)y^{2g-2}.$$

Combining the polylogarithm formula for the Gromov–Witten free energy (84) with the power series expansion of Li_{3-2g} for $g \geq 2$ (77) gives

$$\widehat{F}_g^{X_{S^3}} = \frac{B_{2g}}{2g(2g-2)}(it)^{2-2g} + \frac{B_{2g}}{g(2g-2)}\sum_{\substack{h \text{ even}\\ h\geq 0}}\binom{2g+h-3}{h}(2\pi)^{2-2g-h}\zeta(2g+h-2)(it)^h\,. \tag{86}$$

We next need to consider the $g = 1$ and $g = 0$ terms. The polylogarithm formula together with Lemma 19.18 gives

$$\widehat{F}_1^{X_{S^3}} = t/24 - \tfrac{1}{12}\ln t + \sum_{\substack{h \text{ even}\\ h\geq 2}}\frac{1}{6h}(2\pi)^{-h}\zeta(h)(it)^h\,. \tag{87}$$

and

$$\widehat{F}_0^{X_{S^3}} = \zeta(3) - \zeta(2)t + 3t^2/4 + t^3/12 - \frac{t^2}{2}\ln t - \sum_{\substack{h \text{ even}\\ h\geq 4}}\frac{2}{h(h-1)(h-2)}(2\pi)^{2-h}\zeta(h-2)(it)^h. \tag{88}$$

Note that in the definition of the Gromov–Witten free energy (82) the sum over degrees begins with $d = 0$ and we have not included $d = 0$. The degree 0 are constant maps and there is a question as to whether they should be included. The answer turns out to be 'yes'. Indeed, notice that the second term in (86) is almost completely identical to the first term in (81). The only difference is that summation there starts at $h = 2$ as opposed to $h = 0$. The extra $h = 0$ term reads

$$\frac{B_{2g}}{g(2g-2)}(2\pi)^{2-2g}\zeta(2g-2).$$

Using the result from Exercise 19.12 together with the fact that $\chi(X_{S^3}) = 2$ we compute

$$\frac{B_{2g}}{g(2g-2)}(2\pi)^{2-2g}\zeta(2g-2) = \frac{(-1)^g(2g-1)B_{2g}B_{2g-2}\,\chi(X_{S^3})}{2(2g-2)(2g)!}.$$

But for $g \geq 2$ this is exactly the negative of the degree zero invariants $N_{g,0}(X)$ computed in equation (19)! Genus 0 and 1 contributions line up as well. There are no contracted genus zero or one stable curves fixed by the torus action so $N_{0,0} = N_{1,0} = 0$. Combining the formulas we get the degree zero term

$$N_{g,0} = -\frac{B_{2g}}{g(2g-2)}(2\pi)^{2-2g}\zeta(2g-2) \tag{89}$$

that exactly cancels the extra $h = 0$ term. In other words, a major discrepancy between the Gromov–Witten and the Chern–Simons free energies takes care of itself if we include the degree zero terms (as we should have from the beginning). Also note that the sum over genus would diverge without the cancelation afforded by the degree zero term. This is why the definition of the Gromov–Witten free energy includes the constant terms.

As we explained $F_1 = \widehat{F}_1$, $F_0 = \widehat{F}_0$ and (86) turns into

$$\begin{aligned} F_g(X_{S^3}) &= \frac{B_{2g}}{2g(2g-2)}(it)^{2-2g} \\ &\quad + \frac{B_{2g}}{g(2g-2)}\sum_{\substack{h \text{ even} \\ h\geq 2}} \binom{2g+h-3}{h}(2\pi)^{2-2g-h}\zeta(2g+h-2)(it)^h. \end{aligned} \tag{90}$$

We are all set for the final comparison of the free energies on both sides of the duality.

19.4 The final comparison

As we warned in the introduction the match between the Gromov–Witten and the Chern–Simons free energies will not be exact. The discrepancy may be due to the fact that as physicists insist, we should really consider the $\mathrm{U}(N)$ not $\mathrm{SU}(N)$ Chern–Simons theory which is expected to insert some additional normalizing factors into the partition function (see Mariño [103; 105]). Combining equations (80), (81)) and the asymptotic expansion for the Barnes term (75) gives the following expression for the unnormalized

Chern–Simons free energy,

$$\begin{aligned}
F^{\mathrm{CS}}_{S^3}(N,x) &= \tfrac{1}{2}N(N-1)\ln x + \tfrac{1}{2}(1-N)\ln(k+N) + \tfrac{1}{2}N^2\ln N \\
&\quad - \tfrac{1}{2}\ln N - \tfrac{3}{4}N^2 - \tfrac{1}{12}\ln N - \zeta'(0)N + \zeta'(-1) \\
&\quad - \sum_{\substack{h \text{ even} \\ h\geq 4}} \frac{2}{h(h-1)(h-2)}(2\pi)^{2-h}\zeta(h-2)N^h x^{h-2} \\
&\quad + \sum_{\substack{h \text{ even} \\ h\geq 2}} \frac{1}{6h}(2\pi)^{-h}\zeta(h)N^h x^h \\
&\quad + \sum_{g=2}^{\infty}\sum_{\substack{h \text{ even} \\ h\geq 2}} \binom{2g+h-3}{h}\frac{B_{2g}}{g(2g-2)}(2\pi)^{2-2g-h}\zeta(2g-2+h)N^h x^{2g-2+h} \\
&\quad + \sum_{g=2}^{\infty}\frac{B_{2g}}{2g(2g-2)}N^{2-2g}.
\end{aligned}$$

In the same way we combined terms to get the Chern–Simons free energy, Definition 12.1 and equations (88), (87) and (90) give the following expression for the full Gromov–Witten free energy of the resolved conifold,

$$\begin{aligned}
F^{\mathrm{GW}}_{X_{S^3}}(t,y) &= \tfrac{1}{24}t - \tfrac{1}{12}\ln t + \zeta(3)y^{-2} - \zeta(2)ty^{-2} + 3t^2y^{-2}/4 \\
&\quad + t^3y^{-2}/12 - \tfrac{1}{2}t^2y^{-2}\ln t \\
&\quad - \sum_{\substack{h \text{ even} \\ h\geq 4}} \frac{2}{h(h-1)(h-2)}(2\pi)^{2-h}\zeta(h-2)(it)^h y^{-2} \\
&\quad + \sum_{\substack{h \text{ even} \\ h\geq 2}} \frac{1}{6h}(2\pi)^{-h}\zeta(h)(it)^h \\
&\quad + \sum_{g=2}^{\infty}\frac{B_{2g}}{g(2g-2)}\sum_{\substack{h \text{ even} \\ h\geq 2}} \binom{2g+h-3}{h}(2\pi)^{2-2g-h}\zeta(2g+h-2)(it)^h y^{2g-2} \\
&\quad + \sum_{g=2}^{\infty}\frac{B_{2g}}{2g(2g-2)}(it)^{2-2g}y^{2g-2}.
\end{aligned}$$

Note that some of the extra terms that appeared 'on the Chern–Simons side' in the original paper [65] show up 'on the Gromov–Witten side' with opposite signs in our

presentation. This is because we chose not to reexpress the genus 0 and 1 contributions in the Chern–Simons free energy via the polylogarithm identities. By inspection, under the substitution $it = \pm Nx$ and $y = x$ all the infinite sums match exactly. In light of the complicated definitions and expressions for the free energies this is a remarkable coincidence. Notice that the sums represent exactly the perturbative part of the Chern–Simons free energy and thus contain all of the information about the perturbative invariants.

Analytically, we are comparing series expansions of two functions near the origin $(t, y) = (0, 0)$. It may seem odd that we should choose the origin since $(t, y) = \left(\frac{2\pi iN}{k+N}, \frac{2\pi}{k+N}\right)$ converges to $(2\pi i, 0)$ at large N. However, as one can see for example from (82) the free energy is periodic in t with period $2\pi i$ so the coefficients are the same as the coefficients at the origin. Another issue is that originally in (82) we assumed the real part of t to be positive. The problem with analytically continuing the free energies to a punctured neighborhood of the origin is that the logarithmic terms in both expressions are ambiguous. However, for us the free energies are just a bookkeeping device for the invariants on both sides of the duality. Since logarithmic and other mismatching terms outside the infinite sums carry no apparent geometric information they do not pose a serious problem. Finally, notice that the infinite sums are real-valued which allows us to package the comparison into the following nice form.

Theorem 19.22 *The full Gromov–Witten free energy and the unnormalized Chern–Simons free energy are related by*

$$\mathrm{Re}\big(F^{GW}_{X_{S^3}}(iNx, x) - F^{CS}_{S^3}(N, x)\big) = \tfrac{5}{12}\ln x + \zeta(3)x^{-2} - \tfrac{1}{2}\ln(2\pi) - \zeta'(-1)\,.$$

Proof Combining the expressions for the free energies gives,

$$\begin{aligned}\mathrm{Re}\big(F^{\mathrm{GW}}_{X_{S^3}}(iNx, x)\big) - F^{\mathrm{CS}}_{S^3}(N, x) = \\ \tfrac{1}{2}(N-1)\ln(k+N) - \tfrac{1}{2}N(N-1)\ln x - \tfrac{1}{2}N^2\ln N + 3N^2/4 + \tfrac{1}{12}\ln N + \zeta'(0)N - \zeta'(-1) \\ -\mathrm{Re}\big(\tfrac{1}{12}\ln(ixN)\big) + \zeta(3)x^{-2} - \tfrac{3}{4}N^2 + \tfrac{1}{2}N^2\mathrm{Re}(\ln(ixN)).\end{aligned}$$

Using the fact that $x = \frac{2\pi}{k+N}$ to write $\ln(k+N)$ and the value $\zeta'(0) = -\frac{1}{2}\ln(2\pi)$ from Sondow [146] allows one to combine like terms further to obtain the result. □

We conclude our presentation of the Gopakumar–Vafa duality with this remarkable equality.

Remark 19.23 Large N duality is said to be exact when the full free energies are equal. It is said to hold to a leading order when the genus zero terms agree. Given this the term $\zeta(3)x^{-2}$ in the comparison theorem is slightly disturbing. This term prevents the duality from holding at the level of the genus zero contributions. In the physical theory this term is canceled by additional genus zero corrections in degree zero. Ooguri and Vafa [121] obtained a perfect agreement of the two sides using the *physical normalization* of the S^3 Chern–Simons free energy (which can only be computed on a case by case basis comparing exact answers to perturbative expansions). Thus, we expect that one will obtain an exact correspondence after enough examples have been computed to find a general form of the correct normalization.

20 New results (2003–2006)

In this section we describe some recent directions of research inspired by the Large N Duality and discuss some difficulties and open problems encountered within them. Obviously this account is biased by our background and interests, and we apologize in advance for any inaccuracies and/or omissions. As in the history Section 1 the dates in the text refer to arxiv submissions while references are given wherever possible to journal publications.

20.1 Computations of the Gromov–Witten invariants

Computational verification of the Gopakumar–Vafa Large N Duality depends largely on one's ability to compute the Gromov–Witten invariants for as large a class of threefolds as possible. Toric threefolds seem to be natural candidates to start with since holomorphic torus actions are a part of their definition and the full power of virtual localization can be applied. However, as the sample computations in Section 7.3 show, the complexity of expressions obtained through virtual localization often grows very rapidly with degree and genus and quickly becomes unmanageable.

Aganagic, Mariño and Vafa introduced an interesting way to attack this computation for local toric Fano surfaces [5]. Iqbal found a nice reformulation of these results [79] and Zhou gave a mathematical proof of the results. This Aganagic, Mariño and Vafa paper led to a breakthrough by Aganagic, Klemm, Mariño and Vafa in [3], where an effective algorithm was offered that produces explicit combinatorial answers (without Hodge integrals, etc) for all toric Calabi–Yau threefolds. The idea is that any toric Calabi–Yau threefold (which is necessarily non-compact) can be presented by a labeled planar trivalent graph that can be cut into trivalent vertices 'with legs' representing $\mathbb{C}^3$ patches. Labels on the edges provide the gluing data that specifies the threefold.

The topological vertex is an explicit function $C_{\vec{\mu},\vec{n}}(\lambda)$ of the edge labels at each vertex, three partitions (μ^1,μ^2,μ^3) and three integers (n^1,n^2,n^3) associated to each vertex of the graph. The generating function of the Gromov–Witten invariants of the threefold can then be written as a 'state sum' of these $C_{\vec{\mu},\vec{n}}(\lambda)$ taken over additional labelings. The authors of [3] provided an explicit combinatorial expression for $C_{\vec{\mu},\vec{n}}(\lambda)$ based on a derivation assuming large N duality.

This algorithm has been almost proved mathematically by Li, Liu, Liu and Zhou [94] based on gluing formulas for relative Gromov–Witten invariants. However, mathematical redefinition leads to a seemingly different expression $\widetilde{C}_{\vec{\mu},\vec{n}}(\lambda)$ for the topological vertex. The equality

$$C_{\vec{\mu},\vec{n}}(\lambda)=\widetilde{C}_{\vec{\mu},\vec{n}}(\lambda)$$

has been verified for the case when one of the partitions μ^i is empty or when all partitions have length ≤ 6 but the general case remains open.

Another interesting class of Calabi–Yau threefolds is given by local curves. Those are the total spaces of rank 2 complex vector bundles N over a complex curve Σ with $c_1(N)=2g(\Sigma)-2$, for example the resolved conifold is a local $\mathbb{C}\mathrm{P}^1$. Although not toric in general these threefolds always admit 'degenerate' holomorphic torus actions that leave the entire zero section fixed (as opposed to isolated points in the usual case). In [36] J Bryan and R Pandharipande used relative Gromov–Witten invariants and the TQFT approach to construct a recursive algorithm that computes the invariants of any local curve.

Next in complexity is the case of local surfaces, that is, total spaces of canonical bundles K_S to complex surfaces S. The work of D-E Diaconescu, B Florea and others on the invariants of local del Pezzo surfaces [46] culminated in the joint work with N Saulina [48] that extends the toric topological algorithm to the case of local ruled surfaces with a finite number of reducible fibers. As in the case of local curves the authors make use of degenerate torus actions that fix finitely many curves and augment the toric formalism by the corresponding correction terms. The derivation of the combinatorial formulae uses physical arguments as in [3] and mathematical justification of the ruled vertex is an open problem. Another open problem is to generalize this algorithm to arbitrary Calabi–Yau threefolds with degenerate torus actions.

From the point of view of Large N Duality it also important to understand the pre-duals of the above threefolds, that is, analogs of T^*S^3 for the resolved conifold and identify the correct pre-dual theories. In the known examples as originally considered by Aganagic and Vafa [7] several 2–cycles are collapsed and then replaced by Lagrangian 3–cycles via resolving a singular deformation. These pre-duals are thus of a more general form than T^*M [5; 46]. The corresponding theories combine elements of both

the Chern–Simons and the Gromov–Witten theories in agreement with Witten's original idea that the Chern–Simons theory on M is the correct 'Gromov–Witten theory' on T^*M (see Witten [161] and Grassi and Rossi [67, Appendix 9]). It was the formalism from [5] that led to the discovery of the topological vertex.

20.2 Intermediate theories

In this section we discuss some theories that have recently emerged and could provide a bridge between the two sides of the large N duality. The most developed of these is the Donaldson–Thomas theory that has already been used to prove some of the duality's structural predictions.

Gopakumar and Vafa predicted in [65] that the properly normalized partition function of the Gromov–Witten invariants $Z_X(\lambda, v)$ on a Calabi–Yau threefold X is a rational function of the variable $q = -e^{i\lambda}$ and expands into a series in q with integral coefficients (BPS states or Gopakumar–Vafa invariants). Intuitively the integers should correspond to counts of embedded curves in X. Classically, embedded curves are described by ideal sheaves on X, that is, torsion-free rank one sheaves with trivial determinants. S Donaldson and R Thomas introduced in [50] a new class of invariants $\widetilde{N}_{\chi,\beta}$ that count the number of ideal sheaves with a given holomorphic Euler characteristic χ and the associated curve class β. The Donaldson–Thomas theory is 'better' than the Gromov–Witten theory in the sense that no orbifolds occur as moduli spaces and the numbers $\widetilde{N}_{\chi,\beta}$ are integers. In [106; 107] D Maulik, N Nekrasov, A Okounkov and R Pandharipande conjectured that the Donaldson–Thomas partition function $\widetilde{Z}_X(q, v) = \sum_{\chi,\beta} \widetilde{N}_{\chi,\beta} q^\chi v^\beta$ turns into $Z_X(\lambda, v)$ after the change of variables $q = -e^{i\lambda}$. This automatically implies the rationality and integrality predictions for all Calabi–Yau threefolds. Moreover, the authors offered the 'equivariant vertex' algorithm analogous to the topological vertex for computing the Donaldson–Thomas invariants. Combined with [94] and [36] their result proves the Gromov–Witten/Donaldson–Thomas duality for toric Calabi–Yau threefolds and local curves respectively. Obviously it is desirable to extend the duality to local surfaces and more general threefolds.

It is expected that the Donaldson–Thomas theory admits a gauge-theoretic interpretation and if so it could serve as a link in a chain mathematically connecting Gromov–Witten theory to Chern–Simons theory. In particular there are some promising connections discovered between the BPS states and Yang–Mills theory in two dimensions [6] and in four dimensions [119].

Another possible intermediary is the symplectic field theory (SFT) of Y Eliashberg, A Givental and H Hofer [53]. In general SFT studies invariants of a contact manifold $\mathcal{C}$ by considering moduli of pseudoholomorphic curves in its symplectization $\mathcal{C} \times \mathbb{R}$. One can

naturally associate a contact manifold to any 3–dimensional manifold M, namely the cosphere bundle $S(T^*M)$ with symplectization $S(T^*M)\times\mathbb{R}\simeq T^*M\setminus M$. Unlike T^*M itself $T^*M\setminus M$ does admit nontrivial pseudoholomorphic curves that may serve as Witten's 'instantons at infinity' [161]. Another attractive trait of SFT is that it reconstructs some knot invariants, for example the Alexander polynomial, from the Gromov–Witten invariants (see L Ng's paper [116] in this volume). The main challenge in applying SFT to Large N Duality is the scarcity of effective algorithms for computing the invariants.

We should also mention an older approach to proving large N duality suggested by B Acharya [1] and developed by M Atiyah, J Maldacena and C Vafa [16] (see also Grassi and Rossi [67]) by lifting both sides to the M–theory on a 7–dimensional manifold with G_2 holonomy. However, so far the M–theory approach (M for mystery) has not been very fruitful mathematically because the geometry of G_2 manifolds is much less understood than that of Calabi–Yau threefolds.

20.3 Construction of large N duals

There are very few known large N dual pairs despite the large number of known Calabi–Yau threefolds with computable Gromov–Witten invariants. For a while after 1998 the original $T^*S^3/\mathcal{O}(-1)\oplus\mathcal{O}(-1)$ example remained the only one. In 2001 F Cachazo, K Intriligator and C Vafa constructed a family of examples [37] with the deformed conifold T^*S^3 replaced by the following hypersurface in $\mathbb{C}^4$:

$$W'(x)^2+f(x)+y^2+z^2+w^2=0.$$

Here $W(x)$, $f(x)$ are polynomials of degrees n, $n-2$ respectively and the deformed conifold is recovered for $W(x)=x^2/2$, $f=\text{const}$. However, they do not provide new examples of the form T^*M, which is where it is easiest to compute the Chern–Simons side. In fact, the only known examples of this form come from the spherical quotients.

The first published version of Large N Duality for the lens spaces $L(p,1)=S^3/\mathbb{Z}_p$ appeared in Giveon, Kehagias and Partouche[64] (see also Mariño [4] for other credits). The idea is to extend the group $\Gamma=\mathbb{Z}_p$ action to T^*S^3, pull it through the transition to $\mathcal{O}(-1)\oplus\mathcal{O}(-1)$ and then resolve the resulting quotient singularity (see Halmagyi, Okuda and Yasnov [70] for details). The dual quotient is a bundle over $\mathbb{C}\mathrm{P}^1$ fibered by surface singularities $\mathbb{C}^2/\Gamma$ and one can obtain a threefold resolution by resolving the surface singularities in each fiber. For the resolved quotient to be Calabi–Yau the resolution must be crepant (see Harris [73]). This restricts the list of groups Γ to finite subgroups of SU(2) (in particular this is why the lens spaces $L(p,q)$ with $q\neq 1$ are out). The geometry of the transition for such quotients and more general fibrations

of surface ADE–singularities that also include the [37] examples is studied in [38]. One may be able to lift the SU(2) restriction by considering orbifold Gromov–Witten invariants. The crepant resolution conjecture proved for the affine ADE–singularities by F Perroni [122] says roughly that the orbifold Gromov–Witten invariants of a variety are equal to the ordinary Gromov–Witten invariants of its crepant resolution (when such exists). Thus conjecturally Large N Duality for the quotients with $\Gamma \subset \mathrm{SO}(4)$ should relate the Chern–Simons invariants of S^3/Γ to the orbifold Gromov–Witten invariants of $\mathcal{O}(-1) \oplus \mathcal{O}(-1)/\Gamma$.

Computationally only the $L(p,1)$ example has been considered so far [4; 70]. Here the Chern–Simons partition function is known (see for example Hansen and Takata [71], Garoufalidis and Mariño [63] and Turaev [152]) and the Gromov–Witten free energy is computable via the topological vertex since the dual is toric. The duality statement is more complicated than in the S^3 case and one has to split the Chern–Simons partition function into contributions from different (gauge classes of) flat connections before comparing to the Gromov–Witten side [4]. In addition to the analysis in [4], there is an indirect check of the duality based on mirror symmetry in [70]. In general, the precise meaning of the duality for more general manifolds remains an open question.

This justifies interest in explicit computations of the LMO invariant (see Le, Murakami and Ohtsuki [93] and Bar-Natan, Garoufalidis, Rozansky and Thurston [26]) that is believed to capture the contribution of the trivial connection into the full Chern–Simons partition function. M Mariño showed in [104] using physical considerations related to mirror symmetry that the LMO invariant can be expressed as a perturbed Gaussian matrix integral for Seifert fiber spaces. This reduces the computation to a solvable matrix model (see Aganagic, Klemm, Mariño and Vafa [4], Fiorenza and Murri [57] and Mariño [105]). The paper [104] also presented some sample computations. Recently S Garoufalidis and M Mariño [63] gave a mathematical derivation of the matrix model for general rational homology spheres based on the Århus integral presentation of the LMO invariant [26]. An interesting open question is just how much of the Gromov–Witten theory on the dual can be recovered from the LMO invariant.

There is also a duality involving $SO(N)$ or $Sp(N)$ Chern–Simons theories discussed in the physics literature [144; 47; 31; 32].

Finally we mention the symplectic surgery approach of I Smith and R Thomas to constructing large N duals [145]. Their work suggests that in many cases such duals ought to be 'non-Kähler Calabi–Yaus', that is, non-Kähler symplectic manifolds X with $c_1(X) = 0$. If so, this explains why so few duals to cotangent bundles are known despite the abundance of known Kähler Calabi–Yau threefolds.

Part IV Appendices

Appendix A Stacks

In this section we follow the excellent exposition covering stacks in Metzler [109], and just add a couple of motivating examples. Stacks were introduced to encode the structure of an orbifold in the category of schemes, but may also be used to define orbifolds in the smooth, topological and analytic categories. We will provide examples in the smooth category. The main geometric objects that we consider may be represented by a category with additional structure, a contravariant functor or a covariant functor. We will conclude this section with a list of properties that such a covariant functor satisfies if and only if it comes from an object of the category in a natural way. The generalized objects that are used in Gromov–Witten theory are just covariant functors that satisfy a subset of these properties.

Recall that an orbifold is a space locally modeled on the quotient of $\mathbb{R}^n$ by a finite group action together with additional data to measure the stabilizer subgroups. As a first example, consider the quotient of $\mathbb{C}$ by the natural left action of $\mathbb{Z}_3$ by multiplication by cube roots of unity. We can encode this as a category X with objects $\mathrm{Ob}(X) := \mathbb{C}$ and arrows $\mathrm{Ar}(X) := \mathbb{Z}_3 \times \mathbb{C}$ where we consider $(\omega, z) \in \mathrm{Mor}(z, \omega z)$. The underlying space of the associated orbifold is the quotient of the objects obtained by identifying those objects connected by a morphism. The stabilizer group of a point $z \in \mathrm{Ob}(X)$ is just $\mathrm{Mor}(z, z)$. This is a special category because every morphism has an inverse. Such a category is called a groupoid. In fact this has the structure of a smooth groupoid. The structure maps in this example are given by (source – $s\colon \mathrm{Ar}(X) \to \mathrm{Ob}(X)$, $s(\omega, z) = z$; target – $t\colon \mathrm{Ar}(X) \to \mathrm{Ob}(X)$, $t(\omega, z) = \omega z$; inverse – $i\colon \mathrm{Ar}(X) \to \mathrm{Ar}(X)$, $i(\omega, z) = (\omega^{-1}, \omega z)$; composition – $m\colon \mathrm{Mor}(\omega z, \theta\omega z) \times \mathrm{Mor}(z, \omega z) \to \mathrm{Mor}(z, \theta\omega z)$, $m(f, g) = f \circ g$)

Definition A.1 A smooth groupoid is a category X with invertible morphisms such that $\mathrm{Ob}(X)$ and $\mathrm{Ar}(X)$ are smooth manifolds and the various structure maps are smooth.

One good example to keep in mind is the smooth groupoid associated to any atlas on a smooth manifold. Given an atlas, $\mathcal{A} = \{\varphi_\alpha\colon U_\alpha \to V_\alpha\}$ define a smooth groupoid with $\mathrm{Ob}(X^{\mathcal{A}}) := \coprod_\alpha V_\alpha$ and $\mathrm{Ar}(X^{\mathcal{A}}) := \coprod_{\alpha,\beta} \varphi_\alpha(U_\alpha \cap U_\beta)$ with the obvious structure maps and smooth structures.

Exercise A.2 Combine the example of the $\mathbb{Z}_3$ quotient of $\mathbb{C}$ with the smooth groupoid associated to an atlas to define a smooth groupoid modeling an orbifold with underlying

space homeomorphic to S^2, one point with stabilizer $\mathbb{Z}_3$, one point with stabilizer $\mathbb{Z}_2$ and the rest of the points having trivial stabilizer.

We now turn to the second way to encode a geometric object – a contravariant functor. Let **DIFF** be the category of smooth manifolds and **SET** be the category of sets. Given a smooth manifold M, we define a contravariant functor $\underline{M}\colon \mathbf{DIFF} \Longrightarrow \mathbf{SET}$ by $\underline{M}(N) := C^\infty(N, M)$ and $\underline{M}(f\colon N \to P) := f^*\colon C^\infty(P, M) \to C^\infty(N, M)$. It is possible to reconstruct the original manifold (up to diffeomorphism) from the associated functor. We therefore think of a contravariant functor $\mathcal{M}\colon \mathbf{DIFF} \Longrightarrow \mathbf{SET}$ as a generalized manifold. (Recall that we described the moduli stack as a contravariant functor from **SCHEME** to **SET**.) However, we will have to add some restrictions in order to have a reasonable family of generalizations. When we add these restrictions to give the formal definition of a stack we will use a third description of geometric objects. This third description will generalize the first two frameworks.

Given a contravariant functor $\mathcal{M}\colon \mathbf{DIFF} \Longrightarrow \mathbf{SET}$ one can define a new category $\mathbf{D}^{\mathcal{M}}$ with

$$\mathrm{Ob}(\mathbf{D}^{\mathcal{M}}) := \coprod_N \mathcal{M}(N)$$
$$\text{and} \quad \mathrm{Mor}((\alpha, N), (\beta, P)) := \{a \in C^\infty(N, P) | \mathcal{M}(a)(\beta) = \alpha\}.$$

One then defines a covariant functor $F^{\mathcal{M}}\colon \mathbf{D}^{\mathcal{M}} \Longrightarrow \mathbf{DIFF}$ by $F^{\mathcal{M}}(a\colon (\alpha, N) \to (\beta, P)) := a\colon N \to P$. When the contravariant functor is of the form $\underline{M}$ the associated covariant functor will satisfy a number of special properties. The first property that it will satisfy is that it will be a fibered category.

Definition A.3 A fibered category over $\mathbf{C}$ is a covariant functor $F\colon \mathbf{D} \Longrightarrow \mathbf{C}$ such that

(1) For every $f\colon C_0 \to C_1 \in \mathrm{Ar}(\mathbf{C})$ and every D_1 such that $F(D_1) = C_1$ there is an arrow $g\colon D_0 \to D_1$ such that $F(g\colon D_0 \to D_1) = f\colon C_0 \to C_1$.

(2) If $F(g_1) \circ f = F(g_0)$, then there is a unique $g \in \mathrm{Ar}(\mathbf{D})$ such that $F(g) = f$.

Exercise A.4 Check that $F^{\underline{M}}\colon \mathbf{D}^{\underline{M}} \Longrightarrow \mathbf{DIFF}$ is a fibered category.

It is also possible to construct a fibered category associated to a smooth groupoid. Let X be a smooth groupoid and define a category ${}^X\mathbf{D}$ with

$$\mathrm{Ob}({}^X\mathbf{D}) := \coprod_{N \in \mathrm{Ob}(\mathbf{DIFF})} C^\infty(N, \mathrm{Ob}(X)),$$
$$\mathrm{Mor}(f\colon N \to \mathrm{Ob}(X), g\colon P \to \mathrm{Ob}(X)) :=$$
$$\big\{(\varphi\colon N \to P, h\colon N \to \mathrm{Ar}(X)) \mid s(h(p)) = f(p) \text{ and } t(h(p)) = g(\varphi(p))\big\}.$$

The associated covariant functor is given by,

$$^{X}F(\varphi\colon N \to P, h\colon N \to \mathrm{Ar}(X)) := \varphi\colon N \to P\,.$$

Exercise A.5 Check that $^{X}F\colon {}^{X}\mathbf{D} \Longrightarrow \mathbf{DIFF}$ is a fibered category. (Recall that the arrows in a groupoid have inverses.)

Let $F\colon \mathbf{D} \Longrightarrow \mathbf{C}$ be a fibered category and C be an object of $\mathbf{C}$. We can define a fibered category over C, denoted F_C, by $\mathrm{Ob}(F_C) := \{D \in \mathrm{Ob}(\mathbf{D}) | F(D) = C\}$ and $\mathrm{Ar}(F_C) := \{\varphi\colon D_0 \to D_1 | F(\varphi) = \mathrm{id}_C\}$. One can see that F_C is a groupoid.

Exercise A.6 Prove that F_C is a groupoid.

One often wishes to put additional structure on a category. A good motivating example is the category of open sets of a topological space with inclusions as arrows. In this setting, one would like to axiomatize the properties of open covers of the original space. This leads to the notion of a Grothendieck topology and the notion of a site. Once a category has a notion of coverings one can define an analogue of a sheaf. This is one way to introduce stacks. We do not need the definitions of Grothendieck topologies or sites, but we do use the following definition that we quote from Metzler [109] to encode the notion of a covering.

Definition A.7 A basis for a Grothendieck topology on a category $\mathbf{C}$ is a function K which assigns to every object C of $\mathbf{C}$ a collection $K(C)$ of families of arrows with target C, called covering families, such that

(1) if $f\colon C' \to C$ is an isomorphism, then $\{f\}$ is a covering family;

(2) (stability) if $\{f_i\colon C_i \to C\}$ is a covering family, then for any arrow $g\colon D \to C$, the pullbacks $C_i \times_C D$ exist and the family of pullbacks $\pi_2\colon C_i \times_C D \to D$ is a covering family (of D);

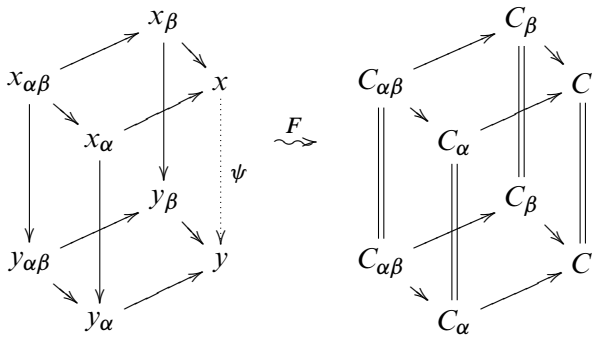
(3) (transitivity) if $\{f_i\colon C_i \to C \,|\, i \in I\}$ is a covering family and for each $i \in I$, one has a covering family $\{g_{ij}\colon D_{ij} \to C_i \mid j \in I_i\}$, then the family of composites $\{f_i g_{ij}\colon D_{ij} \to C \,|\, i \in I, j \in I_i\}$ is a covering family.

In addition to the usual notion of an open cover in the category of open sets, we obtain an example of a basis for a Grothendieck topology on **DIFF** by considering collections of open embeddings whose images cover a given manifold. This example will be used in our definition of prestack and stack.

We now have two different constructions of special fibered categories. One starts with a manifold and passes to an associated contravariant functor and then to the associated

fibered category. The second passes from a manifold to a smooth groupoid to the associated fibered category. The fibered categories constructed in either of these ways satisfy additional conditions summarized in the definition of a prestack. We take the characterization of a prestack given in [109, Lemma 25] as the definition of a prestack.

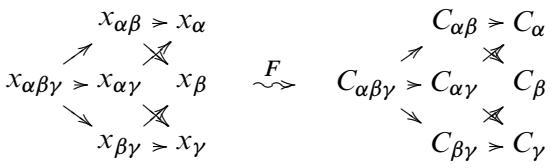
Definition A.8 A fibered category $F\colon \mathbf{C} \Longrightarrow \mathbf{D}$ is a prestack if and only if there is a unique arrow $\psi\colon x \to y$ filling in the dotted arrow in every diagram of the following form with $F(\psi) = 1$. In this diagram $C \in \mathrm{Ob}(\mathbf{C})$, $\{C_\alpha \to C\}$ is a cover of C and x, y are objects of $\mathbf{D}$ that map to C and $x_\alpha \in \mathbf{D}$ map to C_α under F. In addition we denote fibered products as $C_{\alpha\beta} := C_\alpha \times_C C_\beta$ $(x_{\alpha\beta})$.



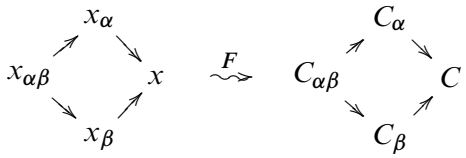
Exercise A.9 Check that $F^{\underline{M}}$ and ${}^{X}F$ are prestacks.

We finally come to the definition of a stack. The definition we give is not the usual one. It is instead the characterization given in Metzler [109, Lemma 32]. We chose to use this as the definition because it was the quickest way to a clean definition. For the more typical description of a stack in terms of descent data see the full exposition in [109] or Behrend [28].

Definition A.10 A prestack $F\colon \mathbf{D} \Longrightarrow \mathbf{C}$ is a stack if and only if for every cover $\{C_\alpha \to C\}$ in $\mathbf{C}$ and x_α mapping to C_α satisfying the following commutative diagram for all index triples



there is an object x in $\mathbf{D}$ mapping to C and arrows $\{x_\alpha \to x\}$ filling in the commutative diagram



Exercise A.11 Prove that $F^{\underline{M}}$ is a stack.

One may be expecting an exercise to prove that ${}^{X}F$ is a stack. However, the fibered category associated to a typical groupoid is not a stack. A good example to consider is the groupoid that arises from the standard two-chart atlas of $\mathbb{C}\mathrm{P}^1$. In this case the smooth groupoid has $\mathrm{Ob}(X) = \mathbb{C} \coprod \mathbb{C}$ corresponding to the two charts and $\mathrm{Ar}(X) = \mathbb{C} \coprod \mathbb{C} \coprod \mathbb{C}^\times \coprod \mathbb{C}^\times$ corresponding to the overlaps. The objects of the domain category of the associated fibered category are smooth maps from smooth manifolds into $\mathrm{Ob}(X)$. The problem arises when one considers a map into $\mathbb{C}\mathrm{P}^1$ with image that is not contained in one of the coordinate charts, for example $x\colon \mathbb{C} \to \mathbb{C}\mathrm{P}^1$ given by $x(z) := [z-1 : z+1]$. By restricting to the charts and their overlaps we obtain objects of the domain of the fibered category $x_\pm\colon \mathbb{C} - \{\mp 1\} \to \mathbb{C}\mathrm{P}^1$ given by $x_\pm(z) := \frac{z \mp 1}{z \pm 1}$. These objects map to the charts under the fibered category and the charts form a cover of $\mathbb{C}\mathrm{P}^1$. The pull-backs of the cover and the $x_\pm$ satisfy the diagram in the hypothesis of the definition of a stack but not the required extension property. It is possible to stackify a prestack by declaring objects to be equivalence classes of objects over elements of covers as in the hypothesis of the definition. This is similar to the sheafification of a presheaf. See Metzler [109] for details. When the prestack associated to the groupoid associated to an atlas is stackified, the domain category has objects that correspond exactly to smooth maps into the given manifold. Thus the constructions of ${}^{X}F$ and $F^{\underline{M}}$ agree when one starts with a smooth manifold and stackifies.

Exercise A.12 Prove that the pull-backs of the cover and the $x_\pm$ satisfy the diagram in the hypothesis of the definition of a stack.

We now need to consider maps of stacks. As motivation consider what can be constructed between the stacks associated to a pair of manifolds from a map between the manifolds, $f\colon M \to N$. Recall that the objects of the domain category of the stack associated to M are just maps $\alpha\colon P \to M$. Such a map can be taken to $f \circ \alpha\colon P \to N$, which is an object of the domain category associated to N. This extends to arrows in the natural way to give a covariant functor $A_f\colon \mathbf{D}^{\underline{M}} \Longrightarrow \mathbf{D}^{\underline{N}}$. This motivates the definition of a map between stacks given below.

Definition A.13 A map between stacks $F\colon \mathbf{D} \Longrightarrow \mathbf{C}$ and $G\colon \mathbf{E} \Longrightarrow \mathbf{C}$ say $A\colon F \to G$ is just a covariant functor $A\colon \mathbf{D} \Longrightarrow \mathbf{E}$ such that $F = G \circ A$.

There are some technical issues that arise when one wishes to define isomorphism of stacks, which are best addressed with 2–categories, see [109].

We now quote some definitions of some properties of stacks and maps of stacks from [109].

Definition A.14 Let $A\colon F' \to F$ be a map of fibered categories over $\mathbf{C}$. We say A is a monomorphism if, for every object C of $\mathbf{C}$, the functor $A_C\colon F'_C \to F_C$ on fibers is fully faithful.

Definition A.15 We say a map of fibered categories $A\colon F' \to F$ is covering (French 'couvrant') if, for every object C of $\mathbf{C}$, and every object D of F_C, there is a covering family $\{f_i\colon C_i \to C\}$ and for every i an object D_i of F'_{C_i} such that $A(D_i) \cong D|_{C_i}$.

If F' and F are stacks, then we refer to a covering map as an epimorphism.

For the next definitions we need the definition of the fibered product or pull-back of a pair of maps of stacks. Given $A\colon F \to H$ and $B\colon G \to H$ one defines the pull-back stack $F \times_H G$ to be the stack with domain category having objects $\mathrm{Ob}(\mathbf{D}_F) \times_{\mathrm{Ob}(\mathbf{D}_H)} \mathrm{Ob}(\mathbf{D}_G)$, that is, ordered pairs (x, y) such that $A(x) = B(y)$. We can generalize the construction of a contravariant functor $\underline{M}$ associated to a smooth manifold to any object C in any category $\mathbf{C}$. If the category is reasonably well behaved the associated covariant functor $F^{\underline{C}}$ will be a stack.

The following definition generalizes the notions of manifold and a submersion between smooth manifolds.

Definition A.16 A stack F is representable if and only if it is isomorphic to a stack of the form $F^{\underline{C}}$ for some object C of the base category $\mathbf{C}$. Let F, G be stacks over $\mathbf{C}$. We say that a map $A\colon F \to G$ is representable if for every object C of $\mathbf{C}$ and map $B\colon F^{\underline{C}} \to G$, the pull-back stack $F^{\underline{C}} \times_G F$ is representable.

In fact if $f\colon M \to N$ is a smooth map, the associated map of stacks $A_f\colon F^{\underline{M}} \to F^{\underline{N}}$ is representable if and only if f is a submersion, see Metzler [109].

Definition A.17 Let F be a stack over $\mathbf{C}$. We say F is locally representable if there is an object C of $\mathbf{C}$ and a representable epimorphism $A\colon F^{\overline{C}} \to F$.

Definition A.18 Let P be a property of maps in $\mathbf{C}$ that is stable under pullback. We say that a representable map $A\colon F \to G$ has property P if, for every object C of $\mathbf{C}$ and map $B\colon F^{\underline{C}} \to G$, the projection $B^*A\colon F^{\underline{C}} \times_G F \to C$ has property P.

We now come to the result characterizing representable stacks, see [109].

Theorem A.19 *A stack F is equivalent to a stack of the form $F^{\underline{M}}$ if and only if*

- **A1:** *The stack F is locally representable by a map $A\colon F^{\underline{N}} \to F$.*
- **A2:** *The map $\Delta\colon F \to F \times F$ is proper.*
- **DM:** *The map A is étale.*
- **R1:** *The stack has trivial automorphisms.*
- **R2:** *The map $\Delta\colon F \to F \times F$ is a closed embedding.*

We now come to the definition of a Deligne–Mumford stack, and the conclusion of this appendix.

Definition A.20 An Artin stack is a stack that satisfies **A1** and **A2**. A Deligne–Mumford stack is an Artin stack that satisfies **DM** (usually assumed to be over the category **SCHEME**). An orbifold is a Deligne–Mumford stack over the category **DIFF**.

Exercise A.21 Prove that the example of the orbifold with underlying space S^2 and two non trivial stabilizers $\mathbb{Z}_3$ and $\mathbb{Z}_2$ really is an orbifold. Analyze the automorphisms of this stack.

Appendix B Graph contributions to N_2

In this appendix we list the contributions of all of the fixed point components to the localization computation of N_2. We label graphs according to the conventions depicted in Figure 7.2. The first contribution is:

$$I(01) = -32(\alpha_0-\alpha_1)^{-2}(\alpha_0-\alpha_2)^4(\alpha_1-\alpha_2)^1(\alpha_0+\alpha_1-2\alpha_2)^{-1}\,.$$

The contribution from $I(02)$ can be obtained by exchanging 1 and 2 in the above expression. The next contribution is:

$$II(010) = 8(\alpha_0-\alpha_1)^{-2}(\alpha_0-\alpha_2)^3(\alpha_1-\alpha_2)^{-1}\,.$$

The next is:

$$II(101) = 8(\alpha_0-\alpha_1)^2(\alpha_0-\alpha_2)^4(\alpha_1-\alpha_2)^{-2}\,.$$

The contributions from $II(020)$ and $II(202)$ can be obtained by exchanging 1 and 2 in the above expressions. The next contribution is:

$$II(012) = (\alpha_0-\alpha_1)^{-1}(\alpha_0-\alpha_2)^3(\alpha_1-\alpha_2)^1(2\alpha_1-\alpha_0-\alpha_2)^{-1}\,.$$

The final contribution is:

$$II(102) = -(\alpha_0-\alpha_1)^1(\alpha_0-\alpha_2)^{-1}(\alpha_1-\alpha_2)^{-2}(2\alpha_0-\alpha_1-\alpha_2)^4\,.$$

A tedious simplification gives

$$N_2 = I(01)+I(02)+II(010)+II(020)+II(101)+II(202)+II(012)+II(102) = 1\,.$$

This is a remarkable check of the localization formula.

Appendix C Quantum invariants from skein theory

When Jones introduced his polynomial invariant, he was motivated by representations of the braid group arising from operator algebras. Shortly thereafter the theory of quantum groups started taking off, and Reshetikhin and Turaev started work on defining link and 3–manifold invariants based on quantum groups. The papers of Witten provided additional inspiration and helped them devise their invariants. Since link invariants only have to be invariant under Reidemeister but not Kirby moves, a modular structure is not required and one can get by using ribbon categories that are much easier to construct. For instance, type I representations of $U_q(\mathfrak{sl}_N\mathbb{C})$ with q a free variable rather than a fixed number form a ribbon category. We denote the resulting colored link invariants by $W_\Lambda^{\mathfrak{sl}_N}$, where Λ is a collection of representations, one for each component of the link. We should note that mathematicians usually label these invariants by the Lie algebra as we do while physicists label the invariants by the corresponding Lie group. There is an easier way suggested by H Wenzl [158] to get the same invariant using skein relations. Even though this definition is easy to understand from first principles it is next to impossible to compute with. In this appendix we first review Wenzl's elementary construction and then indicate how the same invariant can be obtained from quantum groups.

The starting point for Wenzl's definition is the THOMFLYP polynomial. (This generalization of the Jones polynomial was simultaneously discovered by several different authors and described in a joint paper [60]. The first acronym was HOMFLY and this is still in common use. It has since been realized that two other authors also deserve credit,

so some people append PT to get HOMFLYPT. We prefer to use the pseudonym.) This polynomial invariant of links is defined by the recurrence relation

$$\lambda^{\frac{1}{2}}\mathcal{P}(L_+)-\lambda^{-\frac{1}{2}}\mathcal{P}(L_-)=(\mathfrak{q}^{\frac{1}{2}}-\mathfrak{q}^{-\frac{1}{2}})\mathcal{P}(L_0),$$

together with the normalization

$$\mathcal{P}(\text{unknot})=\frac{\lambda^{\frac{1}{2}}-\lambda^{-\frac{1}{2}}}{\mathfrak{q}^{\frac{1}{2}}-\mathfrak{q}^{-\frac{1}{2}}},$$

where $L_\pm$ and L_0 are three links that differ exactly in the neighborhood of one crossing as in Figure C.1. The THOMFLYP polynomial is not in fact a polynomial – rather

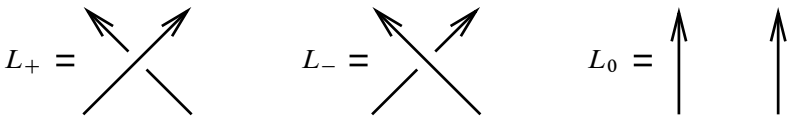


Figure C.1: Terms in the skein relation

it is a rational function of the formal variables $\mathfrak{q}^{\frac{1}{2}}$ and $\lambda^{\frac{1}{2}}$. The Jones polynomial is recovered by substituting $\lambda=\mathfrak{q}^2$ into the THOMFLYP polynomial.

Working backwards from Definition 17.68, it is clear that the framed link invariant should be defined by

$$W^{\mathfrak{sl}_N}_{\square,\dots,\square}(L)=\epsilon^{(N-1/N)\sum_{i=1}^{c}\sum_{j=1}^{c}n_{ij}(L)}\mathcal{P}(L)\,, \tag{91}$$

where $n_{ij}(L)$ are the entries of the linking matrix. By inserting a right-handed twist (the configuration labeled by θ_V in Figure 16.3) into an arbitrary link diagram to get L_+ and a left-handed twist to get L_-, one can compute that

$$\mathcal{P}(L0)=(\mathfrak{q}^{1/2}-\mathfrak{q}^{-1/2})(\lambda^{1/2}-\lambda^{-1/2})^{-1}\mathcal{P}(L)$$

where $L0$ is obtained from L by adjoining a completely unlinked and unknotted component. The physics literature often uses a slightly different normalization. Namely,

$$W^{\mathrm{SU}(N)}_{\square,\dots,\square}(L)=\lambda^{\sum_{i=1}^{c}\sum_{j=1}^{c}n_{ij}(L)}\mathcal{P}(L)\,,$$

Exercise C.1 Assume zero framings (self-linking numbers). Compute

$$W^{\mathrm{SU}(N)}_{\square,\square}(\text{left Hopf link})=\left(\frac{\lambda^{\frac{1}{2}}-\lambda^{-\frac{1}{2}}}{\mathfrak{q}^{\frac{1}{2}}-\mathfrak{q}^{-\frac{1}{2}}}\right)^2+\lambda^{-1}-1\,.$$

Compute $W^{\mathrm{SU}(N)}_{\square,\square}$(right Hopf link) and $W^{\mathrm{SU}(N)}_{\square}$(left (2,3) torus knot). (The left Hopf link is depicted in Figure 16.2)

The answer to the left $(2,3)$ torus knot may be found in Mariño [102].

There is a different invariant of framed links that may be constructed out of The THOMFLYP polynomial. It is

$$P^{\alpha}(L) := \alpha^{\sum_{i=1}^{c}\sum_{j=1}^{c} n_{ij}}\, P(L).$$

To go further, we specialize by evaluating these polynomial invariants at $\mathfrak{q}^{\frac{1}{2}} = e^{i\pi/(N+k)}$, $\lambda^{\frac{1}{2}} = e^{iN\pi/(N+k)}$. Let γ denote a multi-index of length $c(L)$, that is, a $c(L)$–tuple of positive integers. We will let $|\gamma|$ denote the sum of the components of γ, and define L^{γ} to be the framed link obtained from L by replacing the p^{th} component of L by γ_p parallel copies. Define a number by

$$\Gamma := \left(\frac{1}{N}\lim_{p\to\infty}\left(\frac{\lambda^{\frac{1}{2}}-\lambda^{-\frac{1}{2}}}{\mathfrak{q}^{\frac{1}{2}}-\mathfrak{q}^{-\frac{1}{2}}}\right)^{-p}\mathcal{P}((1\text{–framed unknot})^{p})\right)^{-1}.$$

Finally, Wenzl defines an invariant by

$$\tau_k^{SU(N)}(M,\varnothing) := N^{-1}\left(\frac{\Gamma}{|\Gamma|}\right)^{\sigma(L_M)-c(L)+1}\lim_{p\to\infty} p^{-c(L)}\Gamma^{c(L)}\sum_{\max(\gamma)\le p}\Theta^{-|\gamma|}\mathcal{P}(L_M^{\gamma}),$$

where

$$\Theta := \left(\frac{\lambda^{\frac{1}{2}}-\lambda^{-\frac{1}{2}}}{\mathfrak{q}^{\frac{1}{2}}-\mathfrak{q}^{-\frac{1}{2}}}\right).$$

The advantage of this definition is that all of the ingredients are elementary. The disadvantages are that one has to work to prove that it is well-defined, and it is not obvious how one can compute based on this definition. Notice that the 1–framed unknot squared is just the right Hopf link, and then try to compute $\mathcal{P}((1\text{–framed unknot})^3)$. This is fairly difficult. Computing the invariant for higher cables without some trick looks hopeless. Wenzl proves that this invariant is indeed well-defined and is equal to the Reshetikhin–Turaev invariant (see Wenzl [158]).

To prove that Wenzl's invariant is equal to the quantum group invariant one notices that $\mathcal{Q}_{\lambda}^{q}\otimes\mathcal{Q}_{\mu}^{q} = \sum_{\nu\in I} N_{\lambda\mu}^{\nu}\mathcal{Q}_{\nu}^{q}$, where $N_{\lambda\mu}^{\nu}$ are constants determined by the structure of a modular category. Via repeated application of this formula one may replace computations in arbitrary representations by computations for various cables in the fundamental representation and then apply the skein relation.

Appendix D Representation theory of Lie groups and Lie algebras

This appendix reviews some aspects of classical representation theory that are used in this paper. The information we summarize here can be found in the wonderful books by Fulton and Harris [62] and Humphreys [78]. We will denote the $N \times N$ unitary matrices ($A^\dagger A = I$) by $U(N)$. The special unitary matrices are those with unit determinant, this group will be denoted by $\mathrm{SU}(N)$. The Lie algebras of these two groups are the algebra of Hermitian matrices ($A^\dagger + A = 0$), denoted $\mathfrak{u}_N$, and trace-free Hermitian matrices, denoted by $\mathfrak{su}_N$, respectively with the standard bracket ($[A, B] = AB - BA$) as a product. The complexifications of these two algebras are the collection of all complex matrices ($\mathfrak{gl}_N\mathbb{C}$) and the subalgebra of trace-free matrices ($\mathfrak{sl}_N\mathbb{C}$) respectively. Let E_{ij} denote the matrix with a 1 in the j^{th} column of the i^{th} row and zeros elsewhere.

Young diagrams and irreducible representations A representation of a group (algebra) is a homomorphism into the automorphisms (endomorphisms) of a vector space. All of the groups and algebras defined above admit an obvious representation with vector space $\mathbb{C}^N$ called the fundamental or defining representation.

Remark D.1 Of course when we talk about differentiating a representation we are talking about a representation of a Lie group. Representations of groups and algebras are closely related but there is not an exact correspondence. For example, the fundamental representation of $GL_N\mathbb{C}$ can be differentiated to produce a representation of the algebra $\mathfrak{gl}_N\mathbb{C}$. The same representation can be conjugated to give a different representation of $GL_N\mathbb{C}$; however, differentiating the conjugate representation will not produce a corresponding algebra representation because it will not be complex linear. In the other direction one can see that there is a two-complex dimensional representation of the Lie algebra $\mathfrak{so}_3$ via the isomorphism $\mathfrak{su}_2$. However there is no corresponding representation of the three-dimensional rotation group $\mathrm{SO}(3)$. The classical groups that we are considering are all matrix groups, so each may be viewed as a subset of $\mathbb{C}^{N^2}$ via the fundamental representation. The automorphisms of any complex vector space can be described as a subset of complex m–space in the same way. It therefore makes sense to consider polynomial representations, that is, those that can be described by polynomials. For simply-connected matrix groups there is a perfect correspondence between finite-dimensional polynomial representations of the group and finite dimensional representations of the Lie algebra given by differentiation.

There are several obvious ways to construct new representations from (sets of) old representations: the dual, direct sum, tensor product etc. If $\rho_k\colon G \to \mathrm{Aut}(V_k)$ for

$k = 1, 2$ are group representations, then the tensor product representation $\rho_1 \otimes \rho_2 \colon G \to \mathrm{Aut}(V_1 \otimes V_2)$ is given by $(\rho_1 \otimes \rho_2)(g)(x \otimes y) = \rho_1(g)(x) \otimes \rho_2(g)(y)$. The tensor product of algebra representations $\mu_k \colon \mathfrak{g} \to \mathrm{End}(V_k)$ for $k = 1, 2$ is given by $(\mu_1 \otimes \mu_2)(A)(x \otimes y) = \mu_1(A)(x) \otimes y + x \otimes \mu_2(A)(y)$. This is obtained by differentiating the tensor product of group representations.

Additional representations can be constructed by symmetric or anti-symmetric tensors. All of the symmetries that we need can by constructed using Young symmetrizers arising from Young diagrams. A typical Young diagram is displayed on the left in Figure D.1. A Young diagram represents a partition of a positive integer, $\ell = \ell_1 + \ell_2 + \cdots + \ell_k$ with $\ell_j \geq \ell_{j+1}$. It is standard to denote such a partition by λ and we use the same notation for the Young diagrams as well. The Young diagram in Figure D.1 corresponds to the partition $6 = 4 + 2$. A Young tableau is a Young diagram filled in with natural numbers according to some rules. Young diagrams describe vector spaces of representations while Young tableau describe special basis vectors in those spaces, namely weight vectors. The Young tableau describing what we will later define as the highest weight associated to a Young diagram is obtained by filling in a diagram in a specific manner (top row with ones, second row with two's, etc) starting in the upper left corner and ending in the lower right one (see the right side of Figure D.1).

Figure D.1: Young diagram

The point is that a Young diagram encodes an endomorphism of the ℓ–fold tensor product of any vector space. Construct a specific Young tableau by filling the Young diagram with the numbers 1 through ℓ filling in rows starting in the upper left. If λ is a Young diagram, a_λ will denote the endomorphism of the tensor product that takes a tensor product of vectors to the sum of the permutated tensor products by permutations preserving the numbers in the rows of the Young tableau. Another endomorphism, b_λ, is defined analogously but with the alternating sum (according to the signs of permutations) over permutations that preserve numbers in the columns of the Young tableau. In the example from the figure,

$$b_\lambda(e_{1,2,3,4,5,6}) = e_{1,2,3,4,5,6} - e_{5,2,3,4,1,6} - e_{1,6,3,4,5,2} + e_{5,6,3,4,1,2},$$

where $e_{1,2,3,4,5,6}$ denotes $e_1 \otimes e_2 \otimes e_3 \otimes e_4 \otimes e_5 \otimes e_6$ and $\{e_k\}$ is a basis for V. The map a_λ can be computed similarly; it is a sum of 48 terms.

The Young symmetrizer is the composition of these two endomorphisms $c_\lambda = a_\lambda \circ b_\lambda$ (a sum of 192 terms in our example). The image of a Young symmetrizer c_λ applied to a tensor product of ℓ copies of the fundamental representation is a denoted by V_λ. Computing a basis for V_λ directly from the definition is a little cumbersome. We will first consider two special cases corresponding to a diagram with one row or one column.

Example D.2 For a diagram with just one row, the partition is just $\ell = \ell$. The a_λ endomorphism is just the sum over all permutations and the b_λ is the identity map. It follows that $V_{\ell=\ell}$ is just the symmetric product of ℓ copies of $\mathbb{C}^N$ ($\mathrm{Sym}^\ell \mathbb{C}^N$). Similarly, for a diagram with just one column the partition is just $\ell = 1+1+\cdots+1$. This time a_λ is trivial and b_λ is just the alternating sum over all permutations so $V_{\ell=1+1+\cdots+1}$ is just the ℓ–th exterior power ($\bigwedge^\ell \mathbb{C}^N$).

Exercise D.3 Define the natural action of any of the classical matrix groups or algebras on the set of homogeneous polynomials of degree ℓ. Show that this representation is isomorphic to $\mathrm{Sym}^\ell \mathbb{C}^N$.

Remark D.4 Whereas the ℓ–th symmetric power is nontrivial for all ℓ, the ℓ–th exterior power is trivial unless $\ell \le N$. Furthermore, the group representation on $\wedge^N \mathbb{C}^N$ is given by the determinant, so the corresponding $SU(N)$ representation is trivial unless $\ell < N$. Likewise the algebra representation on the top exterior power is given by the trace so the $\mathfrak{su}_N$ and $\mathfrak{sl}_N\mathbb{C}$ representations are trivial unless $\ell < N$. The same comments hold for any Young diagram: The $\mathrm{GL}_N\mathbb{C}$, $\mathfrak{gl}_N\mathbb{C}$ and $\mathrm{U}(N)$ representations are trivial unless the diagram has less than or equal to N rows and the $\mathrm{SL}_N\mathbb{C}$, $\mathfrak{sl}_N\mathbb{C}$ and $\mathrm{SU}(N)$ representations are trivial unless the diagram has strictly less than N rows.

When λ has just one row, the set of vectors represented by all possible ways of filling in the Young diagram with a non-decreasing sequence of integers between 1 and N inclusive is a basis for V_λ. When λ has just one column, the set of vectors represented by all possible ways of filling in the Young diagram with an increasing sequence of integers between 1 and N inclusive is a basis for V_λ. In general, V_λ has a basis consisting of all ways of filling the Young diagram with a sequence of integers between 1 and N inclusive so that the numbers do not decrease as one reads across rows and so that the numbers strictly increase as one reads down columns. This is exactly the rule specifying the Young tableau alluded to earlier.

The action of a matrix on the vector space V_λ can be described easily: the matrix acts on the tensor product of the vectors with the given indices either using the group or

algebra action as appropriate, the answer is fully expanded and written as a combination of terms in the standard non-decreasing/increasing order.

As we go further, we will concentrate on finite-dimensional representations. A finite-dimensional representation is called decomposable if it can be expressed as a nontrivial direct sum. It is indecomposable otherwise. A representation is reducible if it contains a nontrivial subrepresentation and irreducible otherwise. In all of the cases that we consider, every finite-dimensional representation will be a direct sum of irreducible representations. The irreducible representations correspond to Young diagrams. It is not immediately clear that the representations V_λ are irreducible or that every irreducible representation is of this form, but this is indeed the case. See the book by Fulton and Harris for the complete story [62].

To be specific irreducible polynomial representations of $\mathrm{GL}_N\mathbb{C}$, $\mathfrak{gl}_N\mathbb{C}$ and $\mathrm{U}(N)$ are indexed by Young diagrams (partitions) with less than or equal to N rows and irreducible polynomial representations of $\mathrm{SL}_N\mathbb{C}$, $\mathfrak{sl}_N\mathbb{C}$ and $\mathrm{SU}(N)$ are indexed by Young diagrams (partitions) with strictly less than N rows. It is not accidental that irreducible representations for the different groups and algebras here are indexed by the same sets. The fact is that every irreducible representation of $\mathrm{GL}_N\mathbb{C}$ restricts to an irreducible representation of $\mathrm{U}(N)$, which is its maximal compact subgroup and all of the latter are obtained in this manner. The same thing happens with $\mathrm{SL}_N\mathbb{C}$, $\mathrm{SU}(N)$, $\mathfrak{gl}_N\mathbb{C}$ and $\mathfrak{sl}_N\mathbb{C}$.

We can encode our discussion up to this point in a definition.

Definition D.5 Let λ be a Young diagram with rows of length $\ell_1, \dots, \ell_N$ and columns of length $m_1, \dots, m_{\ell_1}$. The associated standard tableau is obtained by filling in the diagram with the numbers one through $\ell := \ell_1 + \cdots + \ell_N$ across rows starting in the upper left. The associated tableaux are obtained by filling in the diagram with some numbers between one and N such that numbers are non-decreasing along rows and strictly increasing down columns. Let $A_\lambda = \mathfrak{S}_{\ell_1} \times \cdots \times \mathfrak{S}_{\ell_N}$ embedded as the subgroup of the permutation group $\mathfrak{S}_\ell$ preserving the rows of the standard tableau. Let $B_\lambda = \mathfrak{S}_{m_1} \times \cdots \times \mathfrak{S}_{m_{\ell_1}}$ embedded as the subgroup of the permutation group $\mathfrak{S}_\ell$ preserving the columns of the standard tableau. The associated elements of the group ring are $a_\lambda := \sum_{\sigma \in A_\lambda} \sigma$, $b_\lambda := \sum_{\sigma \in B_\lambda} (-1)^\sigma \sigma$ and the Young symmetrizer $c_\lambda = a_\lambda \circ b_\lambda$. The Specht module V_λ is the image of the tensor power $V^{\otimes \ell}$ under the natural action of the Young symmetrizer.

This is starting to get complicated. Every irreducible representation is generated by what are called weight vectors. The weight vectors in one of the V_λ representations are just the vectors represented by Young tableau.

Example D.6 For example in the $\mathfrak{sl}_4\mathbb{C}$ representation given by $V_{\ydiagram{2,2}}$ the vector obtained by filling the first row of the diagram with the sequence $1,1,2,3$ and the second row with $2,4$ is a weight vector. It can be written as $c_{\ydiagram{2,2}}(e_{1,1,2,3,2,4})$. Let's see how several different elements of $\mathfrak{sl}_4\mathbb{C}$ act on the vector from our example. The matrix E_{21} maps e_1 to e_2 and the rest of the vectors to 0. So our weight vector will get mapped to $c_{\ydiagram{2,2}}(e_{2,1,2,3,2,4}) + c_{\ydiagram{2,2}}(e_{1,2,2,3,2,4}) = c_{\ydiagram{2,2}}(e_{1,2,2,3,2,4})$. The matrix E_{14} would replace the 4 in the second row by a 1, but this is just the zero vector. For a more complicated example notice that

$$\begin{aligned} E_{21}c_{\ydiagram{2,2}}(e_{1,1,2,2,3,4}) &= c_{\ydiagram{2,2}}(e_{2,1,2,2,3,4}) + c_{\ydiagram{2,2}}(e_{1,2,2,2,3,4}) \\ &= 2c_{\ydiagram{2,2}}(e_{1,2,2,2,3,4}) + c_{\ydiagram{2,2}}(e_{1,2,2,4,2,3}) - c_{\ydiagram{2,2}}(e_{1,2,2,3,2,4})\,. \end{aligned}$$

Exercise D.7 Verify that $\dim(V_{\ydiagram{2,2}}) = 126$.

We will concentrate on Lie algebra representations for a while. We can see more about these representations once we know more about Lie algebras. The Cartan subalgebra is a maximal abelian (every bracket is zero) subalgebra. The Cartan subalgebra is usually denoted by $\mathfrak{h}$.

Exercise D.8 Show that it is possible to simultaneously diagonalize a set of Hermitian matrices exactly when the matrices in the set commute.

Example D.9 Using the previous exercise by changing the basis one can take the subset of diagonal matrices as the Cartan subalgebra. The Cartan subalgebra of $\mathfrak{gl}_N\mathbb{C}$ is generated by the matrices E_{ii}. For $\mathfrak{sl}_N\mathbb{C}$ we can take $E_{ii} - E_{i+1i+1}$ as a basis for the Cartan subalgebra.

The dimension of the Cartan subalgebra is called the rank of the group or algebra. It is denoted by r in general and is $N-1$ for $\mathfrak{sl}_N\mathbb{C}$.

Just as eigenvalues are important invariants of matrices, weights are important invariants of a representation.

Definition D.10 The weights of an arbitrary representation $\mu\colon \mathfrak{g} \to \mathrm{End}(V)$ are linear functionals $\omega\colon \mathfrak{h} \to \mathbb{C}$ that have an associated non-zero weight vector, $v_\omega \in V$, satisfying $\mu(H)v_\omega = \omega(H)v_\omega$ for every $H \in \mathfrak{h}$. The set of all weight vectors (including zero) is called the weight space and is denoted by V^ω. The set of all weights of all representations is called the weight lattice and is denoted Λ_w.

Thus we see that weights are just eigenvalues of families of operators, weight vectors are just the corresponding eigenvectors and weight spaces are just the corresponding eigenspaces. The sum of weights corresponding to the same representation may not be a weight of the same representation; however, it is a weight in the tensor product of the representation with itself. It helps to consider the weights of all possible representations together. These form a subgroup of the dual to the Cartan subalgebra $\mathfrak{h}^*$.

Exercise D.11 Prove that the set of weights is a subgroup of the dual $\mathfrak{h}^*$. Hint: think about the tensor product and dual of representations.

Exercise D.12 Verify that Young tableaux always represent weight vectors.

Example D.13 Recall the weight vector $v_\lambda = c_{\ydiagram{3,1}}(e_{1,1,2,2,3,4})$ from Example D.6. The elements of the Cartan subalgebra act as follows: $E_{11}-E_{22}$ multiplies our weight vector by zero since there are two ones and two twos, $E_{22}-E_{33}$ multiplies the vector by $1 = 2-1$, and $E_{33}-E_{44}$ multiplies the vector by $0 = 1-1$. Thus v_λ is indeed a weight vector.

It is helpful to write the weights in coordinates. Recall the standard E_{ii} basis of the Cartan subalgebra of $\mathfrak{gl}_N\mathbb{C}$. Let E^*_{ii} denote the dual basis and set

$$I^* := \sum_{i=1}^{N} E^*_{ii},,$$

then the Cartan subalgebra of $\mathfrak{sl}_N\mathbb{C}$ contains exactly the matrices annihilated by I^*. Therefore we can identify its dual $\mathfrak{h}^*_{\mathfrak{sl}_N}$ with the subspace of $\mathfrak{h}^*_{\mathfrak{gl}_N}$ that annihilates I. Thus $\mathfrak{h}^*_{\mathfrak{sl}_N}$ is generated by $E^*_{ii} - E^*_{i+1i+1}$. It is not at all obvious, but the weight lattices of $\mathfrak{gl}_N\mathbb{C}$, $\mathfrak{sl}_N\mathbb{C}$ are generated over $\mathbb{Z}$ by E^*_{ii} and

$$L_i := E^*_{ii} - \frac{1}{N} I^*$$

respectively.

Example D.14 Given this notation, the weight corresponding to the weight vector of our $c_{\ydiagram{3,1}}(e_{1,1,2,2,3,4})$ example is $E^*_{11} + E^*_{22} - \frac{2}{4} I^* = L_1 + L_2$.

Exercise D.15 Assuming that the irreducible representations are exactly V_λ show that the weight lattice of $\mathfrak{sl}_N\mathbb{C}$ is generated by $L_i := E^*_{ii} - \frac{1}{N} I^*$.

Every Lie algebra has one special representation called the adjoint representation.

Definition D.16 The adjoint representation is ad: $\mathfrak{g} \to \mathrm{End}(\mathfrak{g})$ given by $\mathrm{ad}(X)(Y) = [X, Y]$. The weights of the adjoint representation are called roots. Roots are typically denoted by α. The corresponding weight vectors are called root vectors. The set of roots is denoted by Δ and the subgroup of $\mathfrak{h}^*$ generated by the roots is called the root lattice and is denoted Λ_r.

The adjoint representation is not necessarily irreducible. If it is the algebra is called simple. For example, $\mathfrak{sl}_N\mathbb{C}$ is simple and $\mathfrak{gl}_N\mathbb{C}$ is not.

Exercise D.17 Find the Young diagram corresponding to the adjoint representation of $\mathfrak{sl}_N\mathbb{C}$. You can check your answer by computing the dimension.

It is standard to pick a vector in $\mathfrak{h}$ that is not annihilated by any non-zero weight to measure the heights of weights and roots. This vector can be chosen so that any weight evaluated on this vector gives a real number. We can take the matrix

$$\mathrm{Ht} := \pi^{N-1}E_{11} + \pi^{N-2}E_{22} + \cdots E_{NN} - \frac{\pi^N - 1}{\pi - 1}I\,,$$

where $I = \sum E_{ii}$ to define heights for $\mathfrak{sl}_N\mathbb{C}$.

Exercise D.18 The highest weight vector in our sample representation is then the one displayed in Figure D.1.

Definition D.19 The roots that evaluate to a positive number are called the positive roots. The set of positive roots is denoted by Δ^+. The positive roots that cannot be written as a sum of positive roots are called simple roots.

The negative of any root is also a root. Any Lie algebra may be written as

$$\mathfrak{g} = \mathfrak{h} \oplus \bigoplus_{\alpha\in\Delta^+} (\mathfrak{g}_{-\alpha} \oplus \mathfrak{g}_\alpha)\,,$$

where $\mathfrak{g}_{\pm\alpha}$ is the eigenspace of $\mathfrak{h}$ with eigenvalue $\pm\alpha$.

Definition D.20 A weight vector in a representation that is annihilated by $\mathfrak{g}_+ := \bigoplus_{\alpha\in\Delta^+}\mathfrak{g}_\alpha$ is called a highest weight vector. The corresponding weight is called a highest weight.

The good news is that this is all very explicit for $\mathfrak{sl}_N\mathbb{C}$.

Example D.21 The roots of $\mathfrak{sl}_N\mathbb{C}$ are just $\alpha_{ij} := E^*_{ii} - E^*_{jj}$. The positive roots are those with $i < j$, and the simple roots are those with $j = i + 1$. We use the notation $\alpha_i = E^*_{ii} - E^*_{i+1i+1}$ for simple roots. The root vectors are $e_{ij} := E_{ij}$ and $f_{ij} := E_{ji}$ The subspace $\mathfrak{sl}_N\mathbb{C}_+$ is generated by E_{ij} with $i < j$, thus the weight vector from our $c_{\ydiagram{3,1}}(e_{1,1,2,2,3,4})$ example is not a highest weight vector since it is not annihilated by E_{12}.

Remark D.22 The vector in a representation with highest weight as measured by the height Ht is a highest weight vector. In general the highest weights of a fixed representation are the weights with maximal height in each of the irreducible factors.

The notation $\mathfrak{g}_- := \bigoplus_{\alpha\in\Delta^+} \mathfrak{g}_{-\alpha}$ will also be useful. If v_λ is a highest weight vector in a representation, the space V_λ (notice the duplicate notation from our discussion with Young diagrams) that is generated by $(\mathfrak{h}\oplus\mathfrak{g}_-)v_\lambda$ will be an irreducible subrepresentation and the original representation will be the direct sum of all such. This is a very powerful idea that we should elaborate. We begin with an example.

Example D.23 For $\mathfrak{sl}_N\mathbb{C}$ the space $\mathfrak{sl}_N\mathbb{C}_-$ is spanned by the matrices E_{ij} with $i > j$. For the representation described in Example D.6 the highest weight vector $v_{\ydiagram{3,1}} := c_{\ydiagram{3,1}}(e_{1,1,1,1,2,2})$ is clearly a weight vector. The fact that it is a highest weight vector allows one to reconstruct the entire representation. In order to construct a basis for the representation one should just start multiplying the vector by the matrices E_{ij} with $i > j$. For example, one should add $E_{21}c_{\ydiagram{3,1}}(e_{1,1,1,1,2,2}) = 4c_{\ydiagram{3,1}}(e_{1,1,1,2,2,2})$ to the basis. Relations in the Lie algebra determine the action of the rest of the Lie algebra on this vector. The first thing that one can see is that it is a weight vector. This is because the bracket of any element of the Cartan subalgebra with E_{21} is a multiple of E_{21}, for example $[E_{11} - E_{22}, E_{21}] = 2E_{21}$. This implies that

$$\begin{aligned}(E_{11} - E_{22})E_{21}v_{\ydiagram{3,1}} &= [E_{11} - E_{22}, E_{21}]v_{\ydiagram{3,1}} + E_{21}(E_{11} - E_{22})v_{\ydiagram{3,1}}\\ &= 2E21v_{\ydiagram{3,1}} + 2E21v_{\ydiagram{3,1}}.\end{aligned}$$

Using the bracket $[E_{12}, E_{21}] = E_{11} - E_{22}$ one can compute $E_{12}E_{21}v_{\ydiagram{3,1}}$. By continuing in this way adding additional vectors such as $E_{32}E_{21}v_{\ydiagram{3,1}}$ and $E^2_{21}v_{\ydiagram{3,1}}$ etc. The entire representation may be reconstructed.

There are two important things to remember from this example. The first is that any finite-dimensional representation is a direct sum of weight spaces: $V = \oplus_\omega V^\omega$. The V^ω are not representations of the full Lie algebra, they are only representations of the Cartan subalgebra. This weight space decomposition completely determines the action of the Cartan subalgebra according to $\mu(H)v = \omega(H)v$. This follows from the

natural generalization of $[E_{11}-E_{22}, E_{21}] = 2E_{21}$. In fact when $E \in \mathfrak{g}_\alpha$, one has $E^n V^\omega \subseteq V^{\omega+n\alpha}$.

The second thing to remember is how the entire representation could be reconstructed from the highest weight vectors. To formalize this define the universal enveloping algebra $U\mathfrak{g}$ to be the associative unital algebra generated by $\mathfrak{g}$ subject to the relations $[X,Y] = XY - YX$. For $\mathfrak{sl}_N\mathbb{C}$ this is just the matrix algebra structure. The point is that in a general Lie algebra one cannot multiply elements; one can only take brackets. The universal enveloping algebra allows one to multiply elements. Let $U_\pm$ be the subalgebras generated by $\mathfrak{h} \oplus \mathfrak{g}_\pm$ respectively. Specifying a weight vector v_λ with weight λ specifies a U_+–module structure on $\mathbb{C}$ thought of as the vector space generated by v_λ. The formal way to extend this to a representation of the universal enveloping algebra (and thus to the Lie algebra) is as the $U\mathfrak{g}$–module $U\mathfrak{g} \otimes_{U_+} \mathbb{C}$.

Remark D.24 We can summarize these two points by stating that every irreducible representation is generated by words in U_- multiplied by a highest weight vector and the weight space decomposition can be used to determine which words act trivially (see Example 17.14 and the discussion after it).

Summarizing the above discussion we get the following description of the irreducible representations in the cases of most interest to us.

> Irreducible complex representations of $\mathfrak{gl}_N\mathbb{C}$ are indexed by Young diagrams (partitions) with less than or equal to N rows and irreducible complex representations of $\mathfrak{sl}_N\mathbb{C}$ are indexed by Young diagrams (partitions) with strictly less than N rows.

In the main body of the paper we mention that general finite-dimensional representations of quantum groups neither decompose into direct sums of weight spaces nor contain highest weight vectors. There is however a class of sub-representations called tilting modules that do possess these desirable properties.

Following the ideas outlined above and in Example D.23 leads to a classification of all finite-dimensional representations of any Lie algebra, in particular $\mathfrak{gl}_N\mathbb{C}$, $\mathfrak{sl}_N\mathbb{C}$ (see Fulton and Harris [62]).

Exercise D.25 Prove that every finite-dimensional representation can be decomposed as a sum of the V_λ uniquely.

Dominant weights and the Weyl character formula As nice as the above description might be it is not convenient for generalization to quantum groups. We now briefly review a different (but closely related) approach due to E Cartan that uses dominant weights instead of Young diagrams.

Any Lie algebra inherits a bilinear form according to

$$\langle X, Y\rangle = \text{constant}\cdot \operatorname{Tr}(\operatorname{ad}(X)\circ \operatorname{ad}(Y))\,.$$

This pairing is called the Killing form. A Lie algebra is called semisimple exactly when this form is nondegenerate.

Exercise D.26 Prove that every simple algebra, that is, one with irreducible adjoint representation, is semisimple.

There are two standard ways to normalize it. It induces a form on $\mathfrak{h}^*$ and one can require that $\langle\alpha,\alpha\rangle = 2$ for short roots or for long roots. In the case of $\mathfrak{sl}_N\mathbb{C}$ all roots have the same length, so there is a standard interpretation of the Killing form. In fact it is just given by $\langle A, B\rangle = \operatorname{Tr}(AB^\dagger)$. Using the Killing form we may identify the Cartan subalgebra with its dual. Under this identification, the coroots are defined by $\alpha^\vee = 2\langle\alpha,\alpha\rangle^{-1}\alpha$ (so in $\mathfrak{sl}_N\mathbb{C}$ there is no difference between the coroots and the roots.) The coroot lattice $\Lambda_r^\vee$ is the lattice generated by the coroots. A coroot $\alpha^\vee$ is simple if the corresponding root α is simple. If α_i are the positive simple roots then the basis biorthogonal to $\alpha_i^\vee$ is denoted by ω_i (that is, $(\alpha_i^\vee,\omega_j)=\delta_{ij}$) and its elements are called the fundamental weights. The sum of fundamental weights $\rho := \sum_i \omega_i$ also plays an important role and is called the Weyl weight.

Exercise D.27 Show that the fundamental weights for $\mathfrak{sl}_N\mathbb{C}$ are given by $\omega_k = \sum_{i=1}^k L_i = \sum_{i=1}^k E_{ii}^* - \frac{k}{N}I^*$. That is, $\langle\alpha_i^\vee,\omega_j\rangle = \delta_{ij}$. Also show that the Weyl weight can be written as $\rho := \frac{1}{2}\sum_{\alpha\in\Delta^+}\alpha$, which is $\sum_{i=1}^{N-1}(N-i)L_i$ for $\mathfrak{sl}_N\mathbb{C}$.

Exercise D.28 Show that L_k are (all) the weights of the defining representation of $\mathfrak{sl}_N\mathbb{C}$.

We denote the highest root by θ. The dual Coxeter number of a Lie algebra is $h^\vee = \langle\rho,\theta\rangle + 1$. For $\mathfrak{sl}_N\mathbb{C}$ the highest root is $E_{11}^* - E_{NN}^*$ and the dual Coxeter number is just N. In the main body of the paper we work only with $\mathfrak{sl}_N\mathbb{C}$. Thus, we often assume that the Cartan subalgebra is identified with its dual and drop * and $^\vee$ from notation.

Exercise D.29 Show that the Weyl weight is given by

$$\rho = \frac{1}{2}\sum_{k=1}^{N}(N+1-2k)\,E^*_{kk} = \frac{1}{2}\sum_{k=1}^{N}(N+1-2k)\,L_k$$

for $\mathfrak{sl}_N\mathbb{C}$.

The Weyl group is the group generated by reflections in hyperplanes perpendicular to the roots. The reflection corresponding to a simple root is $s_i(\beta) = \beta - 2\frac{\langle\beta,\alpha_i\rangle}{\langle\alpha_i,\alpha_i\rangle}\alpha_i$. For $\mathfrak{sl}_N\mathbb{C}$ the Weyl group is the permutation group $\mathfrak{S}_N$. The reflections s_i act on the E^*_{kk} via the permutation $(i\ i+1)$.

It is time to establish a correspondence between the positive weights of $\mathfrak{sl}_N\mathbb{C}$ and partitions (Young diagrams). Note that the matrices $L_i := E^*_{ii} - \frac{1}{N}I^*$, $1 \le i \le N-1$ may also serve as a basis of Λ_w different from the basis of fundamental weights ω_i. These are related by, $\omega_i = \sum_{j=1}^{i} L_j$. Given a highest weight λ we have

$$\lambda = \sum_{i=1}^{N-1} n_i\omega_i = \sum_{i=1}^{N-1} n_i \sum_{j=1}^{i} L_j = \sum_{j=1}^{N-1}\Big(\sum_{i=j}^{N-1} n_i\Big)L_j = \sum_{j=1}^{N-1}\lambda_j L_j.$$

Since the n_i are non-negative numbers we have $\lambda_j \ge \lambda_{j+1}$, in other words the vector $(\lambda_1,\dots,\lambda_{N-1})$ is a partition. This is the partition that corresponds to the highest weight λ and we abuse notation by using the same letter for the weight, partition and Young diagram. It is possible to construct a Young diagram from a sum of fundamental weights directly. Namely, if $\lambda = \sum_{i=1}^{N-1} n_i\omega_i$ then the diagram has n_{N-1} columns with $N-1$ boxes, n_{N-2} columns with $N-2$ boxes, etc. Since the λ_i are the numbers of boxes in the rows this provides a simple graphical method of converting sums of fundamental weights into partitions and vice versa. Since we already know that Young diagrams index irreducible complex representations of $\mathfrak{sl}_N\mathbb{C}$ it makes sense to distinguish their counterparts among weights.

Definition D.30 A dominant weight is a linear combination of fundamental weights with non-negative integer coefficients. If $\lambda = \sum_{i=1}^{N-1} n_i\omega_i$ is a dominant weight we denote $\ell(\lambda) := \max\{i \,|\, n_i > 0\}$ the length of λ and $|\lambda| := \sum_{i=1}^{N-1} i\,n_i$ the volume of λ. The set of all dominant weights is denoted Λ_w^+.

Exercise D.31 Show that the Young diagram corresponding to λ has $\ell(\lambda)$ rows and $|\lambda|$ boxes. For every non-negative weight find a vector $H \in \mathfrak{h}$ such that $|\lambda| = (\lambda, H)$. Hint: express H as a sum of coroots and use the biorthogonality relation. In case you are wondering: no, $H \ne Ht$.

The language of dominant weights allows one to characterize irreducible representations for all semisimple Lie algebras

> Any dominant weight is the highest weight of some irreducible representation and any highest weight is dominant. Irreducible complex representations of a complex semisimple Lie algebra, for example $\mathfrak{sl}_N\mathbb{C}$ are indexed by dominant weights.

We now elaborate on this a bit.

Definition D.32 If λ is a dominant weight we let $\mathcal{I}_\lambda$ be the left ideal of $U(\mathfrak{sl}_N\mathbb{C})$ generated by e_i and $\alpha_i^\vee - \lambda(\alpha_i^\vee)$. The classical Verma module associated to λ is

$$\widehat{V}_\lambda := U(\mathfrak{sl}_N\mathbb{C})/\mathcal{I}_\lambda\,.$$

The irreducible representation V_λ that we constructed using the Young diagram λ is isomorphic to the maximal irreducible quotient of $\widehat{V}_\lambda$. In particular the maximal irreducible quotient is finite dimensional. Also every finite dimensional representation decomposes into a direct sum of irreducible representations. In addition the irreducible representations are simple. One nice proof of all of these facts uses the Weyl unitary trick. This goes as follows.

Any finite dimensional representation of $\mathfrak{sl}_N\mathbb{C}$ also gives a representation of $\mathfrak{su}_N$ by restriction. This in turn induces a representation of $SU(N)$ by exponentiation. It follows from averaging over $SU(N)$ that any finite dimensional representation of $\mathfrak{sl}_N\mathbb{C}$ admits an invariant inner product. This quickly implies that any representation is a sum of simple representations and that every representation has a weight. By repeated action of the f_i on any weight vector one can find a highest weight vector. This is similar to the computation in example D.23. The same computation shows that any two irreducible representations with the same highest weight are isomorphic.

Exercise D.33 Write out details for the facts outlined in the previous paragraph. Compare with example Example 17.14 and exercise Exercise 17.15 from the main text.

It is instructive to draw a picture of the weight lattice Λ_w and the subset of dominant weights inside of the dual to the Cartan subalgebra. For $\mathfrak{sl}_N\mathbb{C}$ this dual can be canonically identified with the subalgebra itself via the Killing form. Under this identification the lattice in the background of Figure 18.3 is the weight lattice for $\mathfrak{sl}_3\mathbb{C}$.

Exercise D.34 Make a larger picture of the weight lattice for $\mathfrak{sl}_3\mathbb{C}$. Given that the lattice points at $2/\sqrt{3}$ and $(2/\sqrt{3})e^{\pi i/3}$ (viewing the picture in the complex plane) correspond to the fundamental weights ω_1 and ω_2 respectively, plot L_i, the roots, simple roots, Weyl vector and the set of dominant weights your picture.

The final thing we need from the representation theory is the Weyl character formula. Strangely enough, we need it for Lie group rather than Lie algebra representations. This is because the form of quantum groups (more precisely, quantized enveloping algebras) that we are using is 'partially integrated' and analogs of elements in the Cartan subalgebra of $\mathfrak{sl}_N\mathbb{C}$ belong to $\mathrm{SU}(N)$ rather than $\mathfrak{sl}_N\mathbb{C}$. Of course, complex representation spaces of $\mathrm{SU}(N)$ and $\mathfrak{sl}_N\mathbb{C}$ can always be identified since $\mathfrak{sl}_N\mathbb{C}$ is the complexified Lie algebra of $\mathrm{SU}(N)$ (see Fulton and Harris [62]). The difference is only in the operators whose traces are taken for characters.

We sketch a proof of the Weyl character formula here, but want to mention that there is a different proof presented in the book by Simon [143].

Let $\lambda\colon G \to \mathrm{Aut}(\mathbb{C}^N)$ be a group representation then its character is defined to be

$$\chi_\lambda(g) := \mathrm{tr}(\lambda(g)).$$

A function $f\colon G \to \mathbb{C}$ is called a class function if it is constant on conjugacy classes, that is, $f(gxg^{-1}) = f(x)$ for all x and g. Every character is a class function due to the cyclic property of traces.

Exercise D.35 Verify the following basic facts about characters of unitary representations: $\chi_{V\oplus W} = \chi_V + \chi_W$, $\chi_{V^*} = \chi_V^*$ and $\chi_{V\otimes W} = \chi_V\chi_W$.

We will need to use inner products and orthogonality relations for characters. Let G be a compact group and dg be the Haar (normalized bi-invariant) measure on it. For two continuous complex-valued functions define the inner product

$$\langle\psi,\varphi\rangle := \int_G \psi(g)\bar{\varphi}(g)\,dg\,,$$

where $\bar{}$ stands for complex conjugation. We reserve $(\ ,\ )$ for the inner product in $\mathrm{U}(N)$. In the case of the symmetric group dg is just the normalized counting measure and

$$\langle\psi,\varphi\rangle := \frac{1}{N!}\sum_{\sigma\in\mathfrak{S}_N} \psi(\sigma)\bar{\varphi}(\sigma)\,.$$

The symmetric group $\mathfrak{S}_N$ can be treated as a subgroup of $\mathrm{U}(N)$ by identifying permutations with permutation matrices. Thus any function on $\mathrm{U}(N)$ restricts to the symmetric group.

Given two representations V and W one can define a third as $\mathrm{Hom}(V,W)$ with action $(Af)(x) := A(f(A^*x))$. Let $\mathrm{Hom}(V,W)^{\mathrm{U}(N)}$ be the linear subspace fixed by the action of $\mathrm{U}(N)$.

Exercise D.36 Check that the transformation of $\mathrm{Hom}(V,W)^{\mathrm{U}(N)}$ defined by $\Psi(f) = \sum_{\sigma\in\mathfrak{S}_N} \sigma f/N!$ is the identity transformation. Notice that when V and W are irreducible representations the dimension of $\mathrm{Hom}(V,W)^{\mathrm{U}(N)}$ is 1 if they are isomorphic and zero otherwise. By taking the trace of Ψ conclude that $\langle \chi_\lambda, \chi_\mu\rangle = \delta_{\lambda\mu}$.

Exercise D.37 By constructing a continuous analog of Ψ from the previous problem prove that $(\chi_\lambda, \chi_\mu) = \delta_{\lambda\mu}$.

Remark D.38 It is a fact that the characters form a complete orthonormal basis for the L^2 class functions on a compact Lie group. It follows that the two norms are in fact the same.

We will now describe the irreducible characters of $\mathrm{U}(N)$ and derive the Weyl character formula. Since every unitary matrix is conjugate to a diagonal one it suffices to define any class function just on the latter. Of course we have to make sure that if two diagonal matrices are conjugate to each other our function takes the same value on both. Two diagonal matrices are conjugate in $\mathrm{U}(N)$ if and only if their diagonal entries differ by a permutation. Hence any class function is symmetric in the eigenvalues of matrices and conversely, any symmetric function can be extended by conjugation to a class function on the entire group.

This means that we can describe characters as symmetric functions on the eigenvalues of matrices. It is also useful to consider alternating functions, that is, those that satisfy $\omega(x_{\sigma(1)},\dots,x_{\sigma(N)}) = (-1)^\sigma \omega(x_1,\dots,x_N)$ for any permutation σ. One such function of particular interest to us is the Vandermonde determinant $\delta_0(x) := \det(x_j^{N-i})$ (the lower index is a coordinate and the upper index is a power). The point is that ratios (or products) of alternating functions are symmetric.

Exercise D.39 Show that $\det(x_j^{N-i}) = \prod_{i<j}(x_i - x_j)$.

Now we are ready to introduce the symmetric functions that correspond to the characters of $\mathrm{U}(N)$. Set $\delta_\lambda(x) := \det(x_j^{\lambda_i+N-i})$. Note that this is an alternating polynomial and δ_0 is the Vandermonde determinant.

Definition D.40 The ratios

$$S_\lambda(x) := \frac{\delta_\lambda(x)}{\delta_0(x)} = \frac{\det(x_j^{\lambda_i+N-i})}{\det(x_j^{N-i})}$$

are symmetric polynomials called the Schur polynomials.

These were named in honor of I Schur who discovered their connection to character theory in 1901.

Exercise D.41 Prove that these fractions are indeed polynomials.

Just as the characters of irreducible representations are orthonormal, the Schur polynomials are orthonormal:

$$\langle S_\lambda, S_\mu \rangle = \delta_{\lambda\mu}\,.$$

In fact we will see that the Schur polynomials evaluated on the eigenvalues of a matrix are equal to the corresponding characters applied to the matrix. Our proof will use the above orthogonality of the Schur polynomials. A proof of this orthogonality independent of the character formula can be found in Fulton and Harris [62].

To prove that the Schur polynomials are the characters we will need to use the complete symmetric polynomials defined by

$$H_m(x) := \sum_{|\lambda|=m} x^\lambda.$$

Alternatively, they can be described by a generating function:

Exercise D.42 Show that $\prod_{j=1}^N (1-tx_j)^{-1} = \sum_{m=0}^\infty H_m(x)t^m$.

The following formula will prove very useful for manipulating Vandermonde-type determinants.

Lemma D.43 *Let a_i $i = 1,\dots,N$ be a decreasing sequence of integers and set*

$$I_a = \{\, b \mid b_1 \geq a_1 > b_2 \geq \cdots > b_N \geq a_N \geq 0\}\,.$$

Then

$$det(x_j^{a_i}) \prod_{j=1}^N (1-x_j)^{-1} = \sum_{b\in I_a} det(x_j^{b_i})\,.$$

Proof The proof uses induction on N. The case $N = 1$ is standard. Using cofactor expansion along the first row gives

$$\begin{aligned}
&\det(x_j^{a_i}) \textstyle\prod_{j=1}^N (1-x_j)^{-1} \\
&\quad = \textstyle\sum_{k=1}^N (-1)^{k+1} x_k^{a_1} (1-x_k)^{-1} \det(x_j^{a_i})_{i\neq 1, j\neq k} \prod_{j\neq k} (1-x_j)^{-1} \\
&\quad = \textstyle\sum_{k=1}^N (-1)^{k+1} \sum_{b_1\geq a_1} \sum_{b_2\geq a_2>\cdots} \det(x_j^{b_i})_{i\neq 1, j\neq k} \\
&\quad = \textstyle\sum_{b_1\geq a_1, b_2\geq a_2>\cdots} \det(x_j^{b_i}) = \sum_{b\in I_a} \det(x_j^{b_i}).
\end{aligned}$$

The last equality follows by grouping terms with $b_1 > b_2$ with terms with $b_2 > b_1$ when $b_2 \geq a_1$. □

The next lemma employs the previous relation to prove the so-called Pieri formula.

Lemma D.44

$$S_\lambda(x) H_m(x) = \sum_{\substack{|\nu|=m \\ 0\leq \nu_i \leq \lambda_{i-1}-\lambda_i}} S_{\lambda+\nu}(x)\,.$$

Proof We have

$$\begin{aligned}
\sum_{m=0}^\infty S_\lambda(x) H_m(x) t^m &= \left(\det(x_j^{\lambda_i+N-i}) \prod_{j=1}^N (1-tx_j)^{-1} \right) / \det(x_j^{N-i}) \\
&= t^{-|\lambda|} \left(\det(tx_j^{\lambda_i+N-i}) \prod_{j=1}^N (1-tx_j)^{-1} \right) / \det(tx_j^{N-i}) \\
&= t^{-|\lambda|} \left(\sum_\nu \det(tx_j^{\lambda_i+\nu_i+N-i}) \right) / \det(tx_j^{N-i}) \\
&= \left(\sum_\nu \det(x_j^{\lambda_i+\nu_i+N-i}) t^{|\nu|} \right) / \det(x_j^{N-i}) \\
&= \sum_{0\leq \nu_i \leq \lambda_{i-1}-\lambda_i} S_{\lambda+\nu}(x) t^{|\nu|}\,.
\end{aligned}$$

Here $|\lambda| = \sum \lambda_i$, we used Lemma D.43 with $a_i = \lambda_i + N - 1$ and set $\nu_i = b_i - a_i$. The condition that $b \in I_a$ is equivalent to ν being a partition with $0 \leq \nu_i \leq \lambda_{i-1} - \lambda_i$. □

Remark D.45 The Pieri formula has a very nice graphical interpretation – the product $S_\lambda H_m$ is the sum of Schur polynomials with Young diagrams obtained from the Young

diagram of λ in all ways of adding m boxes to it with no two new boxes in the same column. We use this interpretation later for computations.

We can now prove that Schur polynomials are the characters for the representations having a Young diagram with just one row. This will provide the base of induction for the general proof.

Lemma D.46 *Let the weight $m\omega_1$ also denote the corresponding Young diagram with m boxes in one row, the corresponding partition and representation, then $S_{m\omega_1} = H_m = \chi_{m\omega_1}$.*

Proof The first equality comes from just applying the Pieri formula with $\lambda = 0$. To get the second equality note that a basis of $V_{m\omega_1}$ is formed by the vectors $c_{m\omega_1}(e_1^{\alpha_1} \otimes \cdots \otimes e_N^{\alpha_N})$ with $|\alpha| = m$. Here $c_{m\omega_1}$ is the Young symmetrizer and e_i form a basis in $V = \mathbb{C}^N$. Let $g \in \mathrm{U}(N)$ act on $\mathbb{C}^N$ with eigenvalues x_j and eigenvectors e_j; then $m\omega_1(g)c_{m\omega_1}(e_1^{\alpha_1} \otimes \cdots \otimes e_N^{\alpha_N}) = x^\alpha c_{m\omega_1}(e_1^{\alpha_1} \otimes \cdots \otimes e_N^{\alpha_N})$. It follows that

$$\chi_{m\omega_1}(g) = \mathrm{tr}(m\omega_1(g)) = \sum_{|\alpha|=m} x^\alpha = H_m(x)\,,$$

where $x^\alpha := x_1^{\alpha_1} \ldots x_N^{\alpha_N}$. □

The Weyl character formula for $\mathrm{U}(N)$ is the simple equality

$$\chi_\lambda = S_\lambda.$$

The general proof proceeds by induction on the number of boxes in the Young diagram. Lemma D.46 proves the equality for all diagrams with one row. For example

$$\chi_{\square\square\square\square} = S_{\square\square\square\square}.$$

To illustrate the proof for more rows consider the case of four boxes. Using the induction hypothesis and the Pieri formula D.44 with $\lambda = 3\square$, $m = 1$ and $\lambda = 2\square$, $m = 2$ respectively gives

$$\chi_{\square\square\square}\chi_{\square} = S_{\square\square\square\square} + S_{\begin{smallmatrix}\square\square\square\\\square\end{smallmatrix}}$$
$$\chi_{\square\square}\chi_{\square\square} = S_{\square\square\square\square} + S_{\begin{smallmatrix}\square\square\square\\\square\end{smallmatrix}} + S_{\begin{smallmatrix}\square\square\\\square\square\end{smallmatrix}}.$$

Next, using the Pieri formula D.44 with $\lambda = 2\square$ and $m = 1$, multiplying the result by $\chi_{\square} = H_1$, and then using the Pieri formula again gives the next formula. Repeating the process further gives the formula after that.

$$\chi_{\square\square}\chi_{\square}\chi_{\square} = S_{\square\square\square\square} + 2S_{\begin{smallmatrix}\square\square\square\\\square\end{smallmatrix}} + S_{\begin{smallmatrix}\square\square\\\square\square\end{smallmatrix}} + S_{\begin{smallmatrix}\square\square\\\square\\\square\end{smallmatrix}}$$
$$\chi_{\square}\chi_{\square}\chi_{\square}\chi_{\square} = S_{\square\square\square\square} + 3S_{\begin{smallmatrix}\square\square\square\\\square\end{smallmatrix}} + 2S_{\begin{smallmatrix}\square\square\\\square\square\end{smallmatrix}} + 3S_{\begin{smallmatrix}\square\square\\\square\\\square\end{smallmatrix}} + S_{\begin{smallmatrix}\square\\\square\\\square\\\square\end{smallmatrix}}\,.$$

Since a product of characters is the character of the tensor product, which in turn can be written as a sum of irreducible representations, it follows that all our character products can be written as sums of irreducible characters. For example we can write $\chi_{\square\square}\chi_{\square} = \sum_\mu n_\mu \chi_\mu$ for some numbers n_μ. Since we need these numbers for different character products we have to double index them. It is convenient to use as the second index the diagram obtained by stacking the boxes in the product one under another. For instance, the coefficients of the above sum should be denoted $n_{\begin{smallmatrix}\square\\\square\end{smallmatrix}\,\mu}$.

Notice that the representation $V_{\begin{smallmatrix}\square\\\square\end{smallmatrix}}$ is a subspace of $V_{\square\square} \otimes V_{\square}$ by definition and has to appear at least once in the irreducible decomposition. It follows that $n_{\begin{smallmatrix}\square\\\square\end{smallmatrix}\,\begin{smallmatrix}\square\\\square\end{smallmatrix}} \geq 1$. Writing the previous five displayed equations in the matrix form will make the notation and the argument even clearer.

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 1 & 0 \\ 1 & 3 & 2 & 3 & 1 \end{bmatrix} \begin{bmatrix} S_{\square\square\square\square} \\ S_{\begin{smallmatrix}\square\\\square\end{smallmatrix}\square\square} \\ S_{\square\square\square\square} \\ S_{\begin{smallmatrix}\square\\\square\end{smallmatrix}\begin{smallmatrix}\square\\\square\end{smallmatrix}} \\ S_{\begin{smallmatrix}\square\\\square\\\square\\\square\end{smallmatrix}} \end{bmatrix} = \begin{bmatrix} \sum n_{\square\square\square\square\,\mu}\chi_\mu \\ \sum n_{\begin{smallmatrix}\square\\\square\end{smallmatrix}\square\square\,\mu}\chi_\mu \\ \sum n_{\square\square\square\square\,\mu}\chi_\mu \\ \sum n_{\begin{smallmatrix}\square\\\square\end{smallmatrix}\begin{smallmatrix}\square\\\square\end{smallmatrix}\,\mu}\chi_\mu \\ \sum n_{\begin{smallmatrix}\square\\\square\\\square\\\square\end{smallmatrix}\,\mu}\chi_\mu \end{bmatrix}.$$

Inverting the matrix implies that we may write $S_\lambda = \sum m_{\lambda\mu}\chi_\mu$ for some integers $m_{\lambda\mu}$. Taking the norm gives

$$1 = \langle S_\lambda, S_\lambda \rangle = \sum m_{\lambda\mu}^2 .$$

Working from the top down we see that $m_{\lambda\lambda} \geq 1$ at one level implies that $m_{\lambda\mu} = \delta_{\lambda\mu}$ and implies that $m_{\lambda\lambda} \geq 1$ holds at the next level down. It follows that $\chi_\lambda = S_\lambda$. There is a similar matrix for the Schur functions with λ having any fixed arbitrary number of boxes.

So far we only considered irreducible characters of $\mathrm{U}(N)$ but now we can get the $\mathrm{SU}(N)$ case for free. Recall that irreducible representations of $\mathrm{SU}(N)$ are indexed by partitions of length $< N$. But we can treat such a partition as being of length $\leq N$ by setting $\lambda_N = 0$. This produces an irreducible representation of $\mathrm{U}(N)$ that restricts to the original one of $\mathrm{SU}(N)$. Since the characters are just traces the same formula gives them for $\mathrm{SU}(N)$ as long as we treat λ as an N–vector with $\lambda_N = 0$.

Theorem D.47 (Weyl character formula for $\mathrm{SU}(N)$) *Let λ denote any dominant weight (partition) of length $\leq N$ and the corresponding irreducible representation of $SU(N)$. Then*

$$\chi_\lambda = S_\lambda = \frac{\delta_\lambda}{\delta_0}$$

Exercise D.48 The coefficients $n^{\nu}_{\lambda\mu}$ defined by $\chi_\lambda\chi_\mu = \sum^{\nu}_{\lambda\mu} n^{\nu}_{\lambda\mu}\chi_\nu$ are the classical Racah coefficients. Compute $n^{\nu}_{3\omega_1\omega_1}$.

There is a version of the Weyl character formula valid for any semisimple Lie algebra. Even for $\mathrm{SU}(N)$ the formula is more useful to us in this general form. This also gives us a chance to introduce notation that comes in handy when considering quantum groups. Introduce a parameter $q = e^{iz}$, $z \in \mathbb{C}$, then the expression $q^\beta := \exp(iz\beta)$ is defined for all complex numbers and complex-valued matrices, for example by the power series. If $\beta \in \mathfrak{g}$ for some matrix Lie algebra then q^β is in the corresponding Lie group and on weight vectors of a representation its action is given by $q^\beta v_\gamma = q^{\langle\gamma,\beta\rangle} v_\gamma$. The Weyl denominator is defined to be

$$\delta_\lambda(q^\beta) := \sum_{w\in W} (-1)^w q^{\langle w(\lambda+\rho),\beta\rangle}. \tag{92}$$

Theorem D.49 (Weyl character formula) *The character χ_λ of an irreducible representation of a semisimple Lie group is given by*

$$\chi_\lambda = \frac{\delta_\lambda}{\delta_0}. \tag{93}$$

Proof for SU(N) It suffices to consider diagonal matrices $\beta = \sum_{i=1}^N \beta_i E_{ii}$ with real entries and the trace $\sum_{i=1}^N \beta_i = 0$. It is straightforward to check that $\beta_i = \langle L_i, \beta\rangle$ and $q^\beta = \mathrm{diag}(x_1,\dots,x_N) \in \mathrm{SU}(N)$, where we put $x_i := q^{\beta_i}$. The Weyl denominator becomes

$$\begin{aligned}
\delta_\lambda(q^\beta) :=& \sum\nolimits_{\sigma\in\mathfrak{S}_N} (-1)^\sigma q^{\langle\sigma(\lambda+\rho),\beta\rangle}\\
=& \sum\nolimits_{\sigma\in\mathfrak{S}_N} (-1)^\sigma q^{\sum_{i=1}^N(\lambda_i+\rho_i)\langle\sigma(L_i),\beta\rangle}\\
=& \sum\nolimits_{\sigma\in\mathfrak{S}_N} (-1)^\sigma \prod\nolimits_{i=1}^N \left(q^{\langle(L_{\sigma(i)},\beta\rangle}\right)^{\lambda_i+\rho_i}\\
=& \sum\nolimits_{\sigma\in\mathfrak{S}_N} (-1)^\sigma x_{\sigma(1)}^{\lambda_1+\rho_1}\dots x_{\sigma(N)}^{\lambda_N+\rho_N}\\
=& \det(x_j^{\lambda_i+\rho_i})\\
=& \det(x_j^{\lambda_i+N-i}).
\end{aligned}$$

The last equality is the result of Exercise D.27. Our claim now follows directly from Theorem D.47. □

As an application let us compute $\chi_\lambda(q^{2\rho})$, where as before ρ is the Weyl weight. This quantity turns up as a 'quantum dimension' in 17.35. First, for any β we have

$$\delta_0(q^{2\beta}) = \sum_{w\in\mathfrak{S}_N} (-1)^w q^{(w(\rho),2\beta)} = \det\left(q^{(L_j,2\beta)(N-i)}\right) \tag{94}$$
$$= \prod_{i<j}\left(q^{(\beta,L_i-L_j)} - q^{-(\beta,L_i-L_j)}\right) = \prod_{\alpha\in\Delta^+}(q^{(\beta,\alpha)} - q^{-(\beta,\alpha)}).$$

The third equality is the Vandermonde determinant identity from Exercise D.42. Now specifically for $q^{2\rho}$:

$$\delta_\lambda(q^{2\rho}) = \sum_{w\in\mathfrak{S}_N} (-1)^w q^{(w(\lambda+\rho),2\rho)} = \sum_{w\in\mathfrak{S}_N} (-1)^w q^{(2(\lambda+\rho),w(\rho))} \tag{95}$$
$$= \delta_0(q^{2(\lambda+\rho)}) = \prod_{\alpha\in\Delta^+}(q^{(\lambda+\rho,\alpha)} - q^{-(\lambda+\rho,\alpha)}).$$

Combining the last two formulas we prove the following.

Corollary D.50 *Let χ_λ be the character of the irreducible representation of $\mathfrak{sl}_N\mathbb{C}$ with the highest weight λ and ρ be the Weyl weight. Then*

$$\chi_\lambda(q^{2\rho}) = \frac{\delta_\lambda(q^{2\rho})}{\delta_0(q^{2\rho})} = \prod_{\alpha\in\Delta^+} \frac{q^{(\lambda+\rho,\alpha)} - q^{-(\lambda+\rho,\alpha)}}{q^{(\rho,\alpha)} - q^{-(\rho,\alpha)}}\,. \tag{96}$$

Appendix E Exact invariants from conformal field theory

There is a simple idea that leads to the mathematical definition of conformal field theory. To understand this idea, consider the loop group of a Lie group. The loop group is the space of all maps from S^1 to the group. Now consider all formal power series with values in the associated Lie algebra. An element of this 'loop algebra' produces an element of the loop group by a two step process. First consider the formal variable to be an element of S^1, so the formal power series is a map from S^1 to the Lie algebra. Second, exponentiate the answer to obtain a map from S^1 to the group. In this way problems in infinite dimensional geometry may be translated into problems in algebra.

Let $\mathbb{C}((t)) := \{\sum_{k=-M}^{\infty} a_k t^k | a_k \in \mathbb{C}\}$, and $\mathbb{C}[\![t]\!] := \{\sum_{k=0}^{\infty} a_k t^k | a_k \in \mathbb{C}\}$; then the affine Lie algebra associated to $\mathfrak{g}$ is defined to be

$$\widehat{\mathfrak{g}} := (\mathfrak{g}\otimes\mathbb{C}((t)))\oplus\mathbb{C}K,$$

with bracket

$$[A\otimes f+\alpha K, B\otimes g+\beta K]:=[A,B]\otimes fg+\langle A,B\rangle \mathrm{Res}_{t=0}(f'g)K.$$

The affine Lie algebra may be written as a direct sum of subalgebras,

$$\hat{\mathfrak{g}}=(\mathfrak{g}\otimes t\mathbb{C}[\![t]\!])\oplus\mathfrak{g}\oplus\mathbb{C}K\oplus(\mathfrak{g}\otimes t^{-1}\mathbb{C}[t^{-1}]).$$

Any representation of $\mathfrak{g}$, say V, is a $(\mathfrak{g}\otimes t\mathbb{C}[\![t]\!])\oplus\mathfrak{g}\oplus\mathbb{C}K$–module with $(\mathfrak{g}\otimes t\mathbb{C}[\![t]\!])$ acting as zero, $\mathfrak{g}$ acting as usual, and $\mathbb{C}K$ acting as $k\,\mathrm{id}$. The number k is called the level. We will always assume that the level is a positive integer. If $R\subseteq S$ and M is an R–module, then $\mathrm{Ind}_R^S M:=M\otimes_R S$ is an S–module. The Weyl module at level k with highest weight λ is defined to be $\mathrm{Ind}_{(\mathfrak{g}\otimes t\mathbb{C}[\![t]\!])\oplus\mathfrak{g}\oplus\mathbb{C}K}^{\hat{\mathfrak{g}}}V_\lambda$ and is denoted by V_λ^k. Restrict attention to $\mathfrak{sl}_N\mathbb{C}$ for simplicity. An integrable module at level k is an $\mathfrak{sl}_N\mathbb{C}$–module, so that K acts by multiplication by k and $(E_{ii}-E_{jj})\otimes t^n$ acts locally nilpotently. It turns out that the category of level k integrable modules $\mathcal{O}_k^{int}$ is a strict modular category. To define the flip $\times_{V,W}$, one uses a creative way to attach integrable modules to a Riemann surface at a number of points to obtain what is called the space of conformal blocks. This gives rise to the bundle of conformal blocks over the moduli space of marked Riemann surfaces of genus g. This bundle admits a projectively flat connection, and one can solve the parallel transport (a system of differential equations called the Knizhnik–Zamolodchikov equations) to see the effect of interchanging points marked with V and W to define the flip. This is not the obvious structure and the resulting invariants are far from the obvious ones. See Bakalov and Kirillov [21], Di Francesco, Mathieu and Sénéchal [45] and Kohno [86] for more information.

Before moving on it is worthwhile to describe the simple objects in this category. The affine Weyl group at level k is the semidirect product $W_k^a:=W\ltimes k\Lambda_r^\vee$ acting on $\mathfrak{h}^*$ by $(s,\beta^\vee)(\gamma):=s(\gamma)+\beta^\vee$, where W is the Weyl group of the associated Lie algebra. The interior of a fundamental domain for this action is given by

$$I=\{\lambda\in\mathfrak{h}^*|\langle\lambda+\rho,\alpha_i^\vee\rangle>0,\ \langle\lambda+\rho,\theta^\vee\rangle<k+h^\vee\}. \tag{97}$$

The simple objects in $\mathcal{O}_k^{int}$ are in one to one correspondence with the weights in I. Given a weight $\lambda\in I$,

$$L_\lambda^k:=V_\lambda^k:=V_\lambda^k/(U(\hat{\mathfrak{g}})(v_\theta\otimes t^{-1})^{k-\langle\lambda,\theta^\vee\rangle+1}v_{\lambda,k}.$$

Here $U(\cdot)$ denotes the universal enveloping algebra and v_λ,k is the highest weight vector of V_λ^k.

Even though many physicists approach the Witten–Chern–Simons invariants through conformal field theory, we will approach these invariants via quantum groups since it

is the fastest way to supply the definitions. It has been shown that the two approaches agree (see Tsuchiya, Ueno and Yamada [150] and Bakalov and Kirillov [21]).

References

[1] **B Acharya**, *On realising $N = 1$ super Yang–Mills in M–theory* arXiv:hep-th/0011089

[2] **V S Adamchik**, *Symbolic and numeric computations of the Barnes function*, Comput. Phys. Comm. 157 (2004) 181–190 MR2105933

[3] **M Aganagic**, **A Klemm**, **M Mariño**, **C Vafa**, *The topological vertex*, Comm. Math. Phys. 254 (2005) 425–478 MR2117633

[4] **M Aganagic**, **A Klemm**, **M Mariño**, **C Vafa**, *Matrix model as a mirror of Chern-Simons theory*, J. High Energy Phys. (2004) 010, 46 pp. MR2046799

[5] **M Aganagic**, **M Mariño**, **C Vafa**, *All loop topological string amplitudes from Chern–Simons theory*, Comm. Math. Phys. 247 (2004) 467–512 MR2063269

[6] **M Aganagic**, **H Ooguri**, **N Saulina**, **C Vafa**, *Black holes, q–deformed 2d Yang–Mills, and non-perturbative topological strings*, Nuclear Phys. B 715 (2005) 304–348 MR2135642

[7] **M Aganagic**, **C Vafa**, *G_2–manifolds, mirror symmetry and geometric engineering* arXiv:hep-th/0110171

[8] **P Aluffi** (editor), *Quantum cohomology at the Mittag–Leffler Institute*, Notes of Courses Given by Teachers at the School, Scuola Normale Superiore, Pisa (1997) MR1664633

[9] **H H Andersen**, *Tensor products of quantized tilting modules*, Comm. Math. Phys. 149 (1992) 149–159 MR1182414

[10] **H H Andersen**, **J Paradowski**, *Fusion categories arising from semisimple Lie algebras*, Comm. Math. Phys. 169 (1995) 563–588 MR1328736

[11] **G E Andrews**, **R Askey**, **R Roy**, *Special functions*, Encyclopedia of Mathematics and its Applications 71, Cambridge University Press, Cambridge (1999) MR1688958

[12] **T M Apostol**, *Modular functions and Dirichlet series in number theory*, second edition, Graduate Texts in Mathematics 41, Springer, New York (1990) MR1027834

[13] **V I Arnol'd**, *Mathematical methods of classical mechanics*, second edition, Graduate Texts in Mathematics 60, Springer, New York (1989) MR997295

[14] **M Atiyah**, *The geometry and physics of knots*, Lezioni Lincee. [Lincei Lectures], Cambridge University Press, Cambridge (1990) MR1078014

[15] **M F Atiyah**, **R Bott**, *The moment map and equivariant cohomology*, Topology 23 (1984) 1–28 MR721448

[16] **M Atiyah**, **J Maldacena**, **C Vafa**, *An M–theory flop as a large N duality*, J. Math. Phys. 42 (2001) 3209–3220 MR1840340

[17] **M Audin**, *Torus actions on symplectic manifolds*, revised edition, Progress in Mathematics 93, Birkhäuser Verlag, Basel (2004) MR2091310

[18] **M Audin**, **J Lafontaine** (editors), *Holomorphic curves in symplectic geometry*, Progress in Mathematics 117, Birkhäuser Verlag, Basel (1994) MR1274923

[19] **S Axelrod**, **S Della Pietra**, **E Witten**, *Geometric quantization of Chern–Simons gauge theory*, J. Differential Geom. 33 (1991) 787–902 MR1100212

[20] **J Baez**, **J P Muniain**, *Gauge fields, knots and gravity*, Series on Knots and Everything 4, World Scientific Publishing Co., River Edge, NJ (1994) MR1313910

[21] **B Bakalov**, **A Kirillov**, *Lectures on tensor categories and modular functors*, University Lecture Series 21, American Mathematical Society, Providence, RI (2001) MR1797619

[22] **D Bar-Natan**, *On the Vassiliev knot invariants*, Topology 34 (1995) 423–472 MR1318886

[23] **D Bar-Natan**, *Perturbative Chern-Simons theory*, J. Knot Theory Ramifications 4 (1995) 503–547 MR1361082

[24] **D Bar-Natan**, *From astrology to topology via Feynman diagrams and Lie algebras*, from: "Proceedings of the 19th Winter School "Geometry and Physics" (Srni 1999)", Rend. Circ. Mat. Palermo (2) (2000) 11–16 MR1758073

[25] **D Bar-Natan**, **S Garoufalidis**, **L Rozansky**, **D P Thurston**, *Wheels, wheeling, and the Kontsevich integral of the unknot*, Israel J. Math. 119 (2000) 217–237 MR1802655

[26] **D Bar-Natan**, **S Garoufalidis**, **L Rozansky**, **D P Thurston**, *The Århus integral of rational homology 3–spheres I. A highly non trivial flat connection on S^3*, Selecta Math. (N.S.) 8 (2002) 315–339 MR1931167

[27] **D Bar-Natan**, **R Lawrence**, *A rational surgery formula for the LMO invariant*, Israel J. Math. 140 (2004) 29–60 MR2054838

[28] **K Behrend**, *Algebraic Gromov–Witten invariants*, from: "New trends in algebraic geometry (Warwick, 1996)", London Math. Soc. Lecture Note Ser. 264, Cambridge Univ. Press, Cambridge (1999) 19–70 MR1714820

[29] **K Behrend**, *Localization and Gromov–Witten invariants*, from: "Quantum cohomology (Cetraro, 1997)", Lecture Notes in Math. 1776, Springer, Berlin (2002) 3–38 MR1911301

[30] **R Bott**, **L W Tu**, *Differential forms in algebraic topology*, Graduate Texts in Mathematics 82, Springer, New York (1982) MR658304

[31] **V Bouchard**, **B Florea**, **M Mariño**, *Counting higher genus curves with crosscaps in Calabi–Yau orientifolds* `arXiv:hep-th/0405083`

[32] **V Bouchard**, **B Florea**, **M Mariño**, *Topological open string amplitudes on orientifolds* arXiv:hep-th/0411227

[33] **G E Bredon**, *Introduction to compact transformation groups*, Pure and Appliced Mathematics 46, Academic Press, New York (1972) MR0413144

[34] **K S Brown**, *Cohomology of groups*, Graduate Texts in Mathematics 87, Springer, New York (1982) MR672956

[35] **A Bruguières**, *Catégories prémodulaires, modularisations et invariants des variétés de dimension 3*, Math. Ann. 316 (2000) 215–236 MR1741269

[36] **J Bryan**, **R Pandharipande**, *Curves in Calabi–Yau threefolds and topological quantum field theory*, Duke Math. J. 126 (2005) 369–396 MR2115262

[37] **F Cachazo**, **K Intriligator**, **C Vafa**, *A large N duality via a geometric transition*, Nuclear Phys. B 603 (2001) 3–41 MR1838691

[38] **F Cachazo**, **S Katz**, **C Vafa**, *Geometric transitions and $N = 1$ quiver theories* arXiv:hep-th/0108120

[39] **P Candelas**, **X C de la Ossa**, *Comments on conifolds*, Nuclear Phys. B 342 (1990) 246–268 MR1068113

[40] **V Chari**, **A Pressley**, *A guide to quantum groups*, Cambridge University Press, Cambridge (1994) MR1300632

[41] **C H Clemens**, *Double solids*, Adv. in Math. 47 (1983) 107–230 MR690465

[42] **H Cohen**, **L Lewin**, **D Zagier**, *A sixteenth-order polylogarithm ladder*, Experiment. Math. 1 (1992) 25–34 MR1181084

[43] **D A Cox**, **S Katz**, *Mirror symmetry and algebraic geometry*, Mathematical Surveys and Monographs 68, American Mathematical Society, Providence, RI (1999) MR1677117

[44] **P Deligne**, **P Etingof**, **D S Freed**, **L C Jeffrey**, **D Kazhdan**, **J W Morgan**, **D R Morrison**, **E Witten** (editors), *Quantum fields and strings: a course for mathematicians. Vol 1, 2*, American Mathematical Society, Providence, RI (1999) MR1701618

[45] **P Di Francesco**, **P Mathieu**, **D Sénéchal**, *Conformal field theory*, Graduate Texts in Contemporary Physics, Springer, New York (1997) MR1424041

[46] **D-E Diaconescu**, **B Florea**, **A Grassi**, *Geometric transitions, del Pezzo surfaces and open string instantons*, Adv. Theor. Math. Phys. 6 (2002) 643–702 MR1969655

[47] **D-E Diaconescu**, **B Florea**, **A Misra**, *Orientifolds, unoriented instantons and localization* arXiv:hep-th/0305021

[48] **D-E Diaconescu**, **B Florea**, **N Saulina**, *A vertex formalism for local ruled surfaces*, Comm. Math. Phys. 265 (2006) 201–226 MR2217303

[49] **L Dixon**, **J Harvey**, **C Vafa**, **E Witten**, *Strings on orbifolds*, Nuclear Phys. B 261 (1985) 678–686 MR0851703

[50] **S K Donaldson**, **R P Thomas**, *Gauge theory in higher dimensions*, from: "The geometric universe (Oxford, 1996)", Oxford Univ. Press, Oxford (1998) 31–47 MR1634503

[51] **V G Drinfel'd**, *Almost cocommutative Hopf algebras*, Algebra i Analiz 1 (1989) 30–46 MR1025154

[52] **J J Duistermaat**, **G J Heckman**, *On the variation in the cohomology of the symplectic form of the reduced phase space*, Invent. Math. 69 (1982) 259–268 MR674406

[53] **Y Eliashberg**, **A Givental**, **H Hofer**, *Introduction to symplectic field theory*, from: "GAFA 2000 (Tel Aviv, 1999)", Geom. Funct. Anal. Special Volume, Part II (2000) 560–673 MR1826267

[54] **P Etingof**, *Geometry and quantum field theory, Feynman calculus*, MIT Open Courseware, Mathematics 18.238 (2002)

[55] **C Faber**, **R Pandharipande**, *Hodge integrals and Gromov-Witten theory*, Invent. Math. 139 (2000) 173–199 MR1728879

[56] **R Fenn**, **C Rourke**, *On Kirby's calculus of links*, Topology 18 (1979) 1–15 MR528232

[57] **D Fiorenza**, **R Murri**, *Matrix integrals and Feynman diagrams in the Kontsevich model*, Adv. Theor. Math. Phys. 7 (2003) 525–576 MR2030059

[58] **O Forster**, *Lectures on Riemann surfaces*, Graduate Texts in Mathematics 81, Springer, New York (1991) MR1185074

[59] **P J Freyd**, **D N Yetter**, *Braided compact closed categories with applications to low-dimensional topology*, Adv. Math. 77 (1989) 156–182 MR1020583

[60] **P Freyd**, **D Yetter**, **J Hoste**, **W B R Lickorish**, **K Millett**, **A Ocneanu**, *A new polynomial invariant of knots and links*, Bull. Amer. Math. Soc. (N.S.) 12 (1985) 239–246 MR776477

[61] **K Fukaya**, *Morse homotopy and Chern-Simons perturbation theory*, Comm. Math. Phys. 181 (1996) 37–90 MR1410567

[62] **W Fulton**, **J Harris**, *Representation theory: a first course*, Graduate Texts in Mathematics 129, Springer, New York (1991) MR1153249

[63] **S Garoufalidis**, **M Mariño**, *On Chern–Simons matrix models* `arXiv:math.GT/0601390`

[64] **A Giveon**, **A Kehagias**, **H Partouche**, *Geometric transitions, brane dynamics and gauge theories*, J. High Energy Phys. (2001) Paper 21, 54 MR1878692

[65] **R Gopakumar**, **C Vafa**, *On the gauge theory/geometry correspondence*, Adv. Theor. Math. Phys. 3 (1999) 1415–1443 MR1796682

[66] **T Graber**, **R Pandharipande**, *Localization of virtual classes*, Invent. Math. 135 (1999) 487–518 MR1666787

[67] **A Grassi**, **M Rossi**, *Large N dualities and transitions in geometry*, from: "Geometry and physics of branes (Como, 2001)"", Ser. High Energy Phys. Cosmol. Gravit., IOP, Bristol (2003) MR1950959

[68] **P Griffiths**, **J Harris**, *Principles of algebraic geometry*, Pure and Applied Mathematics, Wiley-Interscience, New York (1978) MR507725

[69] **M Gromov**, *Pseudoholomorphic curves in symplectic manifolds*, Invent. Math. 82 (1985) 307–347 MR809718

[70] **N Halmagyi**, **T Okuda**, **V Yasnov**, *Large–N duality, lens spaces and the Chern–Simons matrix model*, J. High Energy Phys. (2004) 014, 10 pp. MR2080896

[71] **S K Hansen**, **T Takata**, *Reshetikhin–Turaev invariants of Seifert 3–manifolds for classical simple Lie algebras*, J. Knot Theory Ramifications 13 (2004) 617–668 MR2080126

[72] **G Hardy**, **E Wright**, *An introduction to the theory of numbers*, Oxford University Press (1979) MR0568909

[73] **J Harris**, *Algebraic geometry, a first course*, Graduate Texts in Mathematics 133, Springer, New York (1992) MR1182558

[74] **J Harris**, **I Morrison**, *Moduli of curves*, Graduate Texts in Mathematics 187, Springer, New York (1998) MR1631825

[75] **R Hartshorne**, *Algebraic geometry*, Graduate Texts in Mathematics 52, Springer, New York (1977) MR0463157

[76] **K Hori**, **S Katz**, **A Klemm**, **R Pandharipande**, **R Thomas**, **C Vafa**, **R Vakil**, **E Zaslow**, *Mirror symmetry*, Clay Mathematics Monographs 1, American Mathematical Society, Providence, RI (2003) MR2003030

[77] **S Hu**, *Lecture notes on Chern–Simons–Witten theory*, World Scientific Publishing Co., River Edge, NJ (2001) MR1852998

[78] **J E Humphreys**, *Introduction to Lie algebras and representation theory*, Graduate Texts in Mathematics 9, Springer, New York (1978) MR499562

[79] **A Iqbal**, *All genus topological string amplitudes and 5–brane webs as Feynman diagrams* `arXiv:hep-th/0207114`

[80] **V F R Jones**, *A polynomial invariant for knots via von Neumann algebras*, Bull. Amer. Math. Soc. (N.S.) 12 (1985) 103–111 MR766964

[81] **V G Kac**, *Infinite-dimensional Lie algebras*, third edition, Cambridge University Press, Cambridge (1990) MR1104219

[82] **C Kassel**, *Quantum groups*, Graduate Texts in Mathematics 155, Springer (1995) MR1321145

[83] **C Kassel**, **M Rosso**, **V G Turaev**, *Quantum groups and knot invariants*, Panoramas et Synthèses 5, Société Mathématique de France (1997) MR1470954

[84] **R Kirby**, *A calculus for framed links in S^3*, Invent. Math. 45 (1978) 35–56 MR0467753

[85] **A A Kirillov, Jr**, *On an inner product in modular tensor categories*, J. Amer. Math. Soc. 9 (1996) 1135–1169 MR1358983

[86] **T Kohno**, *Conformal field theory and topology*, Translations of Mathematical Monographs 210, American Mathematical Society, Providence, RI (2002) MR1905659

[87] **M Kontsevich**, *Intersection theory on the moduli space of curves and the matrix Airy function*, Comm. Math. Phys. 147 (1992) 1–23 MR1171758

[88] **M Kontsevich**, *Feynman diagrams and low-dimensional topology*, from: "First European Congress of Mathematics, Vol. II (Paris, 1992)", Progr. Math. 120, Birkhäuser, Basel (1994) 97–121 MR1341841

[89] **M Kontsevich**, *Enumeration of rational curves via torus actions*, from: "The moduli space of curves (Texel Island, 1994)", Progr. Math. 129, Birkhäuser, Boston (1995) 335–368 MR1363062

[90] **M Kontsevich**, *Vassiliev's knot invariants*, Adv. Soviet Math. 16/2, American Mathematical Society, Providence, RI (2003) 137–150 MR1237836

[91] **S Koshkin**, *Conormal bundles to knots and the Gopakumar–Vafa conjecture* `arXiv:math.DG/0503248`

[92] **J M F Labastida**, **M Mariño**, **C Vafa**, *Knots, links and branes at large N* `arXiv:hep-th/0010102`

[93] **T T Q Le**, **J Murakami**, **T Ohtsuki**, *On a universal perturbative invariant of 3–manifolds*, Topology 37 (1998) 539–574 MR1604883

[94] **J Li**, **C-C M Liu**, **K Liu**, **J Zhou**, *A mathematical theory of the topological vertex* `arXiv:math.AG/0408426`

[95] **J Li**, **Y S Song**, *Open string instantons and relative stable morphisms*, Adv. Theor. Math. Phys. 5 (2001) 67–91 MR1894338

[96] **J Li**, **G Tian**, *Virtual moduli cycles and Gromov–Witten invariants of algebraic varieties*, J. Amer. Math. Soc. 11 (1998) 119–174 MR1467172

[97] **J Li**, **G Tian**, *Virtual moduli cycles and Gromov–Witten invariants of general symplectic manifolds*, from: "Topics in symplectic 4-manifolds (Irvine, CA, 1996)", First Int. Press Lect. Ser., I, Int. Press, Cambridge, MA (1998) 47–83 MR1635695

[98] **C-C M Liu**, *Moduli of J–holomorphic curves with Lagrangian boundary conditions and open Gromov–Witten invariants for an S^1–equivariant pair*, PhD thesis `arXiv:math.SG/0210257`

[99] **G Lusztig**, *Quantum groups at roots of 1*, Geom. Dedicata 35 (1990) 89–113 MR1066560

[100] **S Mac Lane**, *Categories for the working mathematician*, second edition, Graduate Texts in Mathematics 5, Springer, New York (1998) MR1712872

[101] **S Majid**, *A quantum groups primer*, London Mathematical Society Lecture Note Series 292, Cambridge University Press, Cambridge (2002) MR1904789

[102] **M Mariño**, *Enumerative geometry and knot invariants*, from: "Infinite dimensional groups and manifolds", IRMA Lect. Math. Theor. Phys. 5, de Gruyter, Berlin MR2104354

[103] **M Mariño**, *Chern–Simons theory and topological strings*, Rev. Modern Phys. 77 (2005) 675–720 MR2168778

[104] **M Mariño**, *Chern–Simons theory, matrix integrals, and perturbative three-manifold invariants*, Comm. Math. Phys. 253 (2005) 25–49 MR2105636

[105] **M Mariño**, *Chern–Simons theory, matrix models and topological strings*, International Series of Monographs on Physics 131, The Clarendon Press Oxford University Press, Oxford (2005) MR2177747

[106] **I Maulik**, **N Nekrasov**, **A Okounkov**, **R Pandharipande**, *Gromov–Witten theory and Donaldson–Thomas theory I* `arXiv:math.AG/0312059`

[107] **I Maulik**, **N Nekrasov**, **A Okounkov**, **R Pandharipande**, *Gromov–Witten theory and Donaldson–Thomas theory II* `arXiv:math.AG/0406092`

[108] **D McDuff**, **D Salamon**, *J –holomorphic curves and symplectic topology*, American Mathematical Society Colloquium Publications 52, American Mathematical Society, Providence, RI (2004) MR2045629

[109] **D Metzler**, *Topological and smooth stacks* `arXiv:math.DG/0306176`

[110] **J W Milnor**, **J D Stasheff**, *Characteristic classes*, Annals of Mathematics Studies 76, Princeton University Press, Princeton, NJ (1974) MR0440554

[111] **M Müger**, *On the structure of modular categories*, Proc. London Math. Soc. (3) 87 (2003) 291–308 MR1990929

[112] **D Mumford**, *Geometric invariant theory*, Ergebnisse der Mathematik und ihrer Grenzgebiete 34, Springer, Berlin (1965) MR0214602

[113] **D Mumford**, *Picard groups of moduli problems*, from: "Arithmetical Algebraic Geometry (Proc. Conf. Purdue Univ., 1963)", Harper & Row, New York (1965) 33–81 MR0201443

[114] **D Mumford**, *Curves and their Jacobians*, The University of Michigan Press, Ann Arbor, Mich. (1975) MR0419430

[115] **D Mumford**, *Towards an enumerative geometry of the moduli space of curves*, from: "Arithmetic and geometry, Vol II", Progr. Math. 36, Birkhauser, Boston, MA (1983) 271–328 MR0717614

[116] **L Ng**, *Conormal bundles, contact homology and knot invariants*, from: "The interaction of finite-type and Gromov–Witten invariants (BIRS 2003)", Geom. Topol. Monogr. 8 (2006) 129–144

[117] **T Ohtsuki**, *Finite type invariants of integral homology* 3*–spheres*, J. Knot Theory Ramifications 5 (1996) 101–115 MR1373813

[118] **A Okounkov**, **R Pandharipande**, *Hodge integrals and invariants of the unknot*, Geom. Topol. 8 (2004) 675–699 MR2057777

[119] **H Ooguri**, **A Strominger**, **C Vafa**, *Black hole attractors and the topological string*, Phys. Rev. D (3) 70 (2004) 106007, 13 MR2123156

[120] **H Ooguri**, **C Vafa**, *Knot invariants and topological strings*, Nuclear Phys. B 577 (2000) 419–438 MR1765411

[121] **H Ooguri**, **C Vafa**, *Worldsheet derivation of a large* N *duality*, Nuclear Phys. B 641 (2002) 3–34 MR1928179

[122] **F Perroni**, *Orbifold cohomology of ADE–singularities* `arXiv:math.AG/0510528`

[123] **M Polyak**, *Feynman diagrams for pedestrians and mathematicians*, from: "Graphs and patterns in mathematics and theoretical physics", Proc. Sympos. Pure Math. 73, Amer. Math. Soc., Providence, RI (2005) 15–42 MR2131010

[124] **V V Prasolov**, **A B Sossinsky**, *Knots, links, braids and 3–manifolds*, Translations of Mathematical Monographs 154, American Mathematical Society, Providence, RI (1997) MR1414898

[125] **J H Przytycki**, **P Traczyk**, *Invariants of links of Conway type*, Kobe J. Math. 4 (1988) 115–139 MR945888

[126] **J G Ratcliffe**, *Foundations of hyperbolic manifolds*, Graduate Texts in Mathematics 149, Springer, New York (1994) MR1299730

[127] **M Reid**, *The moduli space of* 3*-folds with* $K = 0$ *may nevertheless be irreducible*, Math. Ann. 278 (1987) 329–334 MR909231

[128] **N Y Reshetikhin**, **V G Turaev**, *Ribbon graphs and their invariants derived from quantum groups*, Comm. Math. Phys. 127 (1990) 1–26 MR1036112

[129] **N Reshetikhin**, **V G Turaev**, *Invariants of* 3*–manifolds via link polynomials and quantum groups*, Invent. Math. 103 (1991) 547–597 MR1091619

[130] **D Rolfsen**, *Knots and links*, Mathematics Lecture Series 7, Publish or Perish, Houston, TX (1990) MR1277811

[131] **M Rosso**, *Quantum groups at a root of 1 and tangle invariants*, Internat. J. Modern Phys. B 7 (1993) 3715–3726 MR1241466

[132] **L Rozansky**, *A large* k *asymptotics of Witten's invariant of Seifert manifolds*, Comm. Math. Phys. 171 (1995) 279–322 MR1344728

[133] **D Salamon**, *Lectures on Floer homology*, from: "Symplectic geometry and topology (Park City, UT, 1997)", IAS/Park City Math. Ser. 7, Amer. Math. Soc., Providence, RI (1999) 143–229 MR1702944

[134] **H Samelson**, *Notes on Lie algebras*, second edition, Universitext, Springer, New York (1990) MR1056083

[135] **S Sawin**, *Quantum groups at roots of unity and modularity* `arXiv:math.QA/0308281`

[136] **J Sawon**, *Perturbative expansion of Chern–Simons theory*, from: "The interaction of finite-type and Gromov–Witten invariants (BIRS 2003)", Geom. Topol. Monogr. 8 (2006) 145–166

[137] **E Schrödinger**, *Statistical thermodynamics*, Cambridge University Press, New York (1962) MR0149891

[138] **G B Segal**, *The definition of conformal field theory*, from: "Differential geometrical methods in theoretical physics (Como, 1987)", NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 250, Kluwer Acad. Publ., Dordrecht (1988) 165–171 MR981378

[139] **G Segal**, *Two-dimensional conformal field theories and modular functors*, from: "IXth International Congress on Mathematical Physics (Swansea, 1988)", Hilger, Bristol (1989) 22–37 MR1033753

[140] **G Segal**, *The definition of conformal field theory*, from: "Topology, geometry and quantum field theory", London Math. Soc. Lecture Note Ser. 308, Cambridge Univ. Press, Cambridge (2004) 421–577 MR2079383

[141] **B Siebert**, *Symplectic Gromov–Witten invariants*, from: "New trends in algebraic geometry (Warwick, 1996)", London Math. Soc. Lecture Note Ser. 264, Cambridge Univ. Press, Cambridge (1999) 375–424 MR1714832

[142] **D J Simms**, **N M J Woodhouse**, *Lectures in geometric quantization*, Lecture Notes in Physics 53, Springer, Berlin (1976) MR0672639

[143] **B Simon**, *Representations of finite and compact groups*, Graduate Studies in Mathematics 10, American Mathematical Society, Providence, RI (1996) MR1363490

[144] **S Sinha**, **C Vafa**, SO *and* Sp *Chern–Simons at Large N* `arXiv:hep-th/0012136`

[145] **I Smith**, **R Thomas**, *Symplectic surgeries from singularities*, Turkish J. Math. 27 (2003) 231–250 MR1975340

[146] **J Sondow**, *Analytic continuation of Riemann's zeta function and values at negative integers via Euler's transformation of series*, Proc. Amer. Math. Soc. 120 (1994) 421–424 MR1172954

[147] **E H Spanier**, *Algebraic topology*, Springer, New York (1981) MR666554

[148] **G 't Hooft**, *A planar diagrams theory for strong interactions*, Nuclear Phys. B 75 (1974) 461–470

[149] **C H Taubes**, *Lagrangians for the Gopakumar-Vafa conjecture*, Adv. Theor. Math. Phys. 5 (2001) 139–163 MR1894340

[150] **A Tsuchiya**, **K Ueno**, **Y Yamada**, *Conformal field theory on universal family of stable curves with gauge symmetries*, from: "Integrable systems in quantum field theory and statistical mechanics", Adv. Stud. Pure Math. 19, Academic Press, Boston (1989) 459–566 MR1048605

[151] **V G Turaev**, *The Conway and Kauffman modules of a solid torus*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) 167 (1988) 79–89, 190 MR964255

[152] **V G Turaev**, *Quantum invariants of knots and 3–manifolds*, de Gruyter Studies in Mathematics 18, Walter de Gruyter & Co., Berlin (1994) MR1292673

[153] **V G Turaev**, **H Wenzl**, *Quantum invariants of* 3*–manifolds associated with classical simple Lie algebras*, Internat. J. Math. 4 (1993) 323–358 MR1217386

[154] **R Vakil**, *The moduli space of curves and Gromov–Witten theory* `arXiv:math.AG/0602347`

[155] **A Vistoli**, *Intersection theory on algebraic stacks and on their moduli spaces*, Invent. Math. 97 (1989) 613–670 MR1005008

[156] **C A Weibel**, *An introduction to homological algebra*, Cambridge Studies in Advanced Mathematics 38, Cambridge University Press, Cambridge (1994) MR1269324

[157] **A Weil**, *Elliptic functions according to Eisenstein and Kronecker*, Classics in Mathematics, Springer, Berlin (1999) MR1723749Reprint of the 1976 original

[158] **H Wenzl**, *Braids and invariants of* 3*–manifolds*, Invent. Math. 114 (1993) 235–275 MR1240638

[159] **E Witten**, *Quantum field theory and the Jones polynomial*, Comm. Math. Phys. 121 (1989) 351–399 MR990772

[160] **E Witten**, *Two-dimensional gravity and intersection theory on moduli space*, from: "Surveys in differential geometry (Cambridge, MA, 1990)", Lehigh Univ., Bethlehem, PA (1991) 243–310 MR1144529

[161] **E Witten**, *Chern–Simons gauge theory as a string theory*, from: "The Floer memorial volume", Progr. Math. 133, Birkhäuser, Basel (1995) 637–678 MR1362846

[162] **A Zee**, *Quantum field theory in a nutshell*, Princeton University Press, Princeton, NJ (2003) MR1978227

Index

Department of Mathematics, Kansas State University
Manhattan KS 66506, USA

Department of Mathematics, Northwestern University
2033 Sheridan Road, Evanston IL 60208-2730, USA

dav@math.ksu.edu, koshkin@math.northwestern.edu

http://www.math.ksu.edu/~dav/

Received: 19 May 2006