Preprint typeset in JHEP style - PAPER VERSION

Non-perturbative effects in string theory and AdS/CFT

Marcos Mariño

Département de Physique Théorique et Section de Mathématiques,
Université de Genève, Genève, CH-1211 Switzerland
marcos.marino@unige.ch

Abstract: Lecture notes.

Contents

1. A crash course on non-pertubative effects

For an introduction to non-perturbative effects, one can see [2] and references therein.

1.1 General aspects

Most series obtained through perturbative methods in quantum theory are asymptotic rather than convergent. In fact, they typically have a zero radius of convergence. This involves even very basic examples. Let us consider the quartic oscillator in QM, with Hamiltonian,

$$H = \frac{p^2}{2} + \frac{q^2}{2} + \frac{g}{4} q^4. \tag{1.1}$$

The ground state energy $E_0(g)$, as a function of g, can be calculated by using the Schrödinger equation. Stationary perturbation theory gives an asymptotic series for $E_0(g)$ around $g = 0$,

$$E_0(g) \sim \sum_{n\geq 0} a_n g^n = \frac{1}{2} + \frac{3}{4}\left(\frac{g}{4}\right) - \frac{21}{8}\left(\frac{g}{4}\right)^2 + \frac{333}{16}\left(\frac{g}{4}\right)^3 + \mathcal{O}(g^4). \tag{1.2}$$

Here, we set $\hbar = 1$. It is known that the coefficients in this series, a_n, grow factorially,

$$a_n \sim \left(\frac{3}{4}\right)^n n!, \qquad n \gg 1. \tag{1.3}$$

Of course, the ground state energy has a non-perturbative definition as a solution to the eigenvalue problem

$$H|\psi_n\rangle = E_n|\psi_n\rangle, \qquad n = 0, 1, \cdots. \tag{1.4}$$

Given an asymptotic series

$$\varphi(z) = \sum_{n\geq 0} a_n z^n \tag{1.5}$$

we say that a well-defined function $f(z)$ provides a *non-perturbative definition* of $\varphi(z)$ if $f(z)$ has $\varphi(z)$ as its asymptotic series,

$$f(z) \sim \varphi(z). \tag{1.6}$$

Clearly, such definition is far from unique, since we can add to $f(z)$ terms of the form $\mathrm{e}^{-A/z}$ without changing the asymptotic expansion. This is called a *non-perturbative ambiguity.*

In physics, very often, we have an observable $f(z)$ depending on a certain parameter z, and we only know its asymptotic expansion $\varphi(z)$. At which extent can we reconstruct $f(z)$ from $\varphi(z)$? Clearly, the existence of a non-perturbative ambiguity indicates that this problem has not a unique solution. However, before reconstructing $f(z)$ we should wonder whether we can extract concrete, numerical information from $\varphi(z)$.

A first approach to this problem is to do *optimal truncation*, i.e. to find the partial sum

$$\varphi_N(z) = \sum_{n=0}^{N} a_n z^n \tag{1.7}$$

which gives the best possible estimate of $f(z)$. To do this, one has to find the N that truncates the asymptotic expansion in an optimal way. This procedure is called *optimal truncation.*

Exercise 1.1. Show that, if

$$a_n \sim A^{-n} n!, \qquad n \gg 1. \tag{1.8}$$

the value of N which gives the optimal truncation can be estimated to be

$$N_* = \left|\frac{A}{z}\right|. \tag{1.9}$$

As an application, consider the following integral

$$I(g) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathrm{d}z \, \mathrm{e}^{-z^2/2 - gz^4/4}. \tag{1.10}$$

Show that it has an asymptotic series given by

$$\varphi(g) = \sum_{k=0}^{\infty} a_k g^k, \qquad a_k = (-4)^{-k} \frac{(4k-1)!!}{k!}. \tag{1.11}$$

Compare the exact value of $I(g)$ with the value obtained by optimal truncation of the asymptotic series, for, say $g = 0.02$. $\square$

As you can see, asymptotic series and optimal truncation are useful, but they don't lead to the correct answer. Typically, optimal truncation gives an *exponentially small error*, proportional to

$$\exp(-A/z). \tag{1.12}$$

This is the first incarnation of a non-perturbative effect. However, in order to have more precise theory of non-perturbative effects it is useful to introduce the Borel resummation of the formal power series.

The *Borel transform* of φ, which we will denote by $\widehat{\varphi}(\zeta)$, is defined as the series

$$\widehat{\varphi}(\zeta) = \sum_{n=0}^{\infty} \frac{a_n}{n!} \zeta^n. \tag{1.13}$$

Notice that, due (1.8), the series $\widehat{\varphi}(\zeta)$ has a finite radius of convergence $\rho = |A|$ and it defines an analytic function in the circle $|\zeta| < |A|$. Let us suppose that $\widehat{\varphi}(\zeta)$ has an analytic continuation to a neighbourhood of the positive real axis, in such a way that the Laplace transform

$$s(\varphi)(z) = \int_0^{\infty} \mathrm{e}^{-\zeta} \widehat{\varphi}(z\zeta) \, \mathrm{d}\zeta = z^{-1} \int_0^{\infty} \mathrm{e}^{-\zeta/z} \widehat{\varphi}(\zeta) \, \mathrm{d}\zeta, \tag{1.14}$$

exists in some region of the complex z-plane. In this case, we say that the series $\varphi(z)$ is *Borel summable* and $s(\varphi)(z)$ is called the *Borel sum* of $\varphi(z)$. Notice that, by construction, $s(\varphi)(z)$ has an asymptotic expansion around $z = 0$ which coincides with the original series $\widehat{\varphi}(\zeta)$, since

$$s(\varphi)(z) = z^{-1} \sum_{n \geq 0} \frac{a_n}{n!} \int_0^{\infty} \mathrm{d}\zeta \, \mathrm{e}^{-\zeta/z} \zeta^n = \sum_{n \geq 0} a_n z^n. \tag{1.15}$$

This procedure makes it possible in principle to reconstruct a well-defined function $s(\varphi)(z)$ from the asymptotic series $\varphi(z)$ (at least for some values of z). As we pointed out above, in some cases the formal series $\varphi(z)$ is the asymptotic expansion of a well-defined function $f(z)$ (like in the example of the quartic oscillator). It might then happen that the Borel resummation $s(\varphi)(z)$ agrees with the original function $f(z)$, and in this favorable case, the Borel resummation reconstructs the original non-pertubative answer. This happens for the quartic integral considered in the exercise above, and in the quantum-mechanical quartic oscillator.

Exercise 1.2. Show that the Borel transform of the series (1.11) is given by

$$\widehat{\varphi}(\zeta) = \frac{2K(k)}{\pi(1+4\zeta)^{1/4}}, \qquad k^2 = \frac{1}{2} - \frac{1}{2\sqrt{1+4\zeta}}. \tag{1.16}$$

where $K(k)$ is the elliptic integral of the first kind. Verify numerically that

$$s(\varphi)(g) = I(g) \tag{1.17}$$

for some values of g. □

In some cases, the Borel transform has poles on the positive real axis, and the Borel transform defined above does not exist, strictly speaking. One can then deform the contour of integration and consider contours $\mathcal{C}_{\pm}$ that avoid the singularities and branch cuts by following paths slightly above or below the positive real axis, as in Fig. 1. One defines then the *lateral Borel resummations*,

$$s_{\pm}(\varphi)(z) = z^{-1} \int_{\mathcal{C}_{\pm}} \mathrm{d}\zeta \, \mathrm{e}^{-\zeta/z} \widehat{\varphi}(\zeta). \tag{1.18}$$

In this case one gets a complex number, whose imaginary piece is also $\mathcal{O}(\exp(-A/z))$.

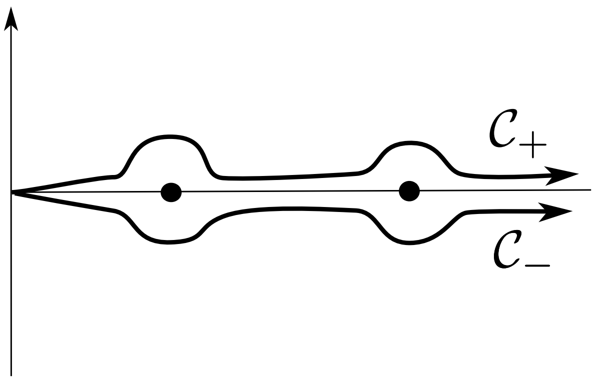

Figure 1: The paths $\mathcal{C}_\pm$ avoiding the singularities of the Borel transform from above (respectively, below).

We can ask now the following question: in cases in which the (lateral) Borel resummation of the perturbative series does not reproduce the right answer, what should we do? Clearly, something else should be added to the perturbative series! In some simple cases, the additional contributions have the form of formal power series,

$$\varphi_\ell(z) = z^{b_\ell} \mathrm{e}^{-\ell A/z} \sum_{n \geq 0} a_{n,\ell} z^n, \qquad \ell = 1, 2, \cdots, \tag{1.19}$$

which can be obtained by doing perturbation theory around *non-trivial saddle points of the (Euclidean) path integral*, like for example instantons. The series $\varphi_\ell(z)$ encodes the non-perturbative effects due to ℓ-instantons.

Example 1.3. A typical example is the double-well potential in QM, with Hamiltonian

$$H = \frac{p^2}{2} + W(x), \qquad W(q) = \frac{g}{2}\left(q^2 - \frac{1}{4g}\right)^2, \qquad g > 0. \tag{1.20}$$

In perturbation theory one finds two degenerate ground states, located around the minima

$$q_\pm = \pm \frac{1}{2\sqrt{g}}. \tag{1.21}$$

The ground state energy obtained in stationary perturbation theory is a formal power series of the form

$$\varphi_0(g) = \frac{1}{2} - g - \frac{9}{2} g^2 - \frac{89}{2} g^3 - \cdots \tag{1.22}$$

This series is obtained by doing a path integral around the constant trajectory $q = q_\pm$. However, one can consider a saddle-point of the Euclidean path integral, given by a path going from q_- to q_+ (or viceversa),

$$q_\pm^{t_0}(t) = \pm \frac{1}{2\sqrt{g}} \tanh\left(\frac{t - t_0}{2}\right). \tag{1.23}$$

This gives a non-perturbative contribution to the ground state energy of the form

$$\varphi_1(g) = -\left(\frac{2}{g}\right)^{1/2} \frac{\mathrm{e}^{-1/6g}}{\sqrt{2\pi}} \left(1 + \mathcal{O}(g)\right). \tag{1.24}$$

□

One should then consider a *trans-series* of the form

$$\Phi(z) = \varphi_0(z) + \sum_{\ell=1}^{\infty} C^{\ell} \varphi_{\ell}(z). \tag{1.25}$$

This series, after appropriate (lateral) Borel resummations and a choice of the constant C, gives sometimes the exact quantity we are looking for, i.e.

$$f(z) = s(\Phi)(z) = s(\varphi)(z) + \sum_{\ell=1}^{\infty} C^{\ell} s(\varphi_{\ell})(z). \tag{1.26}$$

1.2 Non-perturbative aspects of string theory

String theory is characterized by *two* coupling constants: the string length ℓ_s and the string coupling constant $g_{\rm st}$. The first one governs the theory at fixed genus, in an expansion around the point-particle limit. The second one governs the interactions in spacetime, obtained by joining/splitting of strings. Correspondingly, there are *two* types of exponentially small, non-perturbative effects in string theory. *Worldsheet instantons* are of the form

$$\exp\left(-A_{\rm ws}/\ell_s^2\right) \tag{1.27}$$

and they can be obtained in many cases as standard instantons in a non-linear, two-dimensional sigma model

$$x : \Sigma \to X \tag{1.28}$$

describing strings with a fixed genus propagating in a target X. In these cases, $A_{\rm ws}$ is just the area of the embedded string, measured by the metric of the target manifold X. Note that, if the target X has a typical length scale L, the dimensionless parameter controlling the expansion around the particle limit is

$$\frac{L}{\ell_s}. \tag{1.29}$$

When

$$\frac{L}{\ell_s} \gg 1, \tag{1.30}$$

the point-particle approximation (i.e. supergravity) is appropriate.

Spacetime instanton effects are typically of the form

$$\exp\left(-A_{\rm st}/g_{\rm st}\right). \tag{1.31}$$

It was pointed out by Polchinski in [3] that, in type II superstring theory, such effects can be due to D-branes.

Remark 1.4. Traditionally, exponentially small, non-perturbative effects are associated to additional (topological) *sectors* or *non-trivial saddle points* of the theory. In the cases mentioned above, worldsheet instantons are associated to topologically non-trivial sectors in the non-linear sigma model, while spacetime instantons are associated to the inclusion of extended objects in the theory. However, non-perturbative effects are not necessarily associated to saddle-points. In other words, we do not always have a semiclassical intuition for these effects. This is what happens for example with renormalons in QFT.

In these lectures we will focus on a particular quantity, namely the total partition function of a superstring/M-theory in an AdS background. Its logarithm (also known as *total free energy*) has a genus expansion of the form

$$F(\lambda, g_{\rm st}) = \sum_{g\geq 0} F_g(\lambda) g_{\rm st}^{2g-2}, \tag{1.32}$$

where λ is a function of L/ℓ_s, and L is the AdS radius. In this case, the functions $F_g(\lambda)$ have a finite radius of convergence (common to all of them). This happens in other situations, like for example in type II strings on Calabi–Yau compactifications, and it is due to the fact that the expansion in L/ℓ_s has very peculiar features due to non-renormalization theorems: the purely perturbative sector is not a series, but a polynomial in ℓ_s/L, and the expansion around each non-trivial topological sector or worldsheet instanton gets also truncated.

In contrast, the genus expansion turns out to be a divergent expansion. This is a generic property of the genus expansion in string theories [4, 1], namely, for a fixed λ (inside the radius of convergence of $F_g(\lambda)$), the $F_g(\lambda)$ grow like

$$F_g(\lambda) \sim (2g)!(A_{\rm st}(\lambda))^{-2g}, \tag{1.33}$$

where $A_{\rm st}(\lambda)$ is a spacetime instanton action. This behavior raises similar issues to what we have explained above. Is there a well-defined quantity from which the $F_g(\lambda)$ arise as coefficients in an asymptotic expansion? The issue gets complicated by the fact that this is a two-parameter model, in contrast to the one-parameter problems appearing in for example quantum mechanics and field theory.

There is however a two-parameter problem in gauge theory which has exactly the same qualitative properties than string theory: namely, gauge theories in the $1/N$ expansion. The two parameters are the gauge coupling $g_s = g^2$ and the 't Hooft parameter

$$\lambda = g_s N. \tag{1.34}$$

The total free energy of the gauge theory has a $1/N$ expansion which has exactly the same structure as (1.32) (where g_s plays the rôle of $g_{\rm st}$). In some special cases (like for example for superconformal field theories and Chern–Simons theory) the functions $F_g(\lambda)$, obtained by resumming diagrams of genus g in the double-line expansion, have a finite radius of convergence, while the sequence of free energies diverges factorially like (1.33). The corresponding non-perturbative effect in the $1/N$ expansion is called a *large N instanton*. In some cases (but not always), these effects can be obtained by considering the standard instantons of the gauge theory, and incorporating quantum planar corrections. In those cases, if we denote by A the action of an instanton in the gauge theory one has that

$$A(\lambda) \to A \tag{1.35}$$

as $\lambda \to 0$. This parallelism of structures suggests that, when the string theory has a large N dual, spacetime instanton effects due to D-branes can be identified with non-perturbative effects appearing in the large N expansion.

1.3 Non-perturbative effects in M-theory

It was noted in [40] that, in M-theory, worldsheet instantons and D-brane instantons can be unified in terms of *membrane instantons*. To understand this, let us recall that in M-theory there is one single coupling, the Planck length ℓ_p. When compactified on an eleven-dimensional circle of radius R_{11}, one obtains type IIA superstring theory, with couplings ℓ_s and $g_{\rm st}$. The dictionary is

$$\ell_p = g_{\rm st}^{1/3} \ell_s, \qquad R_{11} = g_{\rm st}\ell_s. \tag{1.36}$$

M-theory contains extended three-dimensional objects, called M2-branes or membranes. When compactified on an eleven dimensional circle, these membranes lead to fundamental strings if their world-volume wraps the compact, eleventh dimension (and therefore appear as two-dimensional objects in ten dimensions). They lead to D2 branes when they do not wrap the compact dimension.

In M-theory, we should incorporate sectors with membranes, which should be regarded as "instanton" sectors. A membrane wrapped around a three-cycle $\mathcal{S}$ leads to an exponentially small effect of the form

$$\exp\left(-\frac{{\rm vol}(\mathcal{S})}{\ell_p^3}\right). \tag{1.37}$$

If the cycle $\mathcal{S}$ wraps the eleven dimensional circle, we have (schematically) that

$$ {\rm vol}(\mathcal{S}) = R_{11}{\rm vol}(\Sigma), \tag{1.38}$$

where Σ is a cycle in ten dimensions. Then,

$$\exp\left(-\frac{{\rm vol}(\mathcal{S})}{\ell_p^3}\right) = \exp\left(-\frac{{\rm vol}(\Sigma)}{\ell_s^2}\right), \tag{1.39}$$

which is precisely the weight of a worldsheet instanton. If $\mathcal{S}$ leads to a three-dimensional cycle $\mathcal{M}$ in ten dimensions, we have

$$\exp\left(-\frac{{\rm vol}(\mathcal{S})}{\ell_p^3}\right) = \exp\left[-\frac{\left({\rm vol}(\mathcal{M})/\ell_s^3\right)}{g_{\rm st}}\right], \tag{1.40}$$

which is the expected weight for a spacetime instanton (more precisely, for a D2-brane instanton).

2. Non-perturbative effects in ABJM theory

2.1 A short review of ABJM theory

ABJM theory and its generalization, also called ABJ theory, were proposed in [9, 10] to describe N M2 branes on $\mathbb{C}^4/\mathbb{Z}_k$. They are particular examples of supersymmetric Chern–Simons–matter theories and their basic ingredient is a pair of vector multiplets with gauge groups $U(N_1)$, $U(N_2)$, described by two supersymmetric Chern–Simons theories with opposite levels k, $-k$. In addition, we have four matter supermultiplets Φ_i, $i = 1, \cdots, 4$, in the bifundamental representation of the gauge group $U(N_1)\times U(N_2)$. This theory can be represented as a quiver with two nodes, which stand for the two supersymmetric Chern–Simons theories, and four edges between the nodes representing the matter supermultiplets (see Fig. 2). In addition, there is a superpotential involving the matter fields, which after integrating out the auxiliary fields in the Chern–Simons–matter system, reads (on $\mathbb{R}^3$)

$$W = \frac{4\pi}{k}{\rm Tr}\left(\Phi_1\Phi_2^\dagger\Phi_3\Phi_4^\dagger - \Phi_1\Phi_4^\dagger\Phi_3\Phi_2^\dagger\right). \tag{2.1}$$

In this expression we have used the standard superspace notation for $\mathcal{N}=1$ supermultiplets. When the two gauge groups have identical rank, i.e. $N_1 = N_2 = N$, the theory is called

ABJM theory. The generalization in which $N_1 \neq N_2$ is called ABJ theory. More details on the construction of these theories can be found in [12, 13]. In most of this review we will focus on ABJM theory, which has two parameters: N, the common rank of the gauge group, and k, the Chern–Simons level. Note that, in this theory, all the fields are in the adjoint representation of $U(N)$ or in the bifundamental representation of $U(N) \times U(N)$. Therefore, they have two color indices and one can use the standard 't Hooft rules [14] to perform a $1/N$ expansion. Since k plays the rôle of the inverse gauge coupling $1/g^2$, the natural 't Hooft parameter is given by

$$\lambda = \frac{N}{k}. \tag{2.2}$$

One of the most important aspects of ABJM theory is that, at large N, it describes a nontrivial background of M theory, as it was already postulated in [15]. In the large distance limit in which M-theory can be described by supergravity, this is nothing but the Freund–Rubin background

$$X_{11} = \mathrm{AdS}_4 \times \mathbb{S}^7/\mathbb{Z}_k. \tag{2.3}$$

If we represent $\mathbb{S}^7$ inside $\mathbb{C}^4$ as

$$\sum_{i=1}^{4} |z_i|^2 = 1, \tag{2.4}$$

the action of $\mathbb{Z}_k$ in (2.3) is simply given by

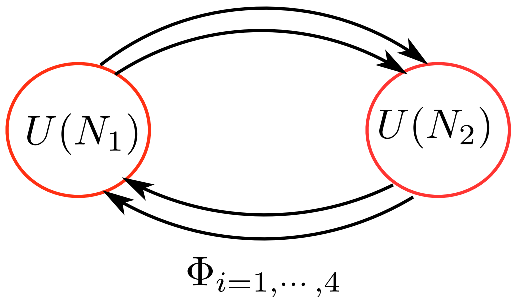

Figure 2: The quiver for ABJ(M) theory. The two nodes represent the $U(N_{1,2})$ Chern–Simons theories (with opposite levels) and the arrows between the nodes represent the four matter multiplets in the bifundamental representation.

$$z_i \to \mathrm{e}^{\frac{2\pi \mathrm{i}}{k}} z_i. \tag{2.5}$$

The metric on $\mathrm{AdS}_4 \times \mathbb{S}^7$ depends on a single parameter, the radius L, and by using metrics on AdS_4 and $\mathbb{S}^7$ of unit radius, we have

$$\mathrm{d}s^2 = L^2 \left(\frac{1}{4} \mathrm{d}s^2_{\mathrm{AdS}_4} + \mathrm{d}s^2_{\mathbb{S}^7} \right). \tag{2.6}$$

As it is well-known, the Freund–Rubin background also involves a non-zero flux for the four-form field strength G of 11d SUGRA, see for example [16] for an early review of eleven-dimensional supergravity on this background.

The AdS/CFT correspondence between ABJM theory and M-theory in the above Freund–Rubin background comes with a dictionary between the gauge theory parameters and the M-theory parameters. The parameter k in the gauge theory has a purely geometric interpretations and it is the same k appearing in the modding out by $\mathbb{Z}_k$ in (2.3) and (2.5). The parameter N corresponds to the number of M2 branes, which lead to the non-zero flux of G, and also determines the radius of the background. One finds,

$$\left(\frac{L}{\ell_p} \right)^6 = 32\pi^2 k N, \tag{2.7}$$

where ℓ_p is the eleven-dimensional Planck constant. It should be emphasized that the above relation is in principle only valid in the large N limit, and it has been argued that it is corrected due to a shift in the M2 charge [17, 18]. According to this argument, the physical charge determining the radius is not N, but rather

$$Q = N - \frac{1}{24}\left(k - \frac{1}{k}\right). \tag{2.8}$$

The geometric, M-theory description in terms of the background (2.3) emerges when

$$N \to \infty, \qquad k \text{ fixed.} \tag{2.9}$$

The corresponding regime in the dual gauge theory will be called the *M-theory regime*. In this regime, one looks for asymptotic expansions of the observables at large N but k fixed. This is the so-called *M-theory expansion* of the gauge theory.

It has been known for a while that the above Freund–Rubin background of M-theory can be used to find a background of type IIA superstring theory of the form

$$X_{10} = \mathrm{AdS}_4 \times \mathbb{CP}^3. \tag{2.10}$$

This is due to the existence of the Hopf fibration,

$$\begin{array}{ccc} \mathbb{S}^1 & \rightarrow & \mathbb{S}^7 \\ & & \downarrow \\ & & \mathbb{CP}^3 \end{array} \tag{2.11}$$

and the circle of this fibration can be used to perform a non-trivial reduction from M-theory to type IIA theory [19]. In order to have a perturbative regime for the type IIA superstring, we need the circle to be small, and this is achieved when k is large. Indeed, by using the standard dictionary relating M-theory and type IIA theory, we find that the string coupling constant $g_{\rm st}$ is given by

$$g_{\rm st}^{-2} = \frac{1}{k^2}\left(\frac{L}{\ell_s}\right)^2, \tag{2.12}$$

where ℓ_s is the string length. On the other hand, we also have from this dictionary that

$$\lambda = \frac{1}{32\pi^2}\left(\frac{L}{\ell_s}\right)^4, \tag{2.13}$$

where λ is the 't Hooft parameter (2.2). We conclude that the perturbative regime of the type IIA superstring corresponds to the 't Hooft $1/N$ expansion, in which

$$N \to \infty, \qquad \lambda = \frac{N}{k} \qquad \text{fixed,} \tag{2.14}$$

i.e. the genus expansion in the 't Hooft regime of the gauge theory corresponds to the perturbative genus expansion of the superstring. In addition, the regime of strong 't Hooft coupling corresponds to the point-particle limit of the superstring, in which α' corrections are suppressed.

A very important aspect of ABJM theory is that there are two different regimes to consider: the M-theory regime (2.9), and the standard 't Hooft regime (2.14). The existence of a well-defined M-theory limit is somewhat surprising from the gauge theory point of view. This limit is more like a thermodynamic limit of the theory, in which the number of degrees of freedom goes to infinity but the coupling constant remains fixed. General aspects of this limit have been discussed in [20].

One of the consequences of the AdS/CFT correspondence is that the partition function of the Euclidean ABJM theory on $\mathbb{S}^3$ should be equal to the partition function of the Euclidean version of M-theory/string theory on the dual AdS backgrounds [21], i.e.

$$Z(\mathbb{S}^3) = Z(X), \tag{2.15}$$

where X is the eleven-dimensional background (2.3) or the ten-dimensional background (2.10), appropriate for the M-theory regime or the 't Hooft regime, respectively. In the M-theory limit we can use the supergravity approximation to compute the M-theory partition function, which is just given by the classical action of eleven-dimensional supergravity evaluated on-shell, i.e. on the metric of (2.3). This requires a regularization of IR divergences but eventually leads to a finite result, which gives a prediction for the behavior of the partition function of ABJM theory at large N and fixed k (see [13] for a review of these isssues). If we define the free energy of the theory as the logarithm of the partition function,

$$F(N,k) = \log Z(N,k), \tag{2.16}$$

one finds, from the supergravity approximation to M-theory [22],

$$F(N,k) \approx -\frac{\pi\sqrt{2}}{3} k^{1/2} N^{3/2}, \qquad N \gg 1. \tag{2.17}$$

The $N^{3/2}$ behavior of the free energy is a famous prediction of AdS/CFT [23] for the large N behavior of a theory of M2 branes. The AdS/CFT prediction (2.15) can be also studied in the 't Hooft regime. In this case, at large N and strong 't Hooft coupling, the superstring partition function is given by the genus zero free partition function at small curvature, i.e. by the type IIA supergravity result. One obtains a prediction for the planar free energy of ABJM theory at strong 't Hooft coupling, of the form

$$-\lim_{N\to\infty} \frac{1}{N^2} F(N,\lambda) \approx \frac{\pi\sqrt{2}}{3\sqrt{\lambda}}, \qquad \lambda \gg 1. \tag{2.18}$$

Interestingly, both predictions are equivalent, in the sense that one can obtain (2.18) from (2.17) by setting $k = N/\lambda$, and viceversa. This is not completely obvious from the point of view of the gauge theory, since it could happen that higher genus corrections in the 't Hooft expansion contribute to the M-theory limit. That this is not the case has been conjectured in [20] to be a general fact and it has been called "planar dominance." It seems to be a general property of Chern–Simons–matter theories with both an M-theory expansion and a 't Hooft expansion. Note as well that, as explained in detail in [13], the behavior of the planar free energy at weak 't Hooft coupling is very different from the prediction (2.18). Therefore, the planar free energy should be a non-trivial interpolating function between the weakly coupled regime and the strongly coupled regime.

Note that, in principle, the AdS/CFT correspondence provides a non-perturbative definition of superstring theory, in the sense mentioned above. Namely, the free energy of ABJM theory on the sphere, $F(N,k)$, is a well-defined quantity, for any integer N and any integer k. In addition, it has an asymptotic expansion in the 't Hooft regime (2.14), of the form

$$F(N,k) = \sum_{g\geq 0} F_g(\lambda) g_s^{2g-2}, \tag{2.19}$$

where

$$g_s = \frac{2\pi \mathrm{i}}{k}. \tag{2.20}$$

This expansion is of course equivalent to a $1/N$ expansion, since $g_s = 2\pi \mathrm{i}\lambda/N$, and the 't Hooft parameter is kept fixed. The coefficients $F_g(\lambda)$ appearing in this expansion are, according to the AdS/CFT conjecture, type IIA superstring free energies at genus g on the AdS background (2.10).

In order to analyze in detail the implications of the large N duality between ABJM theory and M-theory on the AdS background (2.3), it is extremely useful to be able to perform reliable computations on the gauge theory side. The techniques of localization pioneered in [5] have led to a wonderful result for the partition function of ABJM theory on the three-sphere $\mathbb{S}^3$, due to [6]. This result expresses this partition function, which *a priori* is given by a complicated path integral, in terms of a matrix model (i.e. a path integral in zero dimensions). We will refer to it as the ABJM matrix model, and it takes the following form,

$$\begin{aligned} &Z(N,k) \\ &= \frac{1}{N!^2} \int \frac{\mathrm{d}^N \mu}{(2\pi)^N} \frac{\mathrm{d}^N \nu}{(2\pi)^N} \frac{\prod_{i<j} \left[2\sinh\left(\frac{\mu_i-\mu_j}{2}\right)\right]^2 \left[2\sinh\left(\frac{\nu_i-\nu_j}{2}\right)\right]^2}{\prod_{i,j}\left[2\cosh\left(\frac{\mu_i-\nu_j}{2}\right)\right]^2} \exp\left[\frac{\mathrm{i}k}{4\pi}\sum_{i=1}^N(\mu_i^2-\nu_i^2)\right]. \end{aligned} \tag{2.21}$$

In the remaining of this review, we will analyze this matrix model in detail. We will study it in different regimes and we will try to extract lessons and consequences for the AdS/CFT correspondence.

Remark 2.1. A purist might object that defining non-perturbative string theory through a gauge theory, as we do in AdS/CFT, does not really solve the problem, since one still has to define the gauge theory observables non-perturbatively (the purist might agree however that this is a more tractable problem). The matrix model representation solves this problem, at least for some observables, since the partition function $Z(N,k)$ is now manifestly well-defined for integer N and integer k (in fact, any real k).

2.2 The 't Hooft expansion

2.2.1 The planar limit

The free energy of ABJM theory on the three-sphere has a $1/N$ or 't Hooft expansion of the form (2.19). The supergravity result (2.18) already gives a prediction for the behavior of the planar free energy for large λ. In order to test this prediction, it would be useful to have an explicit expression for $F_0(\lambda)$, and eventually for the full series of genus g free energies $F_g(\lambda)$. This is in principle a formidable problem, involving the resummation of double-line diagrams with a fixed genus in the perturbative expansion of the total free energy. However, since the partition function is given by the matrix integral (2.21), we can try to obtain the $1/N$ expansion directly in the matrix model. The large N expansion of matrix models has been extensively studied since the seminal work of Brézin, Itzykson, Parisi and Zuber [24], and there are by now many different techniques to solve this problem. The first step in this calculation is of course to obtain the planar free energy $F_0(\lambda)$, which is the dominant term at large N.

A detailed review of the calculation of the planar free energy of the ABJM matrix model can be found in [13], and we won't repeat it here. We will just summarize the most important aspects of the solution. As usual, at large N, the eigenvalues of the matrix model "condense" around cuts in the complex plane. This means that the equilibrium values of the eigenvalues μ_i, ν_i, $i = 1, \cdots, N$, fall into two arcs in the complex plane as N becomes large. The equilibrium conditions for the eigenvalues μ_i, ν_i can be found immediately from the integrand of the matrix integral:

$$\begin{aligned} \frac{\mathrm{i}k}{2\pi}\mu_i &= -\sum_{j\neq i}^{N} \coth\frac{\mu_i-\mu_j}{2} + \sum_{j=1}^{N} \tanh\frac{\mu_i-\nu_j}{2}, \\ \frac{\mathrm{i}k}{2\pi}\nu_i &= \sum_{j\neq i}^{N} \coth\frac{\nu_i-\nu_j}{2} - \sum_{j=1}^{N} \tanh\frac{\nu_i-\mu_j}{2}. \end{aligned} \tag{2.22}$$

In standard matrix models, it is useful to think about the equilibrium values of the eigenvalues as the result of a competition between a confining one-body potential and a repulsive two-body potential. Here we can not do that, since the one-body potential is imaginary. One way to go around this is to use analytic continuation: we rotate k to an imaginary value, and then at the end of the calculation we rotate it back. This is the procedure followed originally in [8]. We consider then the saddle-point equations

$$\begin{aligned} \mu_i &= \frac{t_1}{N_1}\sum_{j\neq i}^{N_1} \coth\frac{\mu_i-\mu_j}{2} + \frac{t_2}{N_2}\sum_{j=1}^{N_2} \tanh\frac{\mu_i-\nu_j}{2}, \\ \nu_i &= \frac{t_2}{N_2}\sum_{j\neq i}^{N_2} \coth\frac{\nu_i-\nu_j}{2} + \frac{t_1}{N_1}\sum_{j=1}^{N_1} \tanh\frac{\nu_i-\mu_j}{2}. \end{aligned} \tag{2.23}$$

where

$$t_i = g_s N_i, \quad i = 1, 2. \tag{2.24}$$

The planar free energy obtained from these equations will be a function only of t_1 and t_2, and to recover the planar free energy of the original ABJM matrix model we have to set

$$t_1 = -t_2 = \frac{2\pi\mathrm{i}}{k}N. \tag{2.25}$$

The equations (2.23), for real g_s, are equivalent to the original ones (2.22) after rotating k to the imaginary axis, and then performing an analytic continuation $N_2 \to -N_2$. At large N_i, and for real g_s, t_i, the eigenvalues μ_i, $i = 1, \cdots, N_1$ and ν_j, $j = 1, \cdots N_2$, condense around two cuts in the real axis, $\mathcal{C}_{1,2}$ (respectively.) Due to the symmetries of the problem, these cuts are symmetric around the origin. We will denote by $[-A, A]$, $[-B, B]$, respectively.

It turns our that the equations (2.23) are the saddle-point equations for the so-called lens space matrix model studied in [25, 26, 27], whose planar solution is well-known. Let us denote by $\rho_1(\mu)$, $\rho_2(\nu)$ the large N densities of eigenvalues on the cuts $\mathcal{C}_1$, $\mathcal{C}_2$, respectively, normalized as

$$\int_{\mathcal{C}_1} \rho_1(\mu)\mathrm{d}\mu = \int_{\mathcal{C}_2} \rho_2(\nu)\mathrm{d}\nu = 1. \tag{2.26}$$

One finds that these densities are given by

$$\begin{aligned} \rho_1(X)\mathrm{d}X &= -\frac{1}{4\pi\mathrm{i}t_1}\frac{\mathrm{d}X}{X}\left(\omega_0(X+\mathrm{i}\epsilon) - \omega_0(X-\mathrm{i}\epsilon)\right), \qquad X \in \mathcal{C}_1, \\ \rho_2(Y)\mathrm{d}Y &= \frac{1}{4\pi\mathrm{i}t_2}\frac{\mathrm{d}Y}{Y}\left(\omega_0(Y+\mathrm{i}\epsilon) - \omega_0(Y-\mathrm{i}\epsilon)\right), \qquad Y \in \mathcal{C}_2, \end{aligned} \tag{2.27}$$

where the function $\omega_0(X)$, known as the planar resolvent, turns out to have the explicit expression [26, 27]

$$\omega_0(Z) = \log\left(\frac{\mathrm{e}^{-t}}{2}\left[f(Z) - \sqrt{f^2(Z) - 4\mathrm{e}^{2t}Z^2}\right]\right). \tag{2.28}$$

Notice that e^{ω_0} has a square root branch cut involving the function

$$\sigma(Z) = f^2(Z) - 4\mathrm{e}^{2t}Z^2 = (Z-a)\,(Z-1/a)\,(Z+b)\,(Z+1/b) \tag{2.29}$$

where $a^{\pm 1}$, -$b^{\pm 1}$ are the endpoints of the cuts in the $Z = \mathrm{e}^z$ plane (i.e. $A = \log a$, $B = \log b$). They are determined, in terms of the parameters t_1, t_2 by the normalization conditions for the densities (2.26). We will state the final results in ABJM theory. A detailed derivation can be found in the original papers [7, 8] and in the review [13].

In the ABJM case we have to consider the special case or "slice" given in (2.25), therefore $t = 0$. One can parametrize the endpoints of the cut in terms of a single parameter κ, as

$$a + \frac{1}{a} = 2 + \mathrm{i}\kappa, \qquad b + \frac{1}{b} = 2 - \mathrm{i}\kappa. \tag{2.30}$$

The 't Hooft coupling λ turns out to be a non-trivial function of κ, determined by the normalization of the density. In order for λ to be real and well-defined, κ has to be real as well, and one finds the equation [7]

$$\lambda(\kappa) = \frac{\kappa}{8\pi}\, {}_3F_2\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}; 1, \frac{3}{2}; -\frac{\kappa^2}{16}\right). \tag{2.31}$$

Notice that the endpoints of the cuts are in general complex, i.e. the cuts $\mathcal{C}_1$, $\mathcal{C}_2$ are arcs in the complex plane. This is a consequence of the analytic continuation and it has been verified in numerical simulations of the original saddle-point equations (2.22) [11]. Using similar techniques (see again [13]), one finds a very explicit expression for the derivative of the planar free energy,

$$\partial_\lambda F_0(\lambda) = \frac{\kappa}{4} G^{2,3}_{3,3}\left(\begin{matrix}\frac{1}{2}, & \frac{1}{2}, & \frac{1}{2} \\ 0, & 0, & -\frac{1}{2}\end{matrix}\,\middle|\, -\frac{\kappa^2}{16}\right) + \frac{\pi^2 \mathrm{i}\kappa}{2}\, {}_3F_2\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}; 1, \frac{3}{2}; -\frac{\kappa^2}{16}\right). \tag{2.32}$$

This is written in terms of the auxiliary variable κ, but by using the explicit map (2.31), one can re-express it in terms of the 't Hooft coupling, and one finds the following expansion around $\lambda = 0$,

$$\partial_\lambda F_0(\lambda) = -8\pi^2\lambda\left(\log\left(\frac{\pi\lambda}{2}\right) - 1\right) + \frac{16\pi^4\lambda^3}{9} + \mathcal{O}\left(\lambda^5\right). \tag{2.33}$$

It is easy to see that this reproduces the perturbative, weak coupling expansion of the matrix integral. This also fixes the integration constant, and one can write

$$F_0(\lambda) = \int_0^\lambda \mathrm{d}\lambda'\, \partial_{\lambda'} F_0(\lambda'). \tag{2.34}$$

To study the strong 't Hooft coupling behavior, we notice from (2.31) that large $\lambda \gg 1$ requires $\kappa \gg 1$. More concretely, we find the following expansion at large κ:

$$\lambda(\kappa) = \frac{\log^2(\kappa)}{2\pi^2} + \frac{1}{24} + \mathcal{O}\left(\frac{1}{\kappa^2}\right), \qquad \kappa \gg 1. \tag{2.35}$$

This suggests to define the shifted coupling

$$\hat{\lambda} = \lambda - \frac{1}{24}. \tag{2.36}$$

Notice from (2.8) that this shift is precisely the one needed in order for $\hat{\lambda}$ to be identified with Q/k, at leading order in the string coupling constant. The relationship (2.35) is immediately inverted to

$$\kappa \approx \mathrm{e}^{\pi\sqrt{2\hat{\lambda}}}, \qquad \lambda \gg 1. \tag{2.37}$$

To compute the planar free energy, we have to analytically continue the r.h.s. of (2.32) to $\kappa = \infty$, and we obtain

$$\partial_\lambda F_0(\lambda) = 2\pi^2 \log\kappa + \frac{4\pi^2}{\kappa^2}\, {}_4F_3\left(1, 1, \frac{3}{2}, \frac{3}{2}; 2, 2, 2; -\frac{16}{\kappa^2}\right). \tag{2.38}$$

After integrating w.r.t. λ, we find,

$$F_0(\hat{\lambda}) = \frac{4\pi^3\sqrt{2}}{3}\hat{\lambda}^{3/2} + \frac{\zeta(3)}{2} + \sum_{\ell\geq 1} \mathrm{e}^{-2\pi\ell\sqrt{2\hat{\lambda}}} f_\ell\left(\frac{1}{\pi\sqrt{2\hat{\lambda}}}\right), \tag{2.39}$$

where $f_\ell(x)$ is a polynomial in x of degree $2\ell - 3$ (for $\ell \geq 2$). If we multiply by g_s^{-2}, we find that the leading term in (2.39) agrees precisely with the prediction from the AdS dual in (2.18). The series of exponentially small corrections in (2.39) were interpreted in [8] as coming from worldsheet instantons of type IIA theory wrapping the $\mathbb{CP}^1$ cycle in $\mathbb{CP}^3$. This is a novel type of correction in AdS_4 dualities which is not present in the large N dual to $\mathcal{N} = 4$ super Yang–Mills theory, see [30] for a preliminary investigation of these effects.

An important aspect of the above planar solution is the following. As we explained above, in finding this solution it is useful to take into account the relationship to the lens space matrix model of [25, 26] discovered in [7]. On the other hand, this matrix model computes, in the $1/N$ expansion, the partition function of topological string theory on a non-compact Calabi–Yau (CY) known as local $\mathbb{P}^1 \times \mathbb{P}^1$, and in particular its planar free energy is given by the genus zero free energy or prepotential of this topological string theory. Local $\mathbb{P}^1 \times \mathbb{P}^1$ has two complexified Kähler parameters $T_{1,2}$. It turns out that the ABJM slice in which $N_1 = N_2$ corresponds to the "diagonal" geometry in which $T_1 = T_2$. The relationship of ABJM theory to this topological string theory has been extremely useful in deriving exact answers for many of these quantities, and we will find it again in the sections to follow. For example, the constant term involving $\zeta(3)$ in (2.39) is well-known in topological string theory and it gives the constant map contribution to the genus zero free energy. The series of worldsheet instantons appearing in (2.39) is related to the worldsheet instantons of genus zero in topological string theory, but with one subtlety: the genus zero free energy in (2.39) is the one appropriate to the so-called "orbifold frame" studied in [26], and then it is analytically continued to large λ, which in topological string theory corresponds to the so-called large radius regime. This is not a natural procedure to follow from the point of view of topological strings on local $\mathbb{P}^1 \times \mathbb{P}^1$, where quantities in the orbifold frame are typically expanded around the orbifold point.

2.2.2 Higher genus corrections

The analysis of the previous subsection gives us the leading term in the $1/N$ expansion, but it is of course an important and interesting problem to compute the higher genus free energies with $g \geq 1$. This involves computing subleading $1/N$ corrections to the free energy of the

ABJM matrix model. The computation of such corrections in Hermitian matrix models has a long history, and a general algorithm solving the problem was developed in [31]. However, this algorithm is difficult to implement in practice. In some examples, one can use a more efficient method, developed in the context of topological string theory, which is known as the direct integration of the holomorphic anomaly equations. This method was introduced in [32], and applied to the ABJM matrix model in [8]. The $F_g(\lambda)$ obtained by this method are written in terms of modular forms and they can be obtained recursively, although there is no known closed form expression or generating functional for them. A detailed analysis for the very first g shows that they have the following structure, in terms of the auxiliary variable κ [29][1]:

$$F_g = c_g + f_g\left(\frac{1}{\log \kappa}\right) + \mathcal{O}\left(\frac{1}{\kappa^2}\right), \qquad g \geq 2, \tag{2.40}$$

where

$$c_g = -\frac{4^{g-1}|B_{2g}B_{2g-2}|}{g(2g-2)(2g-2)!} \tag{2.41}$$

involves the Bernoulli numbers B_{2g}, and

$$f_g(x) = \sum_{j=0}^{g} c_j^{(g)} x^{2g-3+j} \tag{2.42}$$

is a polynomial. Physically, the equation (2.40) tells us that the higher genus free energy has a constant contribution, a polynomial contribution in inverse powers of $\lambda^{1/2}$, going like

$$F_g(\lambda) - c_g \approx \lambda^{\frac{3}{2}-g}, \qquad \lambda \gg 1, \quad g \geq 2, \tag{2.43}$$

and an infinite series of corrections due to worldsheet instantons of genus g. The quantities appearing here have a natural interpretation in the context of topological string theory, since the $F_g(\lambda)$ are simply the orbifold higher genus free energies of local $\mathbb{P}^1 \times \mathbb{P}^1$. The constants (2.41) are the well-known constant map contributions to the higher genus free energies, and the worldhseet instantons of type IIA superstring theory appearing in $F_g(\lambda)$ come from the worldsheet instantons of the topological string.

The 't Hooft expansion (2.19) gives an asymptotic series for the free energy, at fixed 't Hooft parameter. General arguments (see [1] for an early statement and [2] for a recent review) suggest that this series diverges factorially. The divergence of the series is controlled by a large N instanton, which is proportional to the exponentially suppressed factor,

$$\exp(-A_{\rm st}(\lambda)/g_s). \tag{2.44}$$

An explicit expression for the instanton action $A_{\rm st}(\lambda)$ was conjectured in [29]. When λ is real and sufficiently large, it is given by

$$A_{\rm st}(\lambda) = \frac{{\rm i}\kappa}{4\pi} G^{2,3}_{3,3}\left(\begin{matrix} \frac{1}{2}, & \frac{1}{2}, & \frac{1}{2} \\ 0, & 0, & -\frac{1}{2} \end{matrix} \,\middle|\, -\frac{\kappa^2}{16}\right) - \frac{\pi\kappa}{2}\, {}_3F_2\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}; 1, \frac{3}{2}; -\frac{\kappa^2}{16}\right) - \pi^2, \tag{2.45}$$

and it is essentially proportional to the derivative of the free energy (2.32). The function $A_{\rm st}(\lambda)$ is complex, and at strong coupling it behaves like,

$$-{\rm i}A_{\rm st}(\lambda) = 2\pi^2\sqrt{2\lambda} + \pi^2{\rm i} + \mathcal{O}\left({\rm e}^{-2\pi\sqrt{2\lambda}}\right), \qquad \lambda \gg 1. \tag{2.46}$$

[1]The constant contribution c_g was not originally included in [8, 29], but this omission was corrected in [28].

Since the genus g amplitudes are real, the complex instanton governing the large order behavior of the $1/N$ expansion must appear together with its complex conjugate, and it leads to an oscillatory asymptotics. If we write

$$A_{\rm st}(\lambda) = |A_{\rm st}(\lambda)| \, {\rm e}^{{\rm i}\theta(\lambda)}. \tag{2.47}$$

we have the behavior,

$$F_g(\lambda) - c_g \sim (2g)! \, |A_{\rm st}(\lambda)|^{-2g} \cos\left(2g\theta(\lambda) + \delta(\lambda)\right), \qquad g \gg 1, \tag{2.48}$$

where $\delta(\lambda)$ is a function of the 't Hooft coupling, which in simple cases is determined by the one-loop corrections around the instanton. The oscillatory asymptotics in (2.48) suggests that the 't Hooft expansion is Borel summable. This was tested in [39] by detailed numerical calculations. However, the Borel resummation of the expansion does not reproduce the correct values of the free energy at finite N and k. The contribution of the complex instanton, which is of order (2.44), should be added in an appropriate way to the Borel-resummed 't Hooft expansion in order to reconstruct the exact answer for the free energy. In practice, this means that one should consider "trans-series" incorporating these exponentially small effects (see for example [2, 38] for an introduction to trans-series.)

However, the resummation of the perturbative free energies in (2.107) in terms of an Airy function suggests another approach to the problem. Conceptually, the resummation of the genus expansion type IIA superstring theory should be achieved when we go to M-theory. The non-perturbative effects appearing in (2.44) should also appear naturally in an M-theory approach: by using (2.46), we see that they have the form, for $\lambda \gg 1$,

$$\exp\left(-\sqrt{2}\pi k^{1/2} N^{1/2}\right), \tag{2.49}$$

and by using the AdS/CFT dictionary, we see that the instanton action depends on the length scales as

$$\left(\frac{L}{\ell_p}\right)^3. \tag{2.50}$$

This is the expected dependence for the action of a membrane instanton in M-theory, which corresponds to a D2-brane in type IIA theory. In [29], it was shown that a D2 brane wrapping the $\mathbb{RP}^3$ cycle inside $\mathbb{CP}^3$ would lead to the correct strong coupling limit of the action (2.46). Therefore, by going to M-theory, we could in principle incorporate not only the worldsheet instantons which were not taken into account in (2.107), but also the non-perturbative effects due to membrane instantons. In fact, it is well-known that in M-theory membrane and worldsheet instantons appear on equal footing [40].

2.3 The M-theory expansion

The M-theory expansion is an expansion when N is large and k is fixed, corresponding to the regime (2.9). The original study of the ABJM matrix model (2.21) in [7, 8] was done in the 't Hooft regime (2.14). It is now time to see if we can understand the matrix model directly in the M-theory regime and solve the problems raised at the end of the previous section: can we resum the genus expansion in some way? Can we incorporate the non-perturbative effects due to membrane instantons?

The first direct study of the M-theory regime of the matrix model (2.21) was performed in [11]. However, including further corrections seems difficult to do in the approach of [11]. This motivates another approach which was started in [42] and has proved to be very useful in understanding the corrections to the strict large N limit.

2.3.1 The Fermi gas approach

There is a long tradition relating matrix integrals to fermionic theories. One reason for this is that the Vandermonde determinant

$$\Delta(\mu) = \prod_{i<j} (\mu_i - \mu_j) \tag{2.51}$$

appearing in these integrals can be regarded, roughly speaking, as the Slater determinant for a theory of N one-dimensional fermions with positions μ_i. For example, the fact that this factor vanishes whenever two particles are at the same point can be regarded as a manifestation of Pauli's exclusion particle.

The rewriting of the ABJM matrix integral in terms of fermionic quantities can be regarded as a variant of this idea. It should be remarked however that the Fermi gas approach that we will explain in this section is not a universal technique which can be applied to any matrix integral with a Vandermonde-like interaction. It rather requires a specific type of eigenvalue interaction, which turns out to be typical of many matrix integrals appearing in the localization of Chern–Simons–matter theories.

The starting point for the Fermi gas approach is the observation that the interaction term in the matrix integral (2.21) can be rewritten by using the Cauchy identity,

$$\begin{aligned} \frac{\prod_{i<j} \left[2\sinh\left(\frac{\mu_i-\mu_j}{2}\right)\right]\left[2\sinh\left(\frac{\nu_i-\nu_j}{2}\right)\right]}{\prod_{i,j} 2\cosh\left(\frac{\mu_i-\nu_j}{2}\right)} &= \det_{ij} \frac{1}{2\cosh\left(\frac{\mu_i-\nu_j}{2}\right)} \\ &= \sum_{\sigma\in S_N} (-1)^{\epsilon(\sigma)} \prod_i \frac{1}{2\cosh\left(\frac{\mu_i-\nu_{\sigma(i)}}{2}\right)}. \end{aligned} \tag{2.52}$$

In this equation, S_N is the permutation group of N elements, and $\epsilon(\sigma)$ is the signature of the permutation σ. After some manipulations spelled out in detail in [43] one obtains [43, 42]

$$Z(N,k) = \frac{1}{N!}\sum_{\sigma\in S_N} (-1)^{\epsilon(\sigma)} \int \frac{\mathrm{d}^N x}{(2\pi k)^N} \frac{1}{\prod_i 2\cosh\left(\frac{x_i}{2}\right) 2\cosh\left(\frac{x_i - x_{\sigma(i)}}{2k}\right)}. \tag{2.53}$$

This expression can be immediately identified [44] as the canonical partition function of a one-dimensional ideal Fermi gas of N particles and with canonical density matrix

$$\rho(x_1,x_2) = \frac{1}{2\pi k} \frac{1}{\left(2\cosh\frac{x_1}{2}\right)^{1/2}} \frac{1}{\left(2\cosh\frac{x_2}{2}\right)^{1/2}} \frac{1}{2\cosh\left(\frac{x_1-x_2}{2k}\right)}. \tag{2.54}$$

Notice that, by using the Cauchy identity again, with $\mu_i = \nu_i$, we can rewrite (2.53) as a matrix integral involving one single set of N eigenvalues,

$$Z(N,k) = \frac{1}{N!}\int \prod_{i=1}^N \frac{\mathrm{d}x_i}{4\pi k} \frac{1}{2\cosh\frac{x_i}{2}} \prod_{i<j}\left(\tanh\left(\frac{x_i-x_j}{2k}\right)\right)^2. \tag{2.55}$$

The canonical density matrix (2.54) is related to the Hamiltonian operator $\hat{H}$ in the usual way,

$$\rho(x_1,x_2) = \langle x_1|\hat{\rho}|x_2\rangle, \qquad \hat{\rho} = \mathrm{e}^{-\hat{H}}, \tag{2.56}$$

where the inverse temperature $\beta = 1$ is fixed. We will come back to the construction of the Hamiltonian shortly.

Since ideal quantum gases are better studied in the grand canonical ensemble, the above representation suggests to look at the grand canonical partition function, defined by

$$\Xi(\mu, k) = 1 + \sum_{N \geq 1} Z(N, k) \mathrm{e}^{N\mu}. \tag{2.57}$$

Here, μ is the chemical potential. The grand canonical potential is

$$\mathcal{J}(\mu, k) = \log \Xi(\mu, k). \tag{2.58}$$

A standard argument (presented for example in [44]) tells us that

$$\mathcal{J}(\mu, k) = -\sum_{\ell \geq 1} \frac{(-\kappa)^\ell}{\ell} Z_\ell, \tag{2.59}$$

where

$$\kappa = \mathrm{e}^\mu \tag{2.60}$$

is the fugacity, and

$$Z_\ell = \mathrm{Tr} \hat{\rho}^\ell = \int \mathrm{d}x_1 \cdots \mathrm{d}x_\ell \, \rho(x_1, x_2) \rho(x_2, x_3) \cdots \rho(x_{\ell-1}, x_\ell) \rho(x_\ell, x_1). \tag{2.61}$$

As is well-known, the canonical and the grand-canonical formulations are equivalent, and the canonical partition function is recovered from the grand canonical one by integration,

$$Z(N, k) = \oint \frac{\mathrm{d}\kappa}{2\pi \mathrm{i}} \frac{\Xi(\mu, k)}{\kappa^{N+1}}. \tag{2.62}$$

Since we are dealing with an ideal gas, all the physics is in principle encoded in the spectrum of the Hamiltonian $\hat{H}$. This spectrum is defined by,

$$\hat{\rho} |\varphi_n\rangle = \mathrm{e}^{-E_n} |\varphi_n\rangle, \qquad n = 0, 1, \ldots, \tag{2.63}$$

or equivalently by the integral equation associated to the kernel (2.54),

$$\int \rho(x, x') \varphi_n(x') \mathrm{d}x' = \mathrm{e}^{-E_n} \varphi_n(x), \qquad n = 0, 1, \ldots. \tag{2.64}$$

It can be verified that this spectrum is indeed discrete and the energies are real. This is because, as it can be easily checked, (2.54) defines a positive, Hilbert–Schmidt operator on $L^2(\mathbb{R})$, and the above properties of the spectrum follow from standard results in the theory of such operators (see, for example, [45]). The thermodynamics is completely determined by the spectrum: the grand canonical partition function is given by the Fredholm determinant associated to the integral operator (2.54),

$$\Xi(\mu, k) = \det(1 + \kappa \hat{\rho}) = \prod_{n \geq 0} \left(1 + \kappa \mathrm{e}^{-E_n}\right). \tag{2.65}$$

In terms of the density of eigenvalues

$$\rho(E) = \sum_{n \geq 0} \delta(E - E_n), \tag{2.66}$$

we also have the standard formula

$$\mathcal{J}(\mu, k) = \int_0^\infty \mathrm{d}E\, \rho(E)\, \log\left(1 + \kappa \mathrm{e}^{-E}\right). \tag{2.67}$$

What can we learn from the ABJM partition function in the Fermi gas formalism? The first thing we can do is to derive the strict large N limit of the free energy, including the correct coefficient. To do this, we have to be more precise about the Hamiltonian of the theory, which is defined implicitly by (2.56) and (2.54). Let us first write the density matrix (2.54) as

$$\hat{\rho} = \mathrm{e}^{-\frac{1}{2}U(\hat{x})}\mathrm{e}^{-T(\hat{p})}\mathrm{e}^{-\frac{1}{2}U(\hat{x})}. \tag{2.68}$$

In this equation, $\hat{x}, \hat{p}$ are canonically conjugate operators,

$$[\hat{x}, \hat{p}] = \mathrm{i}\hbar, \tag{2.69}$$

and

$$\hbar = 2\pi k. \tag{2.70}$$

Note that $\hbar$ is the inverse coupling constant of the gauge theory/string theory, therefore semi-classical or WKB expansions in the Fermi gas correspond to strong coupling expansions in gauge theory/string theory. Finally, the potential $U(x)$ in (2.68) is given by

$$U(x) = \log\left(2\cosh\frac{x}{2}\right), \tag{2.71}$$

and the kinetic term $T(p)$ is given by the same function,

$$T(p) = \log\left(2\cosh\frac{p}{2}\right). \tag{2.72}$$

Exercise 2.2. Show that, for the operator (2.68), we have

$$\langle x'|\hat{\rho}|x\rangle = \frac{1}{2\pi k}\frac{1}{\left(2\cosh\frac{x_1}{2}\right)^{1/2}}\frac{1}{\left(2\cosh\frac{x_2}{2}\right)^{1/2}}\frac{1}{2\cosh\left(\frac{x_1-x_2}{2k}\right)}. \tag{2.73}$$

□

The resulting Hamiltonian is not standard. First of all, the kinetic term leads to an operator involving an infinite number of derivatives (by expanding it around $p = 0$), and it should be regarded as a difference operator, as we will see later. Second, the ordering of the operators in (2.68) shows that the Hamiltonian we are dealing with is not the sum of the kinetic term plus the potential, but it includes $\hbar$ corrections due to non-trivial commutators. This is for example what happens when one considers quantum theories on the lattice: the standard Hamiltonian is only recovered in the continuum limit, which sets the commutators to zero. All these complications can be treated appropriately, and we will address some of them in this expository article, but let us first try to understand what happens when N is large.

The potential in (2.71) is a confining one, and at large x it behaves linearly,

$$U(x) \sim \frac{|x|}{2}, \qquad |x| \to \infty. \tag{2.74}$$

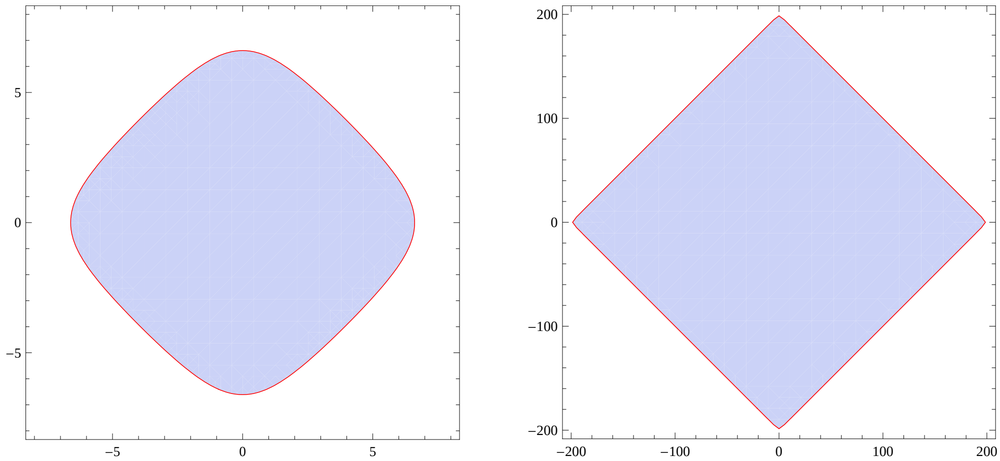

Figure 3: The Fermi surface (2.76) for ABJM theory in the $q = x$-p plane, for $E = 4$ (left) and $E = 100$ (right). When the energy is large, the Fermi surface approaches the polygon (2.77).

When the number of particles in the gas, N, is large, the typical energies are large, and we are in the semiclassical regime. In that case, we can ignore the quantum corrections to the Hamiltonian and take its classical limit

$$H_{\rm cl}(x,p) = U(x) + T(p). \tag{2.75}$$

Standard semiclassical considerations indicate that the number of particles N is given by the area of the Fermi surface, defined by

$$H_{\rm cl}(x,p) = E, \tag{2.76}$$

divided by $2\pi\hbar$, the volume of an elementary cell. However, for large E, we can replace $U(x)$ and $T(p)$ by their leading behaviors at large argument, so that the Fermi surface is well approximated by the polygon,

$$\frac{|x|}{2} + \frac{|p|}{2} = E. \tag{2.77}$$

This can be seen in Fig. 3, where we show the Fermi surface computed from (2.76) for two values of the energies, a moderate one and a large one. For the large one, the Fermi surface is very well approximated by the polygon of (2.77). The area of this polygon is $8E^2$. Therefore, by using the relation between the grand potential and the average number of particles,

$$\frac{\partial \mathcal{J}(\mu,k)}{\partial \mu} = \langle N(\mu,k)\rangle \approx \frac{8\mu^2}{2\pi\hbar}, \tag{2.78}$$

we obtain immediately

$$\mathcal{J}(\mu,k) \approx \frac{2\mu^3}{3\pi^2 k}. \tag{2.79}$$

To compute the free energy, we note that, at large N, the contour integral (2.62) can be computed by a saddle–point approximation, which leads to the standard Legendre transform,

$$F(N,k) \approx \mathcal{J}(\mu_*,k) - \mu_* N, \tag{2.80}$$

where μ_* is the function of N and k defined by (2.78), i.e.

$$\mu_* \approx \frac{\sqrt{2}}{2}\pi k^{1/2} N^{1/2}. \tag{2.81}$$

In this way, we immediately recover the result (2.17) from (2.80). In particular, the scaling 3/2 is a simple consequence from the analysis: it is the expected scaling for a Fermi gas in one dimension with a linearly confining potential and an ultra-relativistic dispersion relation $T(p) \propto |p|$ at large p. This is arguably the simplest derivation of the result (2.17), as it uses only elementary notions in Statistical Mechanics. Note that in this derivation we have considered the M-theory regime in which N is large and k is fixed, and we have focused on the strict large N limit considered in [11] and reviewed in the last section. The main questions is now: can we use the Fermi gas formulation to obtain explicit results for the corrections to the ABJM partition function? In the next sections we will address this question.

2.3.2 The WKB expansion of the Fermi gas

In the Fermi gas approach, the physics of the partition function is encoded in a quantum ideal gas. Although the gas is non-interacting, its one-particle Hamiltonian is complicated, and the energy levels E_n in (2.63) are not known in closed form. What can we do in this situation? As we have seen in (2.70), the parameter k corresponds to the Planck constant of the quantum Fermi gas. Therefore, we can try a systematic development around $k = 0$, i.e. a semiclassical WKB approximation. Such an approach should give a way of computing corrections to (2.79) and (2.17). Of course, we are not *a priori* interested in the physics at small k, but rather at *finite* k, and in particular at integer k, which corresponds to the non-perturbative definition of the theory. However, the expansion at small k gives interesting clues about the problem and it can be treated systematically.

There are two ways of working out the WKB expansion: we can work directly at the level of the grand potential, or we can work at the level of the energy spectrum. Let us first consider the problem at the level of the grand potential. It turns out that, in order to perform a systematic semiclassical expansion, the most useful approach is Wigner's phase space formulation of Quantum Mechanics (in fact, this formulation was originally introduced by Wigner in order to understand the semiclassical expansion of thermodynamic quantities.) A detailed application of this formalism to the ABJM Fermi gas can be found in [42, 46]. The main idea of the method is to map quantum-mechanical operators to functions in classical phase space through the Wigner transform. Under this map, the product of operators famously becomes the $\star$ or Moyal product (see for example [47] for a review, and [48] for an elegant summary with applications). This approach is particularly useful in view of the nature of our Hamiltonian $\hat{H}$, which includes quantum corrections. The Wigner transform of $\hat{H}$ has the structure

$$H_{\mathrm{W}}(x,p) = H_{\mathrm{cl}}(x,p) + \sum_{n\geq 1} \hbar^{2n} H_{\mathrm{W}}^{(n)}(x,p), \tag{2.82}$$

where $H_{\mathrm{cl}}(x,p)$ is the classical Hamiltonian introduced in (2.75). Proceeding in this way, we obtain a systematic $\hbar$ expansion of all the quantities of the theory. The WKB expansion of the grand potential reads,

$$\mathcal{J}^{\mathrm{WKB}}(\mu,k) = \sum_{n\geq 0} \mathcal{J}_n(\mu) k^{2n-1}. \tag{2.83}$$

Note that this is principle an approximation to the full function $\mathcal{J}(\mu,k)$, since it does not take into account non-perturbative effects in $\hbar$. The functions $\mathcal{J}_n(\mu)$ in this expansion can be in principle computed in closed form, although their calculation becomes more and more cumbersome as n grows. The leading term $n = 0$ is however relatively easy to compute [42]. We first notice that the traces (2.61) have a simple semiclassical limit,

$$Z_\ell \approx \int \frac{\mathrm{d}x\mathrm{d}p}{2\pi\hbar} \mathrm{e}^{-\ell H_{\rm cl}(x,p)}, \qquad k \to 0, \tag{2.84}$$

which is just the classical average, with an appropriate measure which includes the volume of the elementary quantum cell $2\pi\hbar$. By using the integral

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}\xi}{\left(2\cosh\frac{\xi}{2}\right)^\ell} = \frac{\Gamma^2(\ell/2)}{\Gamma(\ell)}, \tag{2.85}$$

we find

$$kZ_\ell \approx \frac{1}{2\pi}\frac{\Gamma^4(\ell/2)}{\Gamma^2(\ell)}, \qquad k \to 0, \tag{2.86}$$

and

$$\mathcal{J}_0(\mu) = -\sum_{\ell=1}^{\infty} \frac{(-\kappa)^\ell}{4\pi^2}\frac{\Gamma^4(\ell/2)}{\ell\Gamma^2(\ell)}. \tag{2.87}$$

This expression is convenient when κ is small, i.e. for $\mu \to -\infty$. To make contact with the large N limit, we need to consider the limit of large, positive chemical potential, $\mu \to +\infty$. This can be done by using a Mellin–Barnes integral, and one finds

$$\mathcal{J}_0(\mu) = \frac{2\mu^3}{3\pi^2} + \frac{\mu}{3} + \frac{2\zeta(3)}{\pi^2} + J_0^{\rm M2}(\mu), \tag{2.88}$$

where

$$J_0^{\rm M2}(\mu) = \sum_{\ell=1}^{\infty} \left(a_{0,\ell}\mu^2 + b_{0,\ell}\mu + c_{0,\ell}\right) \mathrm{e}^{-2\ell\mu}. \tag{2.89}$$

The leading, cubic term in μ in (2.88) is the one we found in (2.79). The subleading term in μ gives a correction of order $N^{1/2}$ to the leading behavior (2.17). The function $J_0^{\rm M2}(\mu)$ involves an infinite series of exponentially small corrections in μ. Note that, although this result for $\mathcal{J}_0(\mu)$ is semiclassical in k, it goes beyond the leading result at large N in (2.17). This is because it takes into account the exact classical Fermi surface (2.76), rather than its polygonal approximation (2.77). Therefore, we see that, already at this level, the Fermi gas approach makes it possible to go beyond the strict large N limit.

Exercise 2.3. Consider the expression (2.87) for the semiclassical grand potential, and write it as a Mellin–Barnes integral,

$$\mathcal{J}_0(\kappa) = -\frac{1}{4\pi^2}\int_{\mathcal{I}} \frac{\mathrm{d}s}{2\pi\mathrm{i}} \frac{\Gamma(-s)\Gamma(s/2)^4}{\Gamma(s)}\kappa^s, \tag{2.90}$$

where the contour $\mathcal{I}$ runs parallel to the imaginary axis, see Fig. 4. It can be deformed so that the integral encloses the poles of $\Gamma(-s)$ at $s = n$, $n = 1, \cdots$ (in the clockwise direction). The residues at these poles give back the infinite series in (2.87). We can however deform the contour in the opposite direction, so that it encloses the poles at poles at

$$s = -2m, \qquad m = 0, 1, 2, \cdots \tag{2.91}$$

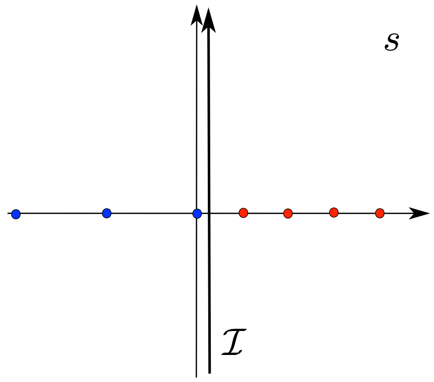

Figure 4: The contour $\mathcal{I}$ in the complex s plane. By closing the contour to the right, we encircle the poles at $s = n$, $n \in \mathbb{Z}_{>0}$. By closing the contour to the left, we encircle the poles at $s = -2n$, $n \in \mathbb{Z}_{\geq 0}$.

Show that

$$\mathcal{J}_0(\kappa) = -\frac{1}{4\pi^2} \sum_{n=0}^{\infty} \mathrm{Res}_{s=-2n} \left[\frac{\Gamma(-s)\Gamma(s/2)^4}{\Gamma(s)} \kappa^s \right]. \tag{2.92}$$

Show that the pole at $s = 0$ gives

$$\frac{2}{3\pi^2}\mu^3 + \frac{1}{3}\mu + \frac{2\zeta(3)}{\pi^2}, \tag{2.93}$$

which is precisely the leading part of (2.88) as $\mu \to \infty$. Compute the contribution of the other poles and show that one finds a series of the form (2.89). $\square$

It is possible to go beyond the leading order of the WKB expansion of the grand potential and compute the corrections appearing in (2.83). The function $\mathcal{J}_1(\mu)$ was also derived in [42] and its large μ expansion has the following form,

$$\mathcal{J}_1(\mu) = \frac{\mu}{24} - \frac{1}{12} + \mathcal{O}\left(\mu^2 \mathrm{e}^{-2\mu}\right). \tag{2.94}$$

Moreover, the following non-renormalization theorem can be proved [42]: for $n \geq 2$, the n-th order correction to the WKB expansion is given by a constant A_n (independent of μ), plus a function which is exponentially suppressed as $\mu \to \infty$, i.e.

$$\mathcal{J}_n(\mu) = A_n + \mathcal{O}\left(\mu^2 \mathrm{e}^{-2\mu}\right), \qquad n \geq 2. \tag{2.95}$$

The exponentially small terms appearing in the functions $\mathcal{J}_n(\mu)$ with $n \geq 1$ have the same structure as for $n = 0$. We then conclude that, in the WKB approximation, i.e. as a power series in k around $k = 0$, the grand potential has the structure

$$\mathcal{J}^{\mathrm{WKB}}(\mu, k) = J^{(\mathrm{p})}(\mu) + J^{\mathrm{M2}}(\mu, k). \tag{2.96}$$

The "perturbative" piece $J^{(\mathrm{p})}(\mu)$ is given by

$$J^{(\mathrm{p})}(\mu) = \frac{C(k)}{3}\mu^3 + B(k)\mu + A(k), \tag{2.97}$$

In this equation, $C(k)$ is given by

$$C(k) = \frac{2}{\pi^2 k}, \tag{2.98}$$

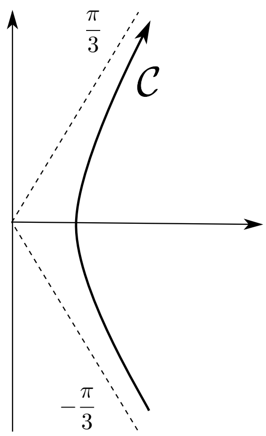

Figure 5: The contour $\mathcal{C}$ in the complex plane of the chemical potential.

as obtained already in (2.79). The coefficient $B(k)$ is given by

$$B(k) = \frac{1}{3k} + \frac{k}{24}, \tag{2.99}$$

where the first term comes from (2.88) and the second term comes from the WKB correction (2.94). Finally, the function $A(k)$ in (2.102) has a power series expansion in k, around $k = 0$, of the form

$$A(k) = \sum_{n \geq 0} A_n k^{2n-1}, \tag{2.100}$$

where

$$A_0 = \frac{2\zeta(3)}{\pi^2}, \qquad A_1 = -\frac{1}{12}, \tag{2.101}$$

as one finds from (2.88) and (2.94). An exact expression for this function was proposed in [28] and it was slightly simplified in [33] to the form,

$$A(k) = \frac{2\zeta(3)}{\pi^2 k}\left(1 - \frac{k^3}{16}\right) + \frac{k^2}{\pi^2}\int_0^\infty \mathrm{d}x \frac{x}{\mathrm{e}^{kx} - 1} \log(1 - \mathrm{e}^{-2x}). \tag{2.102}$$

The second term in the r.h.s. of (2.96) has the structure,

$$J^{\mathrm{M2}}(\mu) = \sum_{\ell=1}^{\infty} \left(a_\ell(k)\mu^2 + b_\ell(k)\mu + c_\ell(k)\right) \mathrm{e}^{-2\ell\mu}, \tag{2.103}$$

where the coefficients have the WKB expansion,

$$a_\ell(k) = \sum_{n \geq 0} a_{n,\ell} k^{2n-1}. \tag{2.104}$$

Similar expansions hold for $b_\ell(k)$, $c_\ell(k)$.

We can now plug the result (2.96) in (2.62). In the μ-plane, this is an integral from $-\pi\mathrm{i}$ to $\pi\mathrm{i}$:

$$Z(N,k) = \frac{1}{2\pi\mathrm{i}} \int_{-\pi\mathrm{i}}^{\pi\mathrm{i}} \frac{\mathrm{d}\mu}{2\pi\mathrm{i}} \mathrm{e}^{\mathcal{J}(\mu,k) - N\mu}. \tag{2.105}$$

If we neglect exponentially small corrections, we can deform the contour $[-\pi\mathrm{i}, \pi\mathrm{i}]$ to the contour $\mathcal{C}$ shown in Fig. 5. Therefore, we find that, up to these corrections,

$$
\begin{aligned}
Z(N,k) &\approx \frac{1}{2\pi\mathrm{i}} \int_{\mathcal{C}} \exp\left(J^{(\mathrm{p})}(\mu) - \mu N\right) \mathrm{d}\mu \\
&= \frac{1}{2\pi\mathrm{i}} \int_{\mathcal{C}} \exp\left[\frac{C(k)}{3}\mu^3 + (B(k) - N)\,\mu + A(k)\right] \mathrm{d}\mu .
\end{aligned}
\tag{2.106}
$$

The above integral can be expressed in terms of the Airy function,

$$
Z(N,k) \approx \mathrm{e}^{A(k)} C^{-1/3}(k) \mathrm{Ai}\left[C^{-1/3}(k)\,(N - B(k))\right]. \tag{2.107}
$$

This expression for the partition function was first obtained in [34], by resumming the results of [8, 29]. It can now be expanded at large N and fixed k (i.e., in the M-theory regime). Let us introduce the parameter

$$
\zeta = 32\pi^2 k\,(N - B(k))\,, \tag{2.108}
$$

Then, we find for the free energy

$$
F(N,k) \approx -\frac{1}{384\pi^2 k}\zeta^{3/2} + \frac{1}{6}\log\left[\frac{\pi^3 k^3}{\zeta^{3/2}}\right] + A(k) + \sum_{n=1}^{\infty} d_{n+1}\pi^{2n} k^n \zeta^{-3n/2}, \tag{2.109}
$$

where the coefficients d_n are just rational numbers,

$$
d_2 = -\frac{80}{3}, \qquad d_3 = 5120, \qquad d_4 = -\frac{18104320}{9}, \qquad d_5 = 1184890880, \qquad \cdots \tag{2.110}
$$

The expression (2.107) is very interesting. First of all, it is a result valid at *all orders* in the $1/N$ expansion and fixed k. It gives an M-theory resummation of the polynomial part of the genus g free energies in (2.40). Moreover, if we assume that the parameter ζ gives the right "renormalized" dictionary between the gauge theory data and the geometry, i.e., if

$$
\left(\frac{L}{\ell_p}\right)^6 = \zeta, \tag{2.111}
$$

then (2.109) is the expected expansion for a free energy in a theory of quantum gravity in eleven dimensions. Indeed, an ℓ-loop term for a vacuum diagram in gravity in d dimensions goes like (see for example [35, 36])

$$
\left(\frac{\ell_p}{L}\right)^{(d-2)(\ell-1)}, \tag{2.112}
$$

which for $d = 11$ agrees with the expansion parameter $\zeta^{-3/2}$ appearing in (2.109). The log term in (2.109) should correspond to a one-loop correction in supergravity, and this was checked by a direct computation in [37], providing in this way a test of the AdS/CFT correspondence beyond the planar limit (in type IIA, this correction comes from the genus one free energy).

We see that the Fermi gas approach leads to a powerful all-orders result in the M-theory regime. In this approach, such a result just requires computing the grand potential at next-to-leading order in the WKB expansion. Although this is a one-loop result, it is exact in k if we neglect exponentially small corrections in μ. Therefore, a one-loop calculation in the grand-canonical ensemble leads to an all-orders result in the canonical ensemble.

Of particular interest are the exponentially small terms in μ in (2.103). What is their meaning? By taking into account that, at large N, μ is given in (2.81), one finds that these corrections to $\mathcal{J}^{\rm WKB}(\mu,k)$ lead to corrections in $Z(N,k)$ precisely of the form (2.49). We recall that these were found originally in the matrix model as non-perturbative effects in the 't Hooft expansion. We conclude that the exponentially small corrections in μ in (2.89), which in the Fermi gas approach appear already in the semi-classical approximation, correspond to non-perturbative corrections to the genus expansion, and should be identified as membrane instanton contributions. One important question is whether we can determine the coefficients appearing in (2.103) as exact functions of k. It turns out that this can be done, as first noted in [50] (and [49]), by relating the spectral problem of the Fermi gas to the quantization of the mirror curve of local $\mathbb{P}^1 \times \mathbb{P}^1$ performed in [52]. The outcome of this analysis is that the coefficients in (2.103) are completely determined by the refined topological string on local $\mathbb{P}^1 \times \mathbb{P}^1$. This gives then the exact membrane instanton contributions, as functions of k. For example, one finds [56]:

$$
\begin{aligned}
a_1(k) &= -\frac{4}{\pi^2 k}\cos\left(\frac{\pi k}{2}\right), \\
b_1(k) &= \frac{2}{\pi}\cos^2\left(\frac{\pi k}{2}\right)\csc\left(\frac{\pi k}{2}\right), \\
c_1(k) &= \left[-\frac{2}{3k}+\frac{5k}{12}+\frac{k}{2}\csc^2\left(\frac{\pi k}{2}\right)+\frac{1}{\pi}\cot\left(\frac{\pi k}{2}\right)\right]\cos\left(\frac{\pi k}{2}\right),
\end{aligned}
\tag{2.113}
$$

and so on. Note that the coefficients $b_1(k)$ and $c_1(k)$ have *poles* at finite k. The meaning of these poles will be discussed shortly.

2.3.3 A conjecture for the exact grand potential

So far we have obtained two differences pieces of information on the matrix model of ABJM theory: on the one hand, the full 't Hooft expansion of the partition function, and on the other hand, the full WKB expansion of the grand potential. Can we put these two pieces of information together? It turns out that the 't Hooft expansion can be incorporated in the grand potential, but this requires a subtle handling of the relationship between the canonical and the grand-canonical ensemble. In the standard thermodynamic relationship, the canonical partition function is given by the integral (2.105). As we have seen in the derivation of (2.106), it is very convenient to extend the integration contour to infinity, along the Airy contour $\mathcal{C}$ shown in Fig. 5. However, this cannot be done without changing the value of the integral: As already noted in [42], if we extend the contour in (2.105) to infinity, we will change the partition function by non-perturbative terms of order

$$
\sim \mathrm{e}^{-\mu/k}. \tag{2.114}
$$

If we want to understand the structure of non-perturbative effects in the ABJM partition function, we have to handle this issue with care. A clever way of proceeding was found in [56]. Following [56], we will introduce an auxiliary object, which we will call the *modified* grand potential, and we will denote it by $J(\mu,k)$. The modified grand potential is *defined* by the equality

$$
Z(N,k) = \int_{\mathcal{C}} \frac{\mathrm{d}\mu}{2\pi \mathrm{i}} \mathrm{e}^{J(\mu,k)-N\mu}, \tag{2.115}
$$

where $\mathcal{C}$ is the Airy contour shown in Fig. 5. As it was noticed in [56], if we know $J(\mu,k)$, we can recover the conventional grand potential $\mathcal{J}(\mu,k)$ by the relation

$$\mathrm{e}^{\mathcal{J}(\mu,k)} = \sum_{n\in\mathbb{Z}} \mathrm{e}^{J(\mu+2\pi \mathrm{i} n,k)}. \tag{2.116}$$

Indeed, if we plug this in (2.105), we can use the sum over n to extend the integration region from $[-\pi\mathrm{i},\pi\mathrm{i}]$ to the full imaginary axis. If we then deform the contour to $\mathcal{C}$, we obtain (2.115). Note that the difference between $J(\mu,k)$ and $\mathcal{J}(\mu,k)$ involves non-perturbative terms of the form (2.114), and it is not seen in a perturbative calculation around $k=0$. Therefore the WKB calculation of the full grand potential still gives the perturbative expansion of the modified grand potential, and we have

$$J(\mu,k) = \mathcal{J}^{\mathrm{WKB}}(\mu,k) + \mathcal{O}\left(\mathrm{e}^{-\mu/k}\right). \tag{2.117}$$

Let us now try to understand how to incorporate the information of the 't Hooft expansion in the grand potential. It is much better to use the modified grand potential, due to the fact that the relationship (2.115) involves an integration going to infinity. Let us denote the 't Hooft contribution to the modified grand potential by $J^{\text{'t Hooft}}(\mu,k)$. Since (2.115) is essentially a Laplace transform, we can easily relate $J^{\text{'t Hooft}}(\mu,k)$ to the 't Hooft expansion of the standard free energy: we simple calculate this Laplace transform by a saddle point calculation as $k\to\infty$. In this way we find the expansion,

$$J^{\text{'t Hooft}}(\mu,k) = \sum_{g=0}^{\infty} k^{2-2g} J_g\left(\frac{\mu}{k}\right), \tag{2.118}$$

which contains exactly the same information than the 't Hooft expansion of the canonical partition function. As usual in the saddle-point expansion, the leading terms, which are the genus zero pieces J_0 and $-F_0/(4\pi^2)$, are related by a Legendre transform: we first solve for the 't Hooft parameter λ, in terms of μ/k, through the equation

$$\frac{\mu}{k} = \frac{1}{4\pi^2}\frac{\mathrm{d}F_0}{\mathrm{d}\lambda}, \tag{2.119}$$

and then we have,

$$J_0\left(\frac{\mu}{k}\right) = -\frac{1}{4\pi^2}\left(F_0(\lambda) - \lambda\frac{\mathrm{d}F_0}{\mathrm{d}\lambda}\right). \tag{2.120}$$

Similarly, the genus one grand potential J_1 is related to the genus one free energy F_1 through the equation

$$J_1\left(\frac{\mu}{k}\right) = F_1\left(\frac{\mu}{k}\right) + \frac{1}{2}\log\left(2\pi k^2 \partial_\mu^2 J_0\left(\frac{\mu}{k}\right)\right), \tag{2.121}$$

which takes into account the one-loop correction to the saddle-point. Since the integration contour in (2.115) goes to infinity, doing the saddle-point expansion with Gaussian integrals is fully justified and no error terms of the form (2.114) are introduced in this way. This is clearly one of the advantages of using the modified grand potential, instead of the standard grand potential.

We should recall now that the genus g free energies F_g are given by the topological string free energies of local $\mathbb{P}^1\times\mathbb{P}^1$, in the so-called orbifold frame. It turns out that the Laplace transform which takes us to the grand potential has an interpretation in topological string theory: as shown in [42] by using the general theory of [57], it is the transformation that takes us from the orbifold frame to the so-called large radius frame. Therefore, we can interpret $J^{\text{'t Hooft}}(\mu,k)$ as the total free energy of the topological string on local $\mathbb{P}^1 \times \mathbb{P}^1$ at large radius. This free energy has a polynomial piece, which reproduces precisely the perturbative piece (2.97), and then an infinite series of corrections of the form,

$$e^{-4\mu/k}. \tag{2.122}$$

These corrections, after the inverse Legendre transform, give back the worldsheet instanton corrections that we found in the 't Hooft expansion of the free energy. Notice however that, from the point of view of the Fermi gas approach, these are non-perturbative in $\hbar$, and correspond to instanton-type corrections in the spectral problem (2.63) [42, 49].

The large radius free energy of local $\mathbb{P}^1 \times \mathbb{P}^1$ is a well studied quantity (in fact, it has been much more studied than the orbifold free energies). In particular, there is a surprising result of Gopakumar and Vafa [58], valid for any CY manifold, which makes it possible to resum the genus expansion in (2.118). In our case, this means that we can resum the genus expansion, order by order in $\exp(-4\mu/k)$ [56]. The result can be written as,

$$J^{\prime\mathrm{tHooft}}(\mu, k) = \frac{C(k)}{3}\mu^3 + B(k)\mu + A(k) + J^{\mathrm{WS}}(\mu, k), \tag{2.123}$$

where

$$J^{\mathrm{WS}}(\mu, k) = \sum_{m \geq 1} (-1)^m d_m(k) e^{-\frac{4m\mu}{k}}, \tag{2.124}$$

and the coefficients $d_m(k)$ are given by

$$d_m(k) = \sum_{g \geq 0} \sum_{m = wd} n_g^d \left(2 \sin \frac{2\pi w}{k}\right)^{2g-2}. \tag{2.125}$$

In this equation, we sum over the positive integers w, d satisfying the constraint $wd = m$. The quantities n_g^d are integer numbers called Gopakumar–Vafa (GV) invariants. One should note that, for any given d, the n_d^g are different from zero only for a finite number of g, therefore (2.125) is well-defined as a formal power series in $\exp(-4\mu/k)$. The GV invariants can be computed by various techniques, and in the case of non-compact CY manifolds, there are algorithms to determine them for all possible values of d and g (like for example the theory of the topological vertex [59].) It is important to note that it is only when we use the modified grand potential that we obtain the results (2.123), (2.124) for the 't Hooft expansion. If we use the standard grand potential, there are additional contributions coming from the "images" of the modified grand potential in the sum (2.116). Note also that the resummation of (2.118) leads to a resummation of the genus g free energies $F_g(\lambda)$, i.e. the terms of the same order in the expansion parameters

$$\exp\left(-2\pi\sqrt{\hat{\lambda}}\right), \qquad \frac{1}{\sqrt{\lambda}} \tag{2.126}$$

can be resummed to all genera.

We have now the most important pieces of the total grand potential, $J^{\mathrm{WS}}(\mu, k)$ and $J^{\mathrm{M2}}(\mu, k)$. Note that, if our expansion parameter is $1/k$, as in the 't Hooft expansion, $J^{\mathrm{WS}}(\mu, k)$ is a resummation of a perturbative series, while $J^{\mathrm{M2}}(\mu, k)$ contains non-perturbative information. Conversely, if our expansion parameter is k, $J^{\mathrm{M2}}(\mu, k)$ is the resummation of a perturbative expansion, while $J^{\mathrm{WS}}(\mu, k)$ is non-perturbative. One would be tempted to conclude that the total, modified grand potential is given by

$$J^{(\mathrm{p})}(\mu, k) + J^{\mathrm{WS}}(\mu, k) + J^{\mathrm{M2}}(\mu, k). \tag{2.127}$$

However, it can be seen that this is not the case: there is a "mixing" of the contributions, which was found experimentally in [60]. In order to take into account this mixing, one introduces an "effective" chemical potential $\mu_{\rm eff}$ through the equation,

$$\mu_{\rm eff} = \mu + \frac{1}{C(k)} \sum_{\ell \geq 1} a_\ell(k) {\rm e}^{-2\ell m u}. \tag{2.128}$$

Note that, from the point of view of the 't Hooft expansion, the corrections appearing here are again non-perturbative. Then, the final proposal for the modified grand potential, putting together all the pieces from [7, 8, 42, 55, 56, 60, 50], is the following:

$$J(\mu, k) = J^{(p)}(\mu, k) + J^{\rm WS}(\mu_{\rm eff}, k) + J^{\rm M2}(\mu, k). \tag{2.129}$$

Thus, the argument in the worldsheet instanton piece has to be corrected by using the effective chemical potential. When the function (2.129) is expanded at large μ, one finds exponentially small corrections in μ of the form,

$$\exp\left\{ -\left(\frac{4n}{k} + 2\ell \right) \mu \right\}. \tag{2.130}$$

In [55] these mixed terms were interpreted as bound states of worldsheet instantons and membrane instantons in the M-theory dual. The appearance of the "effective" chemical potential is relatively easy to understand from the point of view of instanton calculus: as it was pointed out in [49], a quantum-mechanical instanton calculation in the Fermi gas would involve a "corrected" instanton action, given essentially by the A-quantum period. This is precisely what one has in (2.128).

One of the most important properties of the proposal (2.129) is the following. It is easy to see, by looking at the explicit expressions (2.124) and (2.125), that $J^{\rm WS}(\mu, k)$ is singular for any rational k. This is a puzzling result, since it implies that the genus resummation of the free energies F_g is also singular for infinitely many values of k, including the integer values for which the theory is in principle well-defined non-perturbatively. It is however clear that this divergence is an artifact of the genus expansion, since the original matrix integral (2.21), as well as its Fermi gas form (2.55), are perfectly well-defined for any real value of k. What is going on?

It was first proposed in [56] that the divergences in the resummation of the genus expansion of $J^{\rm WS}(\mu, k)$ should be cured by other terms in the modified grand potential, in such a way that the total result is finite. It turns out that $J^{\rm M2}(\mu, k)$ is also singular, as we have seen in (2.113). It turns out that its singularities *cancel* those of $J^{\rm WS}(\mu, k)$ in such a way that the total $J(\mu, k)$ is finite. This remarkable property of $J(\mu, k)$ was called in [55] the HMO cancellation mechanism. Originally, this mechanism was used as a way to understand the structure of $J^{\rm M2}(\mu, k)$ at finite k. In [50] it was shown that this cancellation is a consequence of the structure of the the modified grand potential, and it can be proved by using the underlying geometric structure of $J^{\rm M2}(\mu, k)$ and $J^{\rm WS}(\mu, k)$.

The HMO cancellation mechanism is conceptually important for a deeper understanding of the non-perturbative structure of M-theory. In the M-theory expansion, we have to resum the worldsheet instanton contributions to the free energy at fixed k and large N, which is precisely what the Gopakumar–Vafa representation (2.124) does for us. However, when one does that, the resulting expression is singular and displays an infinite number of poles. We obtain a finite result only when the contribution of membrane instantons has been added. This shows very clearly that a theory based solely on fundamental strings is fundamentally incomplete, and that additional extended objects are needed in M-theory. Of course, this has been clear since the advent of M-theory, but the above calculation shows that the contribution of membranes is not just a correction to the contribution of fundamental strings; it is crucial to remove the poles and to make the amplitude well-defined. Conversely, a theory based only on membrane instantons will be also incomplete and will require fundamental strings in order to make sense.

The result (2.129) is the current proposal for the grand potential of the ABJM matrix model. It is a remarkable exact result. From the point of view of gauge theory, it encodes the full $1/N$ expansion, at all genera, as well as non-perturbative corrections at large N (due presumably to some form of large N instanton). From the point of view of M-theory, it incorporates both membrane instantons and worldsheet instantons. One interesting application of (2.129) is the determination of the exact spectrum for the spectral problem (2.63) [61]. This solves, to a large extent, the problem of determining the exact properties of the ABJM Fermi gas.

The result (2.129) is relatively complicated, since it involves an enormous amount of information, including the all-genus Gopakumar–Vafa invariants of local $\mathbb{P}^1 \times \mathbb{P}^1$, and the quantum periods of this manifold to all orders. However, there are certain cases in which (2.129) can be simplified, as first noted in [64]. These are precisely the cases in which $k = 1$ or $k = 2$ and ABJM theory has enhanced $\mathcal{N} = 8$ supersymmetry. In these cases, one can write an explicit generating functional for the partition functions $Z(N, k)$.

3. Generalizations

3.1 The Fermi gas approach to Chern–Simons–matter theories

Let us briefly mention some generalizations of the Fermi gas approach. The $\mathcal{N} = 3$ models where it can be applied more successfully are the necklace quiver gauge theories constructed in [65, 66]. These theories are given by a Chern–Simons quiver with gauge groups and levels,

$$U(N)_{k_1} \times U(N)_{k_2} \times \cdots U(N)_{k_r}. \tag{3.1}$$

Each node is labelled with the letter $a = 1, \cdots, r$. There are bifundamental chiral superfields A_{aa+1}, B_{aa-1} connecting adjacent nodes, and in addition there can be N_{f_a} matter superfields $(Q_a, \tilde{Q}_a)$ in each node, in the fundamental representation. We will write

$$k_a = n_a k, \tag{3.2}$$

and we will assume that

$$\sum_{a=1}^{r} n_a = 0. \tag{3.3}$$

The matrix model computing the $\mathbb{S}^3$ partition function of such a necklace quiver gauge theory is given by

$$Z(N, n_a, N_{f_a}, k) = \frac{1}{(N!)^r} \int \prod_{a,i} \frac{d\lambda_{a,i}}{2\pi} \frac{\exp\left[\frac{in_a k}{4\pi}\lambda_{a,i}^2\right]}{\left(2\cosh\frac{\lambda_{a,i}}{2}\right)^{N_{f_a}}} \prod_{a=1}^{r} \frac{\prod_{i<j}\left[2\sinh\left(\frac{\lambda_{a,i}-\lambda_{a,j}}{2}\right)\right]^2}{\prod_{i,j} 2\cosh\left(\frac{\lambda_{a,i}-\lambda_{a+1,j}}{2}\right)}. \tag{3.4}$$

To construct the corresponding Fermi gas, we define a kernel corresponding to a pair of connected nodes (a, b) by,

$$K_{ab}(x', x) = \frac{1}{2\pi k} \frac{\exp\left\{i\frac{n_b x^2}{4\pi k}\right\}}{2\cosh\left(\frac{x'-x}{2k}\right)} \left[2\cosh\frac{x}{2k}\right]^{-N_{f_b}}, \tag{3.5}$$

where we set $x = \lambda/k$. If we use the Cauchy identity (2.52), it is easy to see that we can write the grand canonical partition function for this theory in the form (2.65), where now [67]

$$\hat{\rho} = \hat{K}_{r1}\hat{K}_{12}\cdots\hat{K}_{r-1,r} \tag{3.6}$$

is the product of the kernels (3.5) around the quiver. Therefore, we still have a Fermi gas, albeit the Hamiltonian is quite complicated, and we can apply the same techniques that were used for ABJM theory in the previous section. For example, it is possible to analyze it in detail in the thermodynamic limit [42]. One can show that, at large μ, the grand potential of the theory is still of the form

$$\mathcal{J}(\mu, k) \approx \frac{C\mu^3}{3} + B\mu, \qquad \mu \gg 1. \tag{3.7}$$

The coefficient C for a general quiver is also determined, as in ABJM theory, by the volume of the Fermi surface at large energy. This limit is a polygon and one finds,

$$\pi^2 C = \text{ vol}\left\{(x,y)\ :\ \sum_{j=1}^{r}\left|y - \left(\sum_{i=1}^{j-1} k_i\right)x\right| + \left(\sum_{j=1}^{r} N_{f_j}\right)|x| < 1\right\}, \tag{3.8}$$

which can be computed in closed form [68]. The B coefficient can be computed in a case by case basis, although no general formula is known for all $\mathcal{N}=3$ quivers (a general formula is however known for a class of special quivers which preserve $\mathcal{N}=4$ supersymmetry, see [69].)

The result (3.7) has two important consequences. First, the free energy at large N has the behavior

$$F(N) \approx -\frac{2}{3}C^{-1/2}N^{3/2}. \tag{3.9}$$

It can be seen [42], by using the results of [68], that this agrees with the prediction from the M-theory dual (this was also verified in [11, 68]). Second, by using the same techniques as in ABJM theory, we conclude that

$$Z(N) \approx C^{-1/3}(k)\text{Ai}\left[C^{-1/3}\left(N - B\right)\right], \tag{3.10}$$

i.e. the leading Airy behavior of the M-theory expansion of the partition function is a *universal* feature of $\mathcal{N}=3$ Chern–Simons–matter theories. Fermi gas formulation also exist for $\mathcal{N}=3$ theories with other gauge groups [41], and we have again the Airy behavior (3.10). This is arguably one of the most important results obtained from the Fermi gas approach, and it has been conjectured in [42] that any Chern–Simons–matter theory which displays the $N^{3/2}$ behavior in its strict large N limit will have the leading, perturbative behavior (3.10) for its perturbative $1/N$ expansion. The subleading, universal logarithmic prediction for $Z(N)$ implied by (3.10) is also tested by the calculation of [37], and it has been argued that the full Airy function can be obtained from a localization computation in AdS supergravity [70].

Of course, one of the most important advantages of the Fermi gas approach is that it makes it possible to calculate non-perturbative corrections to the 't Hooft expansion. However, for theories other than ABJM theory, this has been in general difficult. An exception is again ABJ theory. The Fermi gas formulation of this theory requires some work, but it has been achieved in the papers [71, 72] (an alternative approach was proposed in [73]). A conjectural formula for the modified grand potential of this theory has been finally proposed in [73, 74]. It has exactly the same properties that we found in ABJM theory: poles appear in the resummation of the genus expansion, and they are cancelled by non-perturbative contributions (presumably coming from membrane instantons), just as in the HMO mechanism of [56].

For other theories which admit a Fermi gas description, the computation of non-perturbative contributions at the level of detail that was done in ABJ(M) theory remains largely an open problem. However, all the existing results indicate that generically the 't Hooft expansion of these theories is radically *insufficient*: not only there are non-perturbative corrections to it, but in addition the resummation of the genus expansion has poles which have to be cured by these non-perturbative effects through a generalization of the HMO mechanism. This pattern has been observed in other models [33, 69].

3.2 Topological strings

One of the key ingredients for the exact determination of the grand potential of ABJM theory has been the relationship to topological string theory. Conversely, one might hope that the structures emerging in the context of Chern–Simons–matter theories are relevant for a better understanding of topological strings. In particular, as already pointed out in [75], the results of [42] suggest a Fermi gas approach to topological string theory on toric CY manifolds which has been pursued in [50, 49, 76]. When the CY X is given by the anti-canonical bundle

$$\mathcal{O}(-K_S) \to S \tag{3.11}$$

over a del Pezzo surface S, a detailed proposal has been presented in [63] and further developed in [77, 78]. The proposal of [63] gives a non-perturbative definition of (closed) topological string theory on X in terms of the operator $\hat{\rho}_S$ obtained by quantizing its mirror curve. The resulting picture is structurally very similar to the theory of the ABJM partition function reviewed in this article. The operator $\hat{\rho}_S$ plays the rôle of the canonical density matrix of the Fermi gas (2.54). For example, in the case of local $\mathbb{P}^2$, the relevant operator is

$$\hat{\rho}_{\mathbb{P}^2} = \left(\mathrm{e}^{\hat{x}} + \mathrm{e}^{\hat{p}} + \mathrm{e}^{-\hat{x}-\hat{p}}\right)^{-1}, \tag{3.12}$$

where $\hat{x}$, $\hat{p}$ are canonically conjugate Heisenberg operators, as in (2.69). In contrast to what happens in ABJM theory, the kernel of this operator is not elementary. However, it can be calculated explicitly in terms of Faddeev's quantum dilogarithm [77].

We can now consider a Fermi gas of N particles whose canonical density matrix is given by $\rho_{\mathbb{P}^2}$. Its canonical partition function $Z(N,\hbar)$ has a matrix model representation of the form

$$Z_S(N,\hbar) = \frac{1}{N!}\sum_{\sigma\in S_N}(-1)^{\epsilon(\sigma)}\int \mathrm{d}^N x \prod_i \rho_S(x_i, x_{\sigma(i)}) = \frac{1}{N!}\int \mathrm{d}^N x \det\left(\rho_S(x_i,x_j)\right). \tag{3.13}$$

The $1/N$ expansion of this matrix model, in the 't Hooft limit

$$N \to \infty, \qquad \frac{N}{\hbar} = \lambda = \text{fixed}, \tag{3.14}$$

gives the all-genus topological string free energies,

$$\log Z(N,\hbar) = \sum_{g\geq 0} \mathcal{F}_g^S(\lambda)\hbar^{2-2g}, \tag{3.15}$$

where λ parametrizes the moduli space of the local del Pezzo. Therefore, the genus expansion of the topological string on X is realized as the asymptotic expansion of a well-defined quantity,

$Z_S(N,\hbar)$. In addition, there are non-perturbative corrections to this expansion which are captured by the refined topological string free energy of the CY manifold, in the NS limit. In the case of local $\mathbb{P}^2$, the matrix model takes the form,

$$Z_{\mathbb{P}^2}(N,\hbar) = \frac{1}{N!}\int_{\mathbb{R}^N} \frac{\mathrm{d}^N u}{(2\pi)^N} \prod_{i=1}^N \mathrm{e}^{-V_{\mathbb{P}^2}(u_i,\hbar)} \frac{\prod_{i<j} 4\sinh\left(\frac{u_i-u_j}{2}\right)^2}{\prod_{i,j} 2\cosh\left(\frac{u_i-u_j}{2}+\frac{\mathrm{i}\pi}{6}\right)}, \tag{3.16}$$

where

$$V_{\mathbb{P}^2}(u,\hbar) = \frac{\hbar u}{2\pi} + \log\left[\frac{\Phi_b\left(\frac{bu}{2\pi}+\frac{\mathrm{i}b}{3}\right)}{\Phi_b\left(\frac{bu}{2\pi}-\frac{\mathrm{i}b}{3}\right)}\right]. \tag{3.17}$$

In this equation, $\Phi_b(x)$ is Faddeev's quantum dilogarithm (where we use the conventions of [77, 78]) and b is related to $\hbar$ by

$$b^2 = \frac{3\hbar}{2\pi}. \tag{3.18}$$

The function $V_{\mathbb{P}^2}(u,\hbar)$ appearing here is relatively complicated. However, it is completely determined by the operator (3.12). It has an asymptotic expansion at large $\hbar$ which can be used to compute the asymptotic expansion of the matrix integral (3.16) in the 't Hooft regime (2.14), as explained in [78]. In fact, this matrix integral is structurally very similar to the generalizations of the $O(2)$ matrix model considered in [79], and might be exactly solvable in the planar limit.

Acknowledgements

This work is supported in part by the Fonds National Suisse, subsidies 200021-156995 and 200020-141329, and by the NCCR 51NF40-141869 "The Mathematics of Physics" (SwissMAP).

References

[1] S. H. Shenker, "The Strength of nonperturbative effects in string theory," in O. Alvarez, E. Marinari, P. Windey (eds.), *Ramdom surfaces and quantum gravity*, Plenum Press, New York, 1991.

[2] M. Mariño, "Lectures on non-perturbative effects in large N gauge theories, matrix models and strings," Fortsch. Phys. **62**, 455 (2014) [arXiv:1206.6272 [hep-th]].

[3] J. Polchinski, "Combinatorics of boundaries in string theory," Phys. Rev. D **50**, 6041 (1994) [arXiv:hep-th/9407031].

[4] D. J. Gross and V. Periwal, "String Perturbation Theory Diverges," Phys. Rev. Lett. **60**, 2105 (1988).

[5] V. Pestun, "Localization of gauge theory on a four-sphere and supersymmetric Wilson loops," Commun. Math. Phys. **313**, 71 (2012) [arXiv:0712.2824 [hep-th]].

[6] A. Kapustin, B. Willett and I. Yaakov, "Exact Results for Wilson Loops in Superconformal Chern-Simons Theories with Matter," JHEP **1003**, 089 (2010) [arXiv:0909.4559 [hep-th]].

[7] M. Mariño and P. Putrov, "Exact Results in ABJM Theory from Topological Strings," JHEP **1006**, 011 (2010) [arXiv:0912.3074 [hep-th]].

[8] N. Drukker, M. Mariño, P. Putrov, "From weak to strong coupling in ABJM theory," Commun. Math. Phys. **306**, 511-563 (2011). [arXiv:1007.3837 [hep-th]].

[9] O. Aharony, O. Bergman, D. L. Jafferis and J. Maldacena, "N=6 superconformal Chern-Simons-matter theories, M2-branes and their gravity duals," JHEP **0810**, 091 (2008) [arXiv:0806.1218 [hep-th]].

[10] O. Aharony, O. Bergman and D. L. Jafferis, "Fractional M2-branes," JHEP **0811**, 043 (2008) [arXiv:0807.4924].

[11] C. P. Herzog, I. R. Klebanov, S. S. Pufu and T. Tesileanu, "Multi-Matrix Models and Tri-Sasaki Einstein Spaces," Phys. Rev. D **83**, 046001 (2011) [arXiv:1011.5487 [hep-th]].

[12] J. Bagger, N. Lambert, S. Mukhi and C. Papageorgakis, "Multiple Membranes in M-theory," Phys. Rept. **527**, 1 (2013) [arXiv:1203.3546 [hep-th]].

[13] M. Mariño, "Lectures on localization and matrix models in supersymmetric Chern-Simons-matter theories," J. Phys. A **44**, 463001 (2011) [arXiv:1104.0783 [hep-th]].

[14] G. 't Hooft, "A Planar Diagram Theory for Strong Interactions," Nucl. Phys. B **72**, 461 (1974).

[15] J. M. Maldacena, "The Large N limit of superconformal field theories and supergravity," Adv. Theor. Math. Phys. **2**, 231-252 (1998). [hep-th/9711200].

[16] M. J. Duff, B. E. W. Nilsson, C. N. Pope, "Kaluza-Klein Supergravity," Phys. Rept. **130**, 1-142 (1986).

[17] O. Bergman and S. Hirano, "Anomalous radius shift in AdS_4/CFT_3," JHEP **0907**, 016 (2009) [arXiv:0902.1743].

[18] O. Aharony, A. Hashimoto, S. Hirano, P. Ouyang, "D-brane Charges in Gravitational Duals of 2+1 Dimensional Gauge Theories and Duality Cascades," JHEP **1001**, 072 (2010). [arXiv:0906.2390 [hep-th]].

[19] B. E. W. Nilsson and C. N. Pope, "Hopf Fibration of Eleven-dimensional Supergravity," Class. Quant. Grav. **1**, 499 (1984).

[20] T. Azeyanagi, M. Fujita and M. Hanada, "From the planar limit to M-theory," Phys. Rev. Lett. **110**, no. 12, 121601 (2013) [arXiv:1210.3601 [hep-th]].

[21] E. Witten, "Anti-de Sitter space and holography," Adv. Theor. Math. Phys. **2**, 253 (1998) [arXiv:hep-th/9802150].

[22] R. Emparan, C. V. Johnson and R. C. Myers, "Surface terms as counterterms in the AdS/CFT correspondence," Phys. Rev. D **60**, 104001 (1999) [arXiv:hep-th/9903238].

[23] I. R. Klebanov and A. A. Tseytlin, "Entropy of near-extremal black p-branes," Nucl. Phys. B **475** (1996) 164 [hep-th/9604089].

[24] E. Brézin, C. Itzykson, G. Parisi, J. B. Zuber, "Planar Diagrams," Commun. Math. Phys. **59**, 35 (1978).

[25] M. Mariño, "Chern-Simons theory, matrix integrals, and perturbative three manifold invariants," Commun. Math. Phys. **253**, 25-49 (2004). [hep-th/0207096].

[26] M. Aganagic, A. Klemm, M. Mariño, C. Vafa, "Matrix model as a mirror of Chern-Simons theory," JHEP **0402**, 010 (2004). [hep-th/0211098].

[27] N. Halmagyi, V. Yasnov, "The Spectral curve of the lens space matrix model," JHEP **0911**, 104 (2009). [hep-th/0311117].

[28] M. Hanada, M. Honda, Y. Honma, J. Nishimura, S. Shiba and Y. Yoshida, "Numerical studies of the ABJM theory for arbitrary N at arbitrary coupling constant," JHEP **1205**, 121 (2012) [arXiv:1202.5300 [hep-th]].

[29] N. Drukker, M. Mariño and P. Putrov, "Nonperturbative aspects of ABJM theory," JHEP **1111**, 141 (2011) [arXiv:1103.4844 [hep-th]].

[30] A. Cagnazzo, D. Sorokin, L. Wulff, “String instanton in $AdS_4 \times \mathbb{CP}^3$,” JHEP **1005**, 009 (2010). [arXiv:0911.5228 [hep-th]].

[31] B. Eynard and N. Orantin, “Invariants of algebraic curves and topological expansion,” Commun. Num. Theor. Phys. **1**, 347 (2007) [math-ph/0702045].

[32] M. x. Huang and A. Klemm, “Holomorphic Anomaly in Gauge Theories and Matrix Models,” JHEP **0709**, 054 (2007) [hep-th/0605195].

[33] Y. Hatsuda and K. Okuyama, “Probing non-perturbative effects in M-theory,” arXiv:1407.3786 [hep-th].

[34] H. Fuji, S. Hirano and S. Moriyama, “Summing Up All Genus Free Energy of ABJM Matrix Model,” JHEP **1108**, 001 (2011) [arXiv:1106.4631 [hep-th]].

[35] C. P. Burgess, “Quantum gravity in everyday life: General relativity as an effective field theory,” Living Rev. Rel. **7**, 5 (2004) [gr-qc/0311082].

[36] A. Sen, “Logarithmic Corrections to Schwarzschild and Other Non-extremal Black Hole Entropy in Different Dimensions,” JHEP **1304**, 156 (2013) [arXiv:1205.0971 [hep-th]].

[37] S. Bhattacharyya, A. Grassi, M. Mariño and A. Sen, “A One-Loop Test of Quantum Supergravity,” Class. Quant. Grav. **31**, 015012 (2014) [arXiv:1210.6057 [hep-th]].

[38] D. Dorigoni, “An Introduction to Resurgence, Trans-Series and Alien Calculus,” arXiv:1411.3585 [hep-th].

[39] A. Grassi, M. Mariño and S. Zakany, “Resumming the string perturbation series,” arXiv:1405.4214 [hep-th].

[40] K. Becker, M. Becker and A. Strominger, “Five-branes, membranes and nonperturbative string theory,” Nucl. Phys. B **456**, 130 (1995) [hep-th/9507158].

[41] M. Mezei and S. S. Pufu, “Three-sphere free energy for classical gauge groups,” JHEP **1402**, 037 (2014) [arXiv:1312.0920 [hep-th], arXiv:1312.0920].

[42] M. Mariño and P. Putrov, “ABJM theory as a Fermi gas,” J. Stat. Mech. **1203**, P03001 (2012) [arXiv:1110.4066 [hep-th]].

[43] A. Kapustin, B. Willett, I. Yaakov, “Nonperturbative Tests of Three-Dimensional Dualities,” JHEP **1010**, 013 (2010). [arXiv:1003.5694 [hep-th]].

[44] R. P. Feynman, *Statistical Mechanics*, Westview Press, Boulder, 1998.

[45] F.G. Tricomi, *Integral equations*, Dover, New York, 1985.

[46] A. Klemm, M. Marino, M. Schiereck and M. Soroush, “ABJM Wilson loops in the Fermi gas approach,” arXiv:1207.0611 [hep-th].

[47] M. Hillery, R. F. O’Connell, M. O. Scully, E. P. Wigner, “Distribution functions in physics: Fundamentals,” Phys. Rept. **106**, 121-167 (1984).

[48] B. Grammaticos, A. Voros, “Semiclassical Approximations For Nuclear Hamiltonians. 1. Spin Independent Potentials,” Annals Phys. **123**, 359 (1979).

[49] J. Kallen and M. Mariño, “Instanton effects and quantum spectral curves,” arXiv:1308.6485 [hep-th].

[50] Y. Hatsuda, M. Mariño, S. Moriyama and K. Okuyama, “Non-perturbative effects and the refined topological string,” JHEP **1409**, 168 (2014) [arXiv:1306.1734 [hep-th]].

[51] C. A. Tracy, H. Widom, “Proofs of two conjectures related to the thermodynamic Bethe ansatz,” Commun. Math. Phys. **179**, 667-680 (1996). [solv-int/9509003].

[52] M. Aganagic, M. C. N. Cheng, R. Dijkgraaf, D. Krefl and C. Vafa, "Quantum Geometry of Refined Topological Strings," JHEP **1211**, 019 (2012) [arXiv:1105.0630 [hep-th]].

[53] N. A. Nekrasov and S. L. Shatashvili, "Quantization of Integrable Systems and Four Dimensional Gauge Theories," arXiv:0908.4052 [hep-th].

[54] Y. Hatsuda, S. Moriyama and K. Okuyama, "Exact Results on the ABJM Fermi Gas," JHEP **1210**, 020 (2012) [arXiv:1207.4283 [hep-th]].

[55] F. Calvo and M. Mariño, "Membrane instantons from a semiclassical TBA," JHEP **1305**, 006 (2013) [arXiv:1212.5118 [hep-th]].

[56] Y. Hatsuda, S. Moriyama and K. Okuyama, "Instanton Effects in ABJM Theory from Fermi Gas Approach," JHEP **1301**, 158 (2013) [arXiv:1211.1251 [hep-th]].

[57] M. Aganagic, V. Bouchard and A. Klemm, "Topological Strings and (Almost) Modular Forms," Commun. Math. Phys. **277**, 771 (2008) [hep-th/0607100].

[58] R. Gopakumar and C. Vafa, "M theory and topological strings. 2.," hep-th/9812127.

[59] M. Aganagic, A. Klemm, M. Mariño and C. Vafa, "The Topological vertex," Commun. Math. Phys. **254**, 425 (2005) [hep-th/0305132].

[60] Y. Hatsuda, S. Moriyama and K. Okuyama, "Instanton Bound States in ABJM Theory," JHEP **1305**, 054 (2013) [arXiv:1301.5184 [hep-th]].

[61] A. Grassi, Y. Hatsuda and M. Mariño, "Spectral determinants and quantization conditions in ABJ(M) theories," to appear.

[62] X. f. Wang, X. Wang and M. x. Huang, "A Note on Instanton Effects in ABJM Theory," arXiv:1409.4967 [hep-th].

[63] A. Grassi, Y. Hatsuda and M. Mariño, "Topological Strings from Quantum Mechanics," arXiv:1410.3382 [hep-th].

[64] S. Codesido, A. Grassi and M. Mariño, "Exact results in N=8 Chern-Simons-matter theories and quantum geometry," arXiv:1409.1799 [hep-th].

[65] D. L. Jafferis and A. Tomasiello, "A Simple class of N=3 gauge/gravity duals," JHEP **0810**, 101 (2008) [arXiv:0808.0864 [hep-th]].

[66] Y. Imamura and K. Kimura, "On the moduli space of elliptic Maxwell-Chern-Simons theories," Prog. Theor. Phys. **120**, 509 (2008) [arXiv:0806.3727 [hep-th]].

[67] I. K. Kostov, "Solvable statistical models on a random lattice," Nucl. Phys. Proc. Suppl. **45A**, 13 (1996) [hep-th/9509124].

[68] D. R. Gulotta, C. P. Herzog and S. S. Pufu, "From Necklace Quivers to the F-theorem, Operator Counting, and T(U(N))," JHEP **1112**, 077 (2011) [arXiv:1105.2817 [hep-th]].

[69] S. Moriyama and T. Nosaka, "Partition Functions of Superconformal Chern-Simons Theories from Fermi Gas Approach," arXiv:1407.4268 [hep-th].

[70] A. Dabholkar, N. Drukker and J. Gomes, "Localization in Supergravity and Quantum $\mathrm{AdS}_4/\mathrm{CFT}_3$ Holography," arXiv:1406.0505 [hep-th].

[71] H. Awata, S. Hirano and M. Shigemori, "The Partition Function of ABJ Theory," Prog. Theor. Exp. Phys. , 053B04 (2013) [arXiv:1212.2966].

[72] M. Honda, "Direct derivation of "mirror" ABJ partition function," JHEP **1312**, 046 (2013) [arXiv:1310.3126 [hep-th]].

[73] S. Matsumoto and S. Moriyama, "ABJ Fractional Brane from ABJM Wilson Loop," JHEP **1403**, 079 (2014) [arXiv:1310.8051 [hep-th]].

[74] M. Honda and K. Okuyama, “Exact results on ABJ theory and the refined topological string,” JHEP **1408**, 148 (2014) [arXiv:1405.3653 [hep-th]].

[75] M. Mariño, “Topological strings at strong string coupling,” talk given in november 2011 at the Banff workshop *New recursion formulae and integrability for Calabi–Yau spaces*, http://www.birs.ca/events/2011/5-day-workshops/11w5114/videos.

[76] M. x. Huang and X. f. Wang, “Topological Strings and Quantum Spectral Problems,” JHEP **1409**, 150 (2014) [arXiv:1406.6178 [hep-th]].

[77] R. Kashaev and M. Mariño, “Operators from mirror curves and the quantum dilogarithm,” arXiv:1501.01014 [hep-th].

[78] M. Mariño and S. Zakany, “Matrix models from operators and topological strings,” arXiv:1502.02958 [hep-th].

[79] I. K. Kostov, “Exact solution of the six vertex model on a random lattice,” Nucl. Phys. B **575**, 513 (2000) [hep-th/9911023].