Two-Dimensional Quantum Yang–Mills Theory and the Makeenko–Migdal Equations

Thierry Lévy

Introduction

These notes, echoing a conference given at the Strasbourg–Zurich seminar in October 2017, are written to serve as an introduction to 2-dimensional quantum Yang–Mills theory and to the results obtained in the last five to ten years about its so-called large N limit.

Quantum Yang–Mills theory, at least in the flavour that we will describe, combines differential geometric and probabilistic ideas. We would like to think, and hope to convince the reader, that this is less a complication than a source of beauty and enjoyment.

Some parts of our presentation will rely more distinctly on a probabilistic or a differential geometric background. We will however always try to keep technicalities aside and to favour explanation over demonstration. This is thus not, in the purest sense, a mathematical text: there will be essentially no proof. On the other hand, we will give fairly detailed examples of some computations that, we hope, are typical of the theory and illustrate it.

Slightly different in aim and content, but also introductory, the notes [26] written with four hands with Ambar Sengupta can serve as counterpoint, or complement, to the present text.

T. Lévy (✉)

Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université, Paris, France

e-mail: thierry.levy@sorbonne-universite.fr

These notes are split in three parts. In the first, we explain the nature of the Yang–Mills holonomy process, which is the main object of interest of the theory. We do it from two perspectives, one differential geometric, and one probabilistic. This leads us to the definition of Wilson loop expectations, which are the most important numerical quantities of the theory.

In the second part, we discuss several approaches to the computation of Wilson loop expectations, and illustrate them on several examples. The large N limit of the theory makes a first appearance in this section, and we derive by hand some concrete instances of the Makeenko–Migdal equations which are the subject of the third part. We also included in the second part a discussion of the holonomy process on the sphere, and of the Douglas–Kazakov phase transition.

In the third part, we describe the Makeenko–Migdal equations. In keeping with the style of these notes, we do not offer a proof of these equations, but we describe as carefully as we can Makeenko and Migdal's original derivation of them. Then, we discuss the amount of information carried by these equations and illustrate their power in the computation of the so-called master field, that is the large N limit of Wilson loop functionals.

1 Quantum Yang–Mills Theory on Compact Surfaces

1.1 The Holonomy Process and the Yang–Mills Action

The central object of study of quantum 2-dimensional Yang–Mills theory is a collection of random unitary matrices indexed by the class $\mathscr{L}_m(M)$ of Lipschitz continuous loops based at some point m on a compact surface M. This collection of random variables is called the *Yang–Mills holonomy process* and it is denoted by

$$(H_\ell)_{\ell \in \mathscr{L}_m(M)} \tag{1}$$

The idea of this collection of random variables arose, along a fairly convoluted path, from physical considerations relating to the description of certain kinds of fundamental interactions.[1] It is, fortunately, not necessary to be familiar with the original motivation of Yang and Mills to understand what the Yang–Mills holonomy process is.

In very broad terms, the basic data of the theory is a compact surface M (for example a disk, a sphere, a cylinder, a torus) and a compact matrix group G (for

---

[1]We will not describe this path, but indicate that it is marked by contributions of Chen Ning Yang and Robert Mills, the classical reference being [48], of Alexander Migdal, who in [32] provided mathematicians with a usable description of a crucial part of Yang–Mills theory, of Leonard Gross who initiated a school of mathematical study of the 2-dimensional Yang–Mills theory [13–15], of Bruce Driver and Ambar Sengupta, who finally gave in [6, 40] the first mathematically rigorous definitions of the Yang–Mills holonomy process. This enumeration is of course much too short not to leave many important contributions aside: a more extensive bibliography can for instance be found in [26].

example U(1), SO(3), U(N)). From this data, an infinite dimensional space of *connections* can be built[2], on which an infinite dimensional symmetry group, the *gauge group* acts[3], with infinite dimensional quotient, and one of the fundamental maps of the theory is the *holonomy map*

$$\{\text{connections}\}\big/\{\text{gauge group}\} \xrightarrow{\text{holonomy}} \text{Maps}(\mathscr{L}_m(M), G)\big/G$$

On the right-hand side, the action of G on the space of maps from $\mathscr{L}_m(M)$ to G is by conjugation. Leaving this action aside, note that the distribution of the holonomy process (1) is a probability measure on the space $\text{Maps}(\mathscr{L}_m(M), G)$. We will call this space the space of *holonomies*.

   One property that makes the holonomy map so important is that it is injective. It is thus legitimate to say that a connection is well described by its holonomy.

   Another fundamental map of the theory is the *Yang–Mills action* S_{YM} which is a non-negative functional traditionally defined on the space of connections, but that can also be defined on the space of holonomies, so that the situation is

$$\{\text{connections}\}\big/\{\text{gauge group}\} \xrightarrow{\text{holonomy}} \text{Maps}(\mathscr{L}_m(M), G)\big/G$$

$$S_{\text{YM}} \searrow \qquad \swarrow S_{\text{YM}}$$

$$[0, \infty]$$

$$(2)$$

The *Yang–Mills measure* is heuristically described as the Boltzmann probability measure, on the space of connections or on the space of holonomies, associated with the Yang–Mills action. The typical formula that one finds in the literature is

$$\mathrm{d}\mu_{\text{YM}}(\omega) = \frac{1}{Z}\, e^{-\frac{1}{2T} S_{\text{YM}}(\omega)}\,\mathrm{d}\omega \qquad (3)$$

where T is a positive real parameter called the coupling constant. Here, ω is meant to stand for a connection or for a holonomy, depending on one's preferred point of view. This expression is however plagued with difficulties: on the infinite dimensional spaces where the Yang–Mills measure is supposed to live, there is no Lebesgue-like reference measure that could reasonably play the role of $\mathrm{d}\omega$, and even if there were, one would not expect the Yang–Mills measure to be absolutely

---

[2]The exact nature of these connections can be ignored for the moment. If $G = \text{U}(1)$, they can be pictured as magnetic potentials on M.

[3]In physical terms, two connections related by a gauge transformation represent two magnetic potentials corresponding to the same magnetic field.

continuous with respect to it; moreover, because of the action of the gauge group, the most sensible value for the normalisation constant would be $Z = +\infty$; and one does finally not expect a typical ω in the sense of the Yang–Mills measure to be regular enough to have a finite Yang–Mills action.

One of the goals of the 2-dimensional quantum Yang–Mills theory is to find a way of sorting out these difficulties and to construct rigorously a probability measure that can honestly be called the Yang–Mills measure. The situation may look rather desperate, but it is uplifting to realise that after replacing the space of connections, or holonomies, by a space of real-valued functions on [0, 1] and the Yang–Mills action by the square of the Sobolev H^1 norm, the analogous problem is almost just as ill-posed but has a very well-known solution, namely the Wiener measure. The main difference between the Wiener and the Yang–Mills cases is the presence in the latter of the gauge symmetry. Symmetry can however be a nuisance or a guide, and it turns out to be possible, in Yang–Mills theory, to make gauge symmetry an ally rather than a foe.

We will now describe more precisely the three maps appearing in the diagram (2). The holonomy map and the Yang–Mills action on the space of connections are differential geometric in nature. We start by describing them, and then turn to the Yang–Mills action on the space of holonomies. It would be unfair to say that the content of Section 1.2 can safely be completely ignored: we will refer to it later, in particular in Section 3.2. However, it is certainly possible to skip it at first reading and to jump to Section 1.3.

1.2 The Yang–Mills Action: Connections

In this section, we assume from the reader some familiarity with the differential geometry of principal bundles. We give brief reminders of the main definitions, but this is of course not the place for a complete exposition. For details, and although some might find it too Bourbakist in style, we recommend the second chapter of the first volume of the classical opus by Kobayashi and Nomizu [21].

1.2.1 The Yang–Mills Action

Although we are concerned in this text with compact surfaces, we will describe the Yang–Mills action in the more general context of compact Riemannian manifolds of arbitrary dimension—this is not more difficult.

Let M be a compact connected Riemannian manifold. Let G be a compact Lie group with Lie algebra $\mathfrak{g}$. Assume that $\mathfrak{g}$ is endowed with a scalar product $\langle \cdot, \cdot \rangle$ that is invariant under the adjoint representation Ad : $G \to \mathrm{GL}(\mathfrak{g})$.[4] Let $\pi : P \to M$

---

[4]The typical example that we have in mind is $G = \mathrm{U}(N)$ and, for all $X, Y \in \mathfrak{u}(N)$ skew-Hermitian $N \times N$ matrices, $\langle X, Y \rangle = N\mathrm{Tr}(X^*Y)$.

be a principal G-bundle over M.[5] Let $\mathscr{A}$ denote the space of connections on P. It is an affine subspace of the space of $\mathfrak{g}$-valued differential 1-forms on P. For every connection $\omega \in \mathscr{A}$, the curvature of ω is the form $\Omega = d\omega + \frac{1}{2}[\omega \wedge \omega]$.[6] This $\mathfrak{g}$-valued 2-form on P vanishes on vertical vectors and is G-equivariant. It can thus be seen as a 2-form on M with values in the adjoint bundle $\mathrm{Ad}(P)$. Using the Hodge operator of the Riemannian structure of M, one can form the $(\mathrm{Ad}(P) \otimes \mathrm{Ad}(P))$-valued form of top degree $\Omega \wedge \star\Omega$ on M. Contracting this form with the Euclidean structure of $\mathrm{Ad}(P)$ induced by the invariant scalar product on $\mathfrak{g}$ yields the real-valued differential form of top degree $\langle \Omega \wedge \star\Omega \rangle$. This form can be integrated[7] to

---

[5]The manifold P is thus acted on, on the right, by G. For small open subsets U of M, the part $\pi^{-1}(U)$ of the manifold P that sits above U is equivariantly diffeomorphic to $U \times G$, with π being the first coordinate map and G acting by translations on the right on the second coordinate. A principal bundle is *trivial* if it is globally isomorphic to $M \times G$.

[6]This definition of the curvature is made slightly ambiguous by the coexistence, in the literature, of two different conventions regarding the definition of the exterior product and the exterior differential of differential forms. Since it took me some time to clarify this elementary point, I want to record it here, to the price of a rather long footnote.

The two conventions could be called 'simplicial' and 'cubical' according to their respective definitions of the exterior product of 1-forms:

$$(\alpha_1 \wedge \ldots \wedge \alpha_k)(X_1, \ldots, X_k) = \begin{cases} \frac{1}{k!} \det\left[(\alpha_i(X_j))_{i,\,j=1\ldots k}\right] & \text{(simplicial)} \\ \det\left[(\alpha_i(X_j))_{i,\,j=1\ldots k}\right] & \text{(cubical)} \end{cases}$$

Each convention is supported by illustrious authors, including, for the simplicial one, Kobayashi and Nomizu [21, p. 35] and Morita [33, Eq. (2.14) p. 70], and for the cubical one, Spivak [43, p. 203]. Since everyone agrees on the formula $d(\alpha \wedge \beta) = d\alpha \wedge \beta + (-1)^{\deg(\alpha)} {}^{\deg(\beta)} \alpha \wedge d\beta$, there must also be two competing definitions of the exterior differential. Specifically, the two definitions are related by the formula $d^{\mathrm{simplicial}}\alpha = \frac{1}{\deg(\alpha)+1} d^{\mathrm{cubical}}\alpha$ (compare, for instance, [21, p. 36] or [33, Thm. 2.9 p. 71] and [43, Thm 13 p. 213]). The formula $d\alpha(X, Y) = X\alpha(Y) - Y\alpha(X) - \alpha([X, Y])$, for instance, belongs to the cubical school.

Returning to the definition of the curvature, it has a different meaning with each convention, but fortunately, the simple relation $\Omega^{\mathrm{simplicial}} = \frac{1}{2}\Omega^{\mathrm{cubical}}$ holds. Let us be more explicit about this definition: the expression $\omega \wedge \omega$ is to be understood as a $\mathfrak{g} \otimes \mathfrak{g}$-valued 2-form, which is then composed by the Lie bracket to yield a $\mathfrak{g}$-valued 2-form. Explicitly, if X and Y are two vector fields defined on an open subset of P, then the curvature of ω is defined on this open set by

$$\Omega^{\mathrm{cubical}}(X, Y) = 2\Omega^{\mathrm{simplicial}}(X, Y) = X\omega(Y) - Y\omega(X) - \omega([X, Y]) + [\omega(X), \omega(Y)]$$

Note that there is universal agreement on what it means for the curvature to vanish.

Finally, since everyone also agrees that Stokes' formula is free of any coefficient, each convention on the definition of the exterior differential entails its own definition of the integral. This is slightly hidden by the fact that everyone agrees on the formula $\int_{[0,1]^n} dx_1 \wedge \ldots \wedge dx_n = 1$ (see [33, Sec. 3.2 (a), p. 104] and [43, Prop. 1 p. 247]), but it must be realised that the differential form that is denoted by $dx_1 \wedge \ldots \wedge dx_n$ is not the same for everyone. Specifically, the relation is $\int^{\mathrm{simplicial}} \alpha = \deg(\alpha)! \int^{\mathrm{cubical}} \alpha$.

Finally, there is agreement on the meaning of the curvature as a linear map from the space of smooth 2-chains in P to $\mathfrak{g}$.

[7]The definition of the Yang–Mills action seems to require an orientation of M. In fact, this orientation is used twice, once to define the Hodge dual $\star\Omega$ of Ω and once to integrate $\langle \Omega, \star\Omega \rangle$

yield the Yang–Mills action of ω:

$$S_{\mathsf{YM}}(\omega) = \frac{1}{2} \int_M \langle \Omega \wedge \star\Omega \rangle \tag{4}$$

In words, the Yang–Mills action of a connection is nothing more than one half of the squared L^2 norm of its curvature.[8]

Let us describe locally, in coordinates, the scalar function that is integrated over M to compute $S_{\mathsf{YM}}(\omega)$. For this, let us consider an open subset U of M on which there exist local coordinates $x_1, \ldots, x_n$ on M and over which P is trivial. Let us choose a section[9] $\sigma : U \to P$ of P over U. Let us define $A = \sigma^*\omega$. Then in the local coordinates on U, the 1-form A writes $A_1 \, dx_1 + \ldots + A_n \, dx_n$, where $A_1, \ldots, A_n$ are maps from U to $\mathfrak{g}$. Then $F = \sigma^*\Omega$ writes

$$F = \sum_{1 \leqslant i < j \leqslant n} \left(\partial_i A_j - \partial_j A_i + [A_i, A_j] \right) dx_i \wedge dx_j$$

and the contribution of U to the Yang–Mills action of ω is

$$\frac{1}{2} \int_U \langle \Omega \wedge \star\Omega \rangle = \frac{1}{2} \sum_{1 \leqslant i < j \leqslant n} \int_U \left\| \partial_i A_j - \partial_j A_i + [A_i, A_j] \right\|^2 \mathrm{dvol}(x)$$

where $\mathrm{dvol}(x)$ is the Riemannian volume measure on M, and $\| \cdot \|$ is the Euclidean norm on $\mathfrak{g}$ associated with the invariant scalar product $\langle \cdot, \cdot \rangle$. The analogy with the squared Sobolev H^1 norm should be even more obvious on this expression.

1.2.2 Gauge Transformations

The gauge group, that we denote by $\mathscr{J}$, is the group of G-equivariant diffeomorphisms of P over the identity of M.[10] It acts by pull-back on $\mathscr{A}$ and a

---

over M. Reversing the orientation changes the Hodge dual and the integral by a sign, so that if M is orientable, the definition of S_{YM} is independent of the choice of orientation of M. Moreover, if M is not orientable, S_{YM} can still be defined using a partition of unity.

[8]Considering that the curvature is a kind of derivative of the connection, the Yang–Mills action stands thus in close analogy with the squared H^1 norm of a real-valued function on $[0, 1]$.

[9]To say that σ is a section of P over U means that $\pi \circ \sigma = \mathrm{id}_U$. The existence of such a section is equivalent to the triviality of the restriction of P over U. In particular, the existence of a global section $\sigma : M \to P$ is equivalent to the triviality of the bundle $\pi : P \to M$. The reader who is more familiar with vector bundles than principal bundles might at first be surprised by this statement, since a vector bundle can admit a global section, even a non-vanishing one, without being trivial. However, the existence of a section for a principal bundle corresponds, for a vector bundle, to the existence of a basis of sections.

[10]An element j of the gauge group is a diffeomorphism $j : P \to P$ that leaves each fibre of P globally stable, and acts on it in a way that commutes with the action of G on the right on P. For the bundle $P = M \times G \to M$, the gauge group can be identified with $\mathscr{J} = C^\infty(M, G)$ acting pointwise on P by multiplication on the left on the second coordinate.

routine verification shows that it leaves S_{YM} invariant. Thus, the Yang–Mills action descends to a function

$$S_{\mathsf{YM}} : \mathscr{A}/\mathscr{G} \to [0, \infty)$$

the study of which is the subject of classical Yang–Mills theory.

Let us display the formulas which give, through a local section of P, the action of the gauge group on a connection and its curvature. These formulas are indeed useful, and ubiquitous in the literature. Let $j : P \to P$ be a gauge transformation. Let $\sigma : U \to P$ be a local section of P over an open subset U of M. Then there exists a unique function $g : U \to G$ such that for every $x \in U$, one has $j(\sigma(x)) = \sigma(x)g(x)$. Then, letting j act on a connection ω yields the new connection $j \cdot \omega = j^*\omega$ and transforms on one hand A into

$$A^g = \sigma^*(j \cdot \omega) = g^{-1}Ag + g^{-1}\,\mathrm{d}g$$

and on the other hand F into

$$F^g = g^{-1}Fg$$

This formula explains the invariance of the Yang–Mills action: without trying to be perfectly precise, one can say that the action of a gauge transformation conjugates the curvature at each point of M by some element of G, and thus leaves its Euclidean norm unchanged.

1.2.3   Some Questions of Classical Yang–Mills Theory

Let us mention, without giving any details, a few examples of the questions that arise in the study of the Yang–Mills action.

- The set $S_{\mathsf{YM}}^{-1}(0)$ is the *moduli space of flat connections*, that is, the quotient of the set of flat connections by the action of the gauge group. It is a finite-dimensional orbifold with a rich geometric structure, the study of which is both an old and an active subject of investigation [11, 12, 20, 28, 29, 45, 46].
- The Yang–Mills action can be understood as arising, through appropriate reformulation and generalisation, from a Lagrangian formulation of Maxwell's equations of the electromagnetic field. The critical points of the Yang–Mills action are thus of special interest: they are, in a sense, the classical physical fields of Yang–Mills theory. They are called *Yang–Mills connections* and a milestone in their study in the 2-dimensional case is [1].
- When M is 4-dimensional, the Yang–Mills action is conformally invariant, in the sense that it depends on the Riemannian metric on M only through its conformal class. There is an extensive literature devoted to Yang–Mills connections on 4-dimensional manifolds [18]. Looking for self-dual Yang–Mills connections on

$\mathbb{R}^4$ that are invariant by translation in two directions, for example, leads to the study of Hitchin equations and Higgs bundles [19].

- From a physical point of view, the Yang–Mills action of a connection is an appropriate measure of its non-triviality. From an analytical point of view, however, it turns out that a natural way of measuring a connection is its Sobolev H^1 norm.[11] The Yang–Mills action is controlled by the H^1 norm, but not conversely. A flat connection, that is, a connection with Yang–Mills action 0, can be given an arbitrarily large H^1 norm by an appropriate gauge transformation. A beautiful theorem of Karen Uhlenbeck states that level sets of the Yang–Mills action, that is, the sets of the form $\{S_{\mathsf{YM}} \leqslant c\}$, $c \in \mathbb{R}_+$, are sequentially weakly compact in H^1 up to gauge transformation: from any sequence of connections with bounded Yang–Mills action, one can extract a subsequence which, after suitable gauge transformation of each term, converges weakly in H^1 [44].
- The Yang–Mills action gives rise to a gradient flow, which formally is the solution of the differential equation $\partial_t \omega_t = -\nabla_{\omega_t} S_{\mathsf{YM}}$. This is the *Yang–Mills flow* [36]. There is currently an active investigation of stochastic perturbations of this flow in cases where M is 2- or 3-dimensional [4, 41].

1.2.4   The Holonomy Map

A fundamental construction associated with a connection is that of the *holonomy*, or *parallel transport*, that it induces. For every continuous and piecewise smooth curve $c : [0, 1] \to M$, the parallel transport along c determined by the connection ω is the G-equivariant mapping $\mathrm{hol}(\omega, c) : P_{c_0} \to P_{c_1}$ which to every point p of P_{c_0} associates the endpoint of the unique continuous curve $\tilde{c} : [0, 1] \to P$ such that $\tilde{c}_0 = p$, $\pi \circ \tilde{c} = c$ and for all $t \in [0, 1]$ at which c is differentiable, $\omega(\dot{\tilde{c}}_t) = 0$.

This parallel transport enjoys the following properties, which are of fundamental importance.

- It is unaffected by a change of parametrisation of the curve.
- If $c : [0, 1] \to M$ is a curve and c^{-1} denotes the same curve traced backwards, that is, $c_t^{-1} = c_{1-t}$, then $\mathrm{hol}(\omega, c^{-1}) = \mathrm{hol}(\omega, c)^{-1}$.
- If c and c' are two curves such that $c_1 = c_0'$, so that the concatenation cc' is well defined, then $\mathrm{hol}(\omega, cc') = \mathrm{hol}(\omega, c') \circ \mathrm{hol}(\omega, c)$.

It will be useful to understand a bit more concretely how this parallel transport can be computed, and how it gives rise to a holonomy in the sense that we gave to this word in Section 1.1.

Assume that the range of the curve c lies in an open subset U of M over which the fibre bundle P is trivial.[12] Let $\sigma : U \to P$ be a section of P over U. Set $A = \sigma^* \omega$.

---

[11] Here, we are talking about connections as elements of $\mathscr{A}$, not of the quotient $\mathscr{A}/\mathscr{J}$.

[12] If c does not lie in such an open subset, it can be split into finitely many pieces which do and the holonomy along c is simply the product of the holonomies along these shorter pieces.

It is a 1-form on U with values in $\mathfrak{g}$. The solution of the differential equation

$$\dot{h}_t = -A(\dot{c}_t)h_t, \quad h_0 = 1_G \tag{5}$$

is a curve $h : [0, 1] \to G$ which starts from the unit element 1_G. The endpoint of this curve computes the parallel transport along c determined by ω in the sense that

$$\mathrm{hol}(\omega, c)(\sigma(c_0)) = \sigma(c_1)h_1$$

This relation is illustrated in Figure 1.

Let us introduce the notation

$$\mathrm{hol}_\sigma(\omega, c) = h_1$$

the holonomy of ω along c read in the section σ. This object has the drawback of depending on the choice of a local section of the bundle, but the great advantage of being fairly concrete, namely an element of G, that is, in many situations, a matrix.

If $j \in \mathscr{J}$ is a gauge transformation of P, recall from Section 1.2.2 that $j \cdot \omega = j^*\omega$ is the pull-back of ω by the diffeomorphism j of P. Then the holonomy of $j \cdot \omega$ along c is related to that of ω by the relation

$$\mathrm{hol}(j \cdot \omega, c) = j^{-1}_{|P_{c_1}} \circ \mathrm{hol}(\omega, c) \circ j_{|P_{c_0}}$$

Through the local section $\sigma : U \to M$, and letting $g : U \to G$ be the function such that $j(\sigma(x)) = \sigma(x)g(x)$ for every $x \in U$, this relation takes the more explicit form

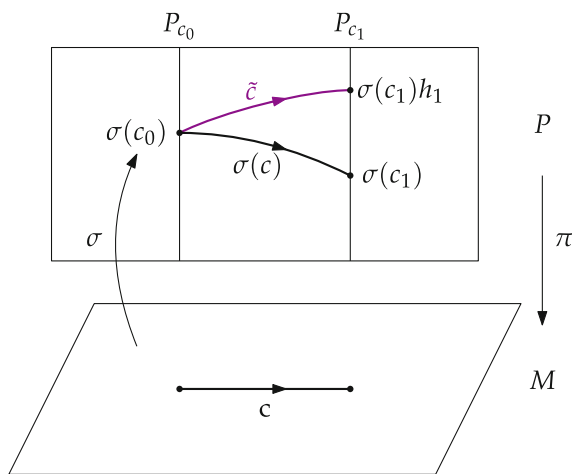$$\mathrm{hol}_\sigma(j \cdot \omega, c) = g^{-1}_{c_1}\mathrm{hol}_\sigma(\omega, c)g_{c_0} \tag{6}$$

Fig. 1 The difference between the horizontal lift of c starting at $\sigma(c_0)$, denoted in this picture by $\tilde{c}$, and $\sigma(c)$, the image of c by the local section σ, is measured by the function h which solves (5)

It follows from (6) that for all loop ℓ on M, that is, all curve which ends at its starting point, the conjugacy class of $\mathrm{hol}_\sigma(\omega, \ell)$ is not affected[13] by a gauge transformation of ω.

More generally, given a base point m on M, and denoting by $\mathscr{L}_m^\infty(M)$ the class of piecewise smooth loops on M based at m, the orbit of

$$(\mathrm{hol}_\sigma(\omega, \ell) : \ell \in \mathscr{L}_m^\infty(M)) \in \mathrm{Maps}(\mathscr{L}_m^\infty(M), G)$$

under the action of G by simultaneous conjugation is not affected by a gauge transformation of ω. This explains how a connection modulo gauge transformation defines a holonomy modulo conjugation.

The following result makes precise the statement that the horizontal arrow of (2) is injective.

Theorem 1.1 *Let m be a point of M. Let σ be a section of P in a neighbourhood of m. For any two connections ω and ω' on P, the following assertions are equivalent.*

1. *There exists a gauge transformation $j \in \mathscr{J}$ such that $j \cdot \omega = \omega'$.*
2. *There exists $g \in G$ such that for all loop $\ell \in \mathscr{L}_m^\infty(M)$, the equality $\mathrm{hol}_\sigma(\omega', l) = g^{-1}\mathrm{hol}_\sigma(\omega, l)g$ holds.*

1.3 The Yang–Mills Action: Holonomies

We will now give an alternative of the Yang–Mills action that is less classical and, most importantly, specific to the 2-dimensional case. To give an idea of the nature of this second description, let us pursue the analogy with the Wiener measure and the Sobolev H^1 norm. Consider a smooth function $b : [0, 1] \to \mathbb{R}$ with $b(0) = 0$. The squared H^1 norm of b can be expressed at least in the following two ways:

$$\|b\|_{H^1}^2 = \int_0^1 |\dot{b}(t)|^2 \, \mathrm{d}t = \sup_{0 \leqslant t_0 < t_1 < \ldots < t_n \leqslant 1} \sum_{k=1}^n \frac{|b(t_k) - b(t_{k-1})|^2}{t_k - t_{k-1}} \tag{7}$$

The integral expression corresponds to the description of the Yang–Mills action that we gave in the last section and is very similar to (4). We will now give another description, similar to the second, more combinatorial one.

---

[13]Incidentally, this class does not depend on the local section σ either.

1.3.1   Holonomies

The main algebraic property of the holonomy of a connection, already mentioned in Section 1.2.4, is that it is a *multiplicative* map from $\mathscr{L}_m^\infty(M)$ to G. Let us formulate this in a slightly different way.

Recall that M is a compact Riemannian manifold and G a compact Lie group. Let $\mathscr{P}(M)$ denote the set of all Lipschitz continuous[14] paths on M, two paths being identified if they differ only by an increasing change of parametrisation. Let us call a function $h : \mathscr{P}(M) \to G$ *multiplicative* if it satisfies the following two properties.

- For all path c, letting c^{-1} denote the same path traced backwards, one has $h(c^{-1}) = h(c)^{-1}$.
- For all paths c and c' such that c finishes where c' starts, so that the concatenated path cc' is defined, one has $h(cc') = h(c')h(c)$.

More generally, given a subset P of $\mathscr{P}(M)$, we say that a function $h : P \to G$ is multiplicative if it satisfies the above conditions whenever all the paths involved belong to the subset P.

Let us denote by $\mathrm{Mult}(\mathscr{P}(M), G)$ (resp. by $\mathrm{Mult}(P, G)$) the subset of $\mathrm{Maps}(\mathscr{P}(M), G)$ (resp. of $\mathrm{Maps}(P, G)$) formed by all multiplicative maps.

There is an action of the *gauge group* $\mathrm{Maps}(M, G)$ on $\mathrm{Mult}(\mathscr{P}(M), G)$ defined as follows. Consider $g : M \to G$ and a multiplicative map $h : \mathscr{P}(M) \to G$. For all path c starting at c_0 and finishing at c_1, define

$$(g \cdot h)(c) = g_{c_1}^{-1} h(c) g_{c_0} \tag{8}$$

an equation that should be compared with (6). It is not difficult to check that the map $g \cdot h$ is still multiplicative.

Let m be a point of M. A multiplicative function can be restricted to $\mathscr{L}_m(M)$ and the action of $\mathrm{Maps}(M, G)$ on this restricted map reduces to the action of G by conjugation. The following fact may seem surprising at first sight, but it is not difficult to prove.

Proposition 1.2 *For all $m \in M$, the restriction map*

$$\mathit{Mult}(\mathscr{P}(M), G)\big/\mathit{Maps}(M, G) \longrightarrow \mathrm{Mult}(\mathscr{L}_m(M), G)\big/G$$

is a bijection.

---

[14]In this text, we consider alternatively paths that are piecewise smooth and paths that are Lipschitz continuous. We do so for reasons of technical convenience, and the reader should not be overly worried by what can safely be regarded as a secondary issue.

We call either side of this bijection the space of *holonomies*. Thanks to the multiplicativity and the gauge symmetry, a holonomy can either be seen as a group-valued function on the set of all paths, or on the set of all loops based at some reference point m on M.

1.3.2   Graphs on Surfaces

We will now assume that M is a 2-dimensional manifold: it is thus a compact surface. We announced an expression of the Yang–Mills action similar to the rightmost term of (7): the role of subdivisions of the interval [0, 1] will be played by *graphs* on M. This will be the occasion of a first encounter with this notion that is central to the construction of the 2-dimensional Yang–Mills measure.

Let us call *edge* an element of $\mathscr{P}(M)$ that is injective — note that this does not depend on the way in which the path is parameterised. A *graph* is a finite set of edges, stable by the reversal map $e \mapsto e^{-1}$, and in which any two edges either form a pair $\{e, e^{-1}\}$, or meet, if at all, at some of their endpoints.

The *vertices* of a graph are the endpoints of its edges. The *faces* of a graph are the connected components of the complement in M of the union of its edges. A graph is conveniently described as a triple $\mathbb{G} = (\mathbb{V}, \mathbb{E}, \mathbb{F})$ consisting of a set of vertices, a set of edges and a set of faces, but it is in fact entirely determined by the set $\mathbb{E}$ of its edges.

A crucial additional assumption is that every face of a graph must be homeomorphic to a disk. This guarantees that the 1-skeleton of the graph correctly represents the topology of the surface, to the extent that a 1-dimensional object can represent a 2-dimensional one.

1.3.3   The Yang–Mills Action

Let $\mathbb{G}$ be a graph on our compact surface M. We will denote by $\mathscr{P}(\mathbb{G})$ the set of paths that can be constructed as concatenations of edges of $\mathbb{G}$. To each face F of $\mathbb{G}$, we can associate in an almost unequivocal way a loop ∂F that winds exactly once around F. To give a perfectly rigorous definition of this loop is less simple than one might expect, but there is nothing counterintuitive in it. It is only *almost* well defined because there is no preferred starting point for this loop. However, if $f : \mathscr{P}(\mathbb{G}) \to G$ is a multiplicative function, then the conjugacy class of the element $h(\partial F)$ of G is well defined. In particular, the Riemannian distance, in G, between the element $h(\partial F)$ and the unit element 1_G, is well defined.[15] This distance is, moreover, not affected by the action of an element of the gauge group Maps(M, G) on h.

---

[15]This distance is defined by the bi-invariant Riemannian metric on G associated with the invariant scalar product chosen on its Lie algebra, see the first lines of Section 1.2.1.

We can now define the Yang–Mills action on the space of holonomies by setting, for all $h \in \mathrm{Mult}(\mathscr{P}(M), G)$,

$$S_{\mathsf{YM}}(h) = \sup \left\{ \sum_{F \in \mathbb{F}} \frac{d_G(1_G, h(\partial F))^2}{\mathrm{area}(F)} \; : \; \mathbb{G} \text{ graph on } M \right\} \tag{9}$$

where the area of a face F is computed using the Riemannian structure on M.

It is manifest on this expression that, in the case where M is a surface, the only part of the Riemannian structure on M that is used in the definition of the Yang–Mills action is the Riemannian volume, in this case the Riemannian area. This is of course also true, be it in a slightly less apparent way, of the definition (4).

Proposition 1.3 *Assume that M is 2-dimensional. Then the definitions (4) and (9) of the Yang–Mills action agree. More precisely, for every connection ω inducing a holonomy h, the equality $S_{\mathsf{YM}}(\omega) = S_{\mathsf{YM}}(h)$ holds.*

1.4 The Yang–Mills Holonomy Process

We will now explain how to construct the Yang–Mills holonomy process. Although the definition of this process is derived, at a heuristic level, from the Yang–Mills action, the process and the action are logically unrelated. We can thus start afresh, from a compact surface M on which we have a Riemannian structure, or at least a measure of area, and a compact Lie group G, on the Lie algebra of which we have an invariant scalar product.

1.4.1 The Configuration Space of Lattice Yang–Mills Theory

One piece of information that we need to retain from the previous sections is the notion of graph on our surface M (see Section 1.3.2). Let us choose a graph $\mathbb{G} = (\mathbb{V}, \mathbb{E}, \mathbb{F})$ on M. The *configuration space* associated with a graph $\mathbb{G}$ on our surface M is the manifold

$$\mathscr{C}_{\mathbb{G}} = \{g = (g_e)_{e \in \mathbb{E}} \in G^{\mathbb{E}} : \forall g \in G, g_{e^{-1}} = g_e^{-1}\} = \mathrm{Mult}(\mathbb{E}, G)$$

of all ways of assigning an element of G to each oriented edge, in a way that is consistent with the orientation reversal.

Recall that we denote by $\mathscr{P}(\mathbb{G})$ the set of paths that can be constructed as concatenations of edges of $\mathbb{G}$. The configuration space $\mathscr{C}_{\mathbb{G}}$ is naturally in one-to-one correspondence with the set $\mathrm{Mult}(\mathscr{P}(\mathbb{G}), G)$ of all multiplicative maps from $\mathscr{P}(\mathbb{G})$ to G.

Choosing an *orientation* of $\mathbb{G}$, that is, a subset $\mathbb{E}^+ \subset \mathbb{E}$ containing exactly one element in each pair $\{e, e^{-1}\}$ allows one to realise the configuration space in the slightly less canonical, but easier to handle, way

$$\mathscr{C}_{\mathbb{G}} = G^{\mathbb{E}^+}$$

This makes it easy, for instance, to endow $\mathscr{C}_{\mathbb{G}}$ with a probability measure, namely the Haar measure on $G^{\mathbb{E}^+}$. The invariance of the Haar measure on the compact group G under the inverse map $x \mapsto x^{-1}$ implies that this measure on $\mathscr{C}_{\mathbb{G}}$ does not depend on the choice of orientation. We denote it by dg.

Every path $c \in \mathscr{P}(\mathbb{G})$ can be uniquely written as a concatenation of edges $c = e_1^{\epsilon_1} \ldots e_n^{\epsilon_n}$ with $e_1, \ldots, e_n \in \mathbb{E}^+$ and $\epsilon_1, \ldots, \epsilon_n \in \{-1, 1\}$. To such a path $c = e_1^{\epsilon_1} \ldots e_n^{\epsilon_n}$ we associate a holonomy map

$$h_c : \mathscr{C}_{\mathbb{G}} \longrightarrow G \qquad\qquad (10)$$
$$g \longmapsto g_{e_n}^{\epsilon_n} \cdots g_{e_1}^{\epsilon_1}$$

Our goal is to endow the configuration space $\mathscr{C}_{\mathbb{G}}$ with an interesting probability measure, so as to make the collection of maps $(h_c)_{c \in \mathscr{P}(\mathbb{G})}$ into a collection of G-valued random variables.

1.4.2   The Driver–Sengupta Formula

In order to define this probability measure, we need to introduce the heat kernel on G, or more accurately the fundamental solution of the heat equation. The invariant scalar product on the Lie algebra $\mathfrak{g}$ determines a bi-invariant Riemannian structure on G, and a Laplace-Beltrami operator Δ. We consider the function $p : \mathbb{R}_+^* \times G \to \mathbb{R}_+^*$ that is the unique positive solution of the heat equation $(\partial_t - \frac{1}{2}\Delta)p = 0$ with initial condition $p(t, x)\, dx \Rightarrow \delta_{1_G}$ as $t \to 0$. We use the notation $p_t(x) = p(t, x)$. A crucial property of this function is that, for all $t > 0$ and all $x, y \in G$, we have $p_t(yxy^{-1}) = p_t(x)$. We refer to this property as the *invariance under conjugation* of the heat kernel.

We mentioned at the end of Section 1.3.3 that, in the 2-dimensional setting, the Yang–Mills action depends on a Riemannian structure of the surface M only through the Riemannian area that it induces. We will denote by $|F|$ the area of a Borel subset F of M.

Given a face F of our graph, recall that we denote by ∂F a path that goes once around this face in the positive direction. Recall also that this path is ill-defined because there is no preferred vertex on the boundary of F from which to start it. However, this indeterminacy only results in an indeterminacy *up to conjugation* for the holonomy map $h_{\partial F}$. Thanks to the invariance under conjugation of the heat kernel, the function $g \mapsto p_t(h_{\partial F}(g))$ is still well defined on $\mathscr{C}_{\mathbb{G}}$ for every $t > 0$.

We can now write the formula which is the basis of the definition of the 2-dimensional Yang–Mills measure. It is due to Bruce Driver in the case where M is the plane, or a disk, and to Ambar Sengupta when M is an arbitrary compact surface. Recall that T is a positive real parameter of the measure. We define, on $\mathscr{C}_{\mathbb{G}}$, the probability measure

$$\mathrm{d}\mu_{\mathsf{YM}}^{\mathbb{G},T}(g) = \frac{1}{Z(\mathbb{G},T)} \prod_{F \in \mathbb{F}} p_{T|F|}(h_{\partial F}(g))\, \mathrm{d}g \qquad \text{(DS)}$$

Here, $Z(\mathbb{G},T)$ is the normalisation constant that makes $\mu_{\mathsf{YM}}^{\mathbb{G},T}$ a probability measure on $\mathscr{C}_{\mathbb{G}}$.

The gauge group $\mathrm{Maps}(\mathbb{V}, G)$ acts on the configuration space $\mathscr{C}_{\mathbb{G}}$ by a formula analogous to (8), and the measure $\mu_{\mathsf{YM}}^{\mathbb{G},T}$ is invariant under this action. Indeed, this action preserves the reference measure $\mathrm{d}g$ and transforms the holonomy along loops, in this case along boundaries of faces, by conjugation, which leaves the value of the fundamental solution of the heat equation on these holonomies unchanged.[16]

---

[16]Let us say a word about the way in which the presence of a boundary to the surface M should be taken into account in (DS), and how to treat the case where M is not orientable. The only place where we used the orientability and orientation of M is when we defined the boundary of a face as a loop winding *positively* around M. However, since the heat kernel also enjoys the invariance property $p_t(x) = p_t(x^{-1})$, it does not matter which orientation we choose around each face of the graph. Thus, (DS) is valid without any modification on a non-orientable surface.

In the case where M has a boundary, this boundary is a finite union of circles. Our assumption that each face of a graph is homeomorphic to a disk implies that each of these circles is a path in any graph on M. In this case, (DS) still makes sense and corresponds to free boundary conditions along the boundary of M. Fixed boundary conditions can be imposed: it is possible to insist that the holonomy along each boundary component belongs to a specific conjugacy class in G. If we wish to set the boundary condition for which the holonomy along a boundary component $c = e_1 \ldots e_n$ belongs to a conjugacy class C of G, the basic ingredient is the unique probability measure $\nu_{n,C}$ on $\mathcal{O}_{n,C} = \{(x_1, \ldots, x_n) \in G^n : x_n \ldots x_1 \in C\}$ invariant under the transitive action of G^n given by

$$(y_1, \ldots, y_n) \cdot (x_1, \ldots, x_n) = (y_1 x_1 y_n^{-1}, y_2 x_2 y_1^{-1}, \ldots, y_n x_n y_{n-1}^{-1})$$

This measure is easily described by the formula

$$\int_{\mathcal{O}_{n,C}} f\, \mathrm{d}\nu_{n,C} = \int_{G^n} f(x_1, \ldots, x_{n-1}, x_n z x_n^{-1} x_1^{-1} \ldots x_{n-1}^{-1})\, \mathrm{d}x_1 \ldots \mathrm{d}x_n$$

for an arbitrary $z \in C$. The way in which (DS) should be modified is that the uniform measure on $\mathscr{C}$ should be replaced, for the edges lying on the boundary of M, by the appropriate copy of a measure of the form $\nu_{n,C}$.

1.4.3   Invariance Under Subdivision

Starting from a graph $\mathbb{G}$ on our surface M, we built the configuration space $\mathscr{C}_{\mathbb{G}}$ and endowed, thanks to the Driver–Sengupta formula, this space with a probability measure, the lattice 2-dimensional Yang–Mills measure on $\mathbb{G}$. In doing so, we automatically produced a collection

$$(h_c)_{c\in\mathscr{P}(\mathbb{G})} \ \text{ or } \ (h_\ell)_{\ell\in\mathscr{L}_m(\mathbb{G})}$$

of G-valued random variables.[17]

The property of this construction that makes it so extremely pleasant is the fact that it is *invariant under subdivision*.

To articulate this fundamental property, let us say that a graph $\mathbb{G}_2$ is *finer* than a graph $\mathbb{G}_1$ if $\mathbb{G}_2$ can be obtained from $\mathbb{G}_1$ by subdividing and adding edges. More precisely, $\mathbb{G}_2$ is finer than $\mathbb{G}_1$ if $\mathbb{E}_1 \subset \mathscr{P}(\mathbb{G}_2)$: each edge of $\mathbb{G}_1$ is a path in $\mathbb{G}_2$. When this happens, there is a natural map

$$\mathscr{C}_{\mathbb{G}_2} \longrightarrow \mathscr{C}_{\mathbb{G}_1}$$
$$g^{(2)} \longmapsto \left(h_e^{(2)}(g^{(2)})\right)_{e\in\mathbb{E}_1}$$

where each edge e of $\mathbb{G}_1$ is seen as a path in $\mathbb{G}_2$ and thus assigned a holonomy by the configuration $g^{(2)}$.

The main result of 2-dimensional lattice Yang–Mills theory is the following.

Theorem 1.4 *Let $\mathbb{G}_1$ and $\mathbb{G}_2$ be two graphs on M. Assume that $\mathbb{G}_2$ is finer than $\mathbb{G}_1$. Then for all $T > 0$, the equality $Z(\mathbb{G}_1, T) = Z(\mathbb{G}_2, T)$ holds and the push-forward of the measure $\mu_{YM}^{\mathbb{G}_2,T}$ by the natural map $\mathscr{C}_{\mathbb{G}_2} \to \mathscr{C}_{\mathbb{G}_1}$ is the measure $\mu_{YM}^{\mathbb{G}_1,T}$.*

This theorem is so important that we are going to give an idea of the mechanism of its proof.

Proof The first observation is that one can always go from a graph to a finer graph by an appropriate succession of elementary operations consisting either in adding a new vertex in the middle of an existing edge or in adding a new edge between two existing vertices. We need to understand why neither of these elementary operations affect the partition function, nor transform essentially the measure.

The subdivision of an edge e into two new edges e' and e'' amounts, in the integral defining the partition function and in the expression defining the discrete Yang–Mills measure, to the replacement of every occurrence of the integration variable g_e by the product of the two new variables $g_{e''}g_{e'}$. The invariance by translation of the Haar measure ensures that this does not affect the result of any computation.

---

[17]Thanks to the multiplicativity of the holonomy and the gauge invariance of the construction of the lattice Yang–Mills measure, the point of view of a collection of random variables indexed by all paths in $\mathbb{G}$ or by the set of loops based at a specific reference point is equivalent, see Proposition 1.2.

The case of the addition of a new edge is more interesting. This edge e splits a face F into two faces F_1 and F_2, the boundaries of which are of the form ea and be^{-1} for some paths a and b. Observe that ba is a loop going along the boundary of F. In the computation of the partition function of the Yang–Mills measure on the finer graph, or of the integral of any functional on the configuration space of the coarser graph with respect to the image of the discrete Yang–Mills measure on the finer graph, we find an integral of a product of many factors, among which the two factors

$$p_{T|F_1|}\big(h_a(g)g_e\big)\, p_{T|F_2|}\big(g_e^{-1}h_b(g)\big)$$

contain the only two occurrences of the integration variable g_e. We can thus easily integrate with respect to g_e, using the convolution property of the heat kernel, namely the equality $p_t * p_s = p_{t+s}$, to find these two factors replaced by

$$p_{T(|F_1|+|F_2|)}\big(h_a(g)h_b(g)\big) = p_{T|F|}\big(h_{ba}(g)\big) = p_{T|F|}\big(h_{\partial F}(g)\big)$$

We are thus left with the partition function, or the integral of our functional, relative to the coarser graph.   □

The partition function $Z(\mathbb{G}, T)$, which is now promoted to a function of T alone, is a very interesting object. Let us give without proof an expression of this function. We use the notation $[a, b] = aba^{-1}b^{-1}$ for the commutator of two elements a and b of G.

Proposition 1.5 *Assume that M is a surface of genus g without boundary. Then for all $T > 0$, the partition function of the 2-dimensional Yang–Mills theory on M is given by*

$$Z_M(T) = \int_{G^{2g}} p_{T|M|}([a_1, b_1] \ldots [a_g, b_g])\, \mathrm{d}a_1 \mathrm{d}b_1 \ldots \mathrm{d}a_g \mathrm{d}b_g$$

1.4.4 The Continuum Limit

Up to some conceptually inessential but technically annoying complications, the invariance by subdivision of the discrete theory allows one to take the limit of the discrete measures as the graphs on the surface become infinitely fine. The technical complications have to do with the fact that, because two edges of two distinct graphs can intersect in a rather pathological way, it is not always true that given two graphs, there exists a third graph that is finer than these two graphs. The net effect of this complication is the persistence, in the theorem asserting the existence and uniqueness of the Yang–Mills holonomy process, of a continuity condition. We say that a sequence of paths $(c_n)_{n \geqslant 1}$ on M *converges* to a path c *with fixed endpoints* if all paths $c, c_1, c_2, \ldots$ start at the same point and finish at the same (possibly

different) point, and if the sequence of the paths $(c_n)_{n \geqslant 1}$ parameterised at unit speed converges uniformly to c.

Theorem 1.6 (The Yang–Mills holonomy process, [23, 40]) *Let M be a compact surface endowed with a smooth[18] measure of area. Let G be a compact Lie group, the Lie algebra of which is endowed with an invariant scalar product. There exists a collection of G-valued random variables $(H_c)_{c \in \mathscr{P}(M)}$ such that*

- *for every graph $\mathbb{G} = (\mathbb{V}, \mathbb{E}, \mathbb{F})$, the distribution of $(H_e)_{e \in \mathbb{E}}$ is the measure $\mu_{YM}^{\mathbb{G}, T}$,*
- *whenever a sequence $(c_n)_{n \geqslant 1}$ of paths converges with fixed endpoints to a path c, the sequence of random variables $(H_{c_n})_{n \geqslant 1}$ converges in probability to H_c.*

Moreover, any two collections of G-valued random variables with these properties have the same distribution.

The Yang–Mills holonomy process $(H_c)_{c \in \mathscr{P}(M)}$ is invariant in distribution under the action of the gauge group. This means that for every function $g : M \to G$, the following equality in distribution holds:

$$\left(g(\overline{c})^{-1} H_c g(\underline{c}) \right)_{c \in \mathscr{P}(M)} \overset{(d)}{=} (H_c)_{c \in \mathscr{P}(M)} \tag{11}$$

where $\underline{c}$ and $\overline{c}$ denote respectively the starting and finishing point of a path c. In particular, the distribution of H_c is uniform on G for every path c that is not a loop. Of course, this huge collection of uniform random variables is correlated in a complicated way, in particular to allow the random variables associated with loops to have non-uniform distributions.

The holonomy process also enjoys a property of invariance under area-preserving maps of M: if $\phi : M \to M$ is an area-preserving diffeomorphism, then ϕ preserves the class $\mathscr{P}(M)$ and the family $(H_{\phi(c)})_{c \in \mathscr{P}(M)}$ has the same distribution as the family $(H_c)_{c \in \mathscr{P}(M)}$. This is because the Driver–Sengupta formula depends only on the combinatorial structure of the graph under consideration, and on the areas of its faces. This is consistent with the fact that the Yang–Mills action, which we originally defined on a Riemannian manifold by (4), depends, if the manifold is 2-dimensional, on the Riemannian structure only through the Riemannian area. We already mentioned this important point in relation with the expression (9) of the Yang–Mills action.

---

[18]By a smooth measure, we mean a measure that admits a smooth positive density with respect to the Lebesgue measure in any coordinate chart.

1.4.5   The Structure of the Holonomy Process

The structure of the Yang–Mills holonomy process can be described fairly concretely provided one understands the structure of the set of loops on a graph.

Let us consider a graph $\mathbb{G}$ on M and a vertex m of this graph. We denote naturally by $\mathscr{L}_m(\mathbb{G})$ the set of loops in $\mathbb{G}$ based at m. The operation of concatenation makes $\mathscr{L}_m(\mathbb{G})$ a monoid, with unit element the constant loop at m. Each element ℓ of this monoid has an 'inverse' ℓ^{-1}, but it is not true, unless ℓ is already the constant loop, that $\ell\ell^{-1}$ is the constant loop. In order to make $\mathscr{L}_m(\mathbb{G})$ a group, into which ℓ^{-1} is truly the inverse of ℓ, it is natural to introduce on it the *backtracking equivalence* relation, for which two loops are equivalent if one can go from one to the other by successively erasing or inserting sub-loops of the form ee^{-1}, where e is an edge of the graph.

Each equivalence class of loops contains a unique loop of shortest length, which is also the unique *reduced* loop in this class, where by a reduced loop we mean one without any sub-loop of the form ee^{-1}.

Moreover, concatenation is compatible with this equivalence relation and the quotient monoid is a group. This quotient monoid can be more concretely described as the set $\mathscr{L}_m^{\mathrm{red}}(\mathbb{G})$ of reduced loops endowed with the operation of concatenation-followed-by-reduction.

With this group of reduced loops in hand, we can make several observations.

- Each element g of the configuration space $\mathscr{C}_{\mathbb{G}}$ induces, by the holonomy map, a map $\mathscr{L}_m^{\mathrm{red}}(\mathbb{G}) \to G$, which sends a loop ℓ to $h_\ell(g)$. This map is a group homomorphism, and the map

$$\mathscr{C}_{\mathbb{G}} \longrightarrow \mathrm{Hom}(\mathscr{L}_m^{\mathrm{red}}(\mathbb{G}), G)$$

  is onto. Moreover, it descends to a bijection

$$\mathscr{C}_{\mathbb{G}}/\mathrm{Maps}(\mathbb{V}, G) \xrightarrow{\sim} \mathrm{Hom}(\mathscr{L}_m^{\mathrm{red}}(\mathbb{G}), G)/G$$

  where the action on the left is that of the gauge group, and on the action on the right is that of G by conjugation.
- Let Γ denote the 1-skeleton of the graph, that is, the union of the ranges of its edges. The map $\mathscr{L}_m^{\mathrm{red}}(\mathbb{G}) \to \pi_1(\Gamma, m)$ which simply sends a reduced loop to its homotopy class is an isomorphism.
- The group $\mathscr{L}_m^{\mathrm{red}}(\mathbb{G})$, being isomorphic to the fundamental group of a graph, or of a 1-dimensional complex, is a free group. The rank of this group is equal to $|\mathbb{E}| - |\mathbb{V}| + 1 = |\mathbb{F}| - \chi(M) + 1 = |\mathbb{F}| + 2g - 1$, where $\chi(M)$ is the Euler characteristic of M and g its genus.

It is useful to recognise that the free group $\mathscr{L}_m^{\mathrm{red}}(\mathbb{G})$ admits nice bases.[19] Let us call *lasso* around a face F of $\mathbb{G}$ any loop of the form $c.\partial F.c^{-1}$, where c is a path from m to a vertex on the boundary of F, and ∂F is a loop going once around F.

It is now quite easy to describe the holonomy process. Let us begin with the case of the plane, or the disk.

Proposition 1.7 *Assume that M is a disk or the plane. Let $\mathbb{G}$ be a graph on M. The free group $\mathscr{L}_m^{\mathrm{red}}(\mathbb{G})$ admits a basis $\{\lambda_F : F \in \mathbb{F}\}$ such that*

- *for each face F, the loop λ_F is a lasso around F,*
- *under the lattice Yang–Mills measure $\mu_{YM}^{\mathbb{G},T}$, the random variables $(H_{\lambda_F} : F \in \mathbb{F})$ are independent, each H_{λ_F} being distributed according to the measure $p_{T|F|}(g)\,\mathrm{d}g$.*

In a sense, the holonomy process has independent increments distributed according to the fundamental solution of the heat equation: it can be described as a 'Brownian motion on G indexed by loops' on the disk, or on the plane. The role of time is played by area, and increments occur along faces of the graph, or lassos, instead of intervals of time.

In the case of a closed surface, the situation is slightly different. In this case, the most natural presentation of the group $\mathscr{L}_m^{\mathrm{red}}(\mathbb{G})$ is not as a free group (which it is), but with one generator too many, and one relation.

Proposition 1.8 *Assume that M is a closed surface of genus g. Let $\mathbb{G}$ be a graph on M. Set $r = |\mathbb{F}|$. The free group $\mathscr{L}_m^{\mathrm{red}}(\mathbb{G})$ admits a presentation*

$$\mathscr{L}_m^{\mathrm{red}}(\mathbb{G}) = \langle \lambda_{F_1}, \ldots, \lambda_{F_r}, a_1, b_1, \ldots, a_g, g_b \mid [a_1, b_1] \ldots [a_g, b_g] = \lambda_{F_1} \ldots \lambda_{F_r} \rangle$$

where

- *the loops $\lambda_{F_1}, \ldots, \lambda_{F_r}$ are lassos around the r faces of $\mathbb{G}$,*
- *the homotopy classes of the loops $a_1, b_1, \ldots, a_g, b_g$ generate $\pi_1(M, m)$,*
- *for every test function $f : G^{2g+r} \to \mathbb{C}$, one has*

$$\int_{\mathscr{C}} f(H_{\lambda_1}, \ldots, H_{\lambda_r}, H_{a_1}, H_{b_1}, \ldots, H_{a_g}, H_{b_g})\,\mathrm{d}\mu_{YM}^{\mathbb{G},T} \tag{12}$$

$$= Z_M(T)^{-1} \int_{G^{2g+r-1}} f(z_1, \ldots, z_{r-1}, z_r, x_1, y_1, \ldots, x_g, y_g) p_{T|F_1|}(z_1) \ldots$$

$$p_{T|F_r|}(z_r)\,\mathrm{d}z_1 \ldots \mathrm{d}z_{r-1}\,\mathrm{d}x_1\,\mathrm{d}y_1 \ldots \mathrm{d}x_g\,\mathrm{d}y_g$$

where in the last integral, z_r stands for

---

[19] Recall that a free group admits bases, that is, subsets by which it is freely generated. Any two bases have the same cardinality, called the rank of the group. Any subgroup of a free group is itself a free group, but the rank of a subgroup can be larger than the rank of the group. In fact, the free group of rank 2 contains subgroups of arbitrary finite or (countably) infinite rank.

$$z_r = (z_{r-1} \ldots z_1 [a_g, b_g] \ldots [a_1, b_1])^{-1}$$

Let us try to spell out the probabilistic content of this result. The presentation of the group $\mathscr{L}_m^{\mathrm{red}}(\mathbb{G})$ that we chose splits it into a homotopically trivial part, giving rise to the random variables $H_{\lambda_1}, \ldots, H_{\lambda_r}$, and a system of generators of the fundamental group of M, associated with the random variables $H_{a_1}, H_{b_1}, \ldots, H_{a_g}, H_{b_g}$. A particular role is played by the homotopically trivial loop $C = [a_1, b_1] \ldots [a_g, b_g]$.

- The distribution of the random variable H_C is such that for every continuous test function $\tilde{f} : G \to \mathbb{C}$,

$$\int_{\mathscr{C}_G} \tilde{f}(H_C) \, \mathrm{d}\mu_{\mathsf{YM}}^{\mathbb{G}, T} = Z_M(T)^{-1} \int_{G^{2g}} (\tilde{f} \, p_{T|M|})([a_1, b_1] \ldots [a_g, b_g]) \, \mathrm{d}a_1 \, \mathrm{d}b_1 \ldots \, \mathrm{d}a_g \, \mathrm{d}b_g$$

  This does not seem to be a particularly well-known distribution. It needs not have a density with respect to the Haar measure: for instance if $G = \mathrm{U}(N)$, it is supported by the Haar-negligible subgroup $\mathrm{SU}(N)$. However, it is, by definition, absolutely continuous with respect to the distribution of the product of g independent commutators of independent uniformly distributed random variables, and this distribution, for example if $G = \mathrm{SU}(N)$ and provided $g \geqslant 2$, is absolutely continuous with respect to the Haar measure. It is also possible to write a Fourier series for this distribution, but it involves Littlewood–Richardson coefficients, or more generally an understanding of the tensor product of irreducible representations of G.

- Conditional on H_C, the families $(H_{\lambda_1}, \ldots, H_{\lambda_r})$ and $(H_{a_1}, H_{b_1}, \ldots, H_{a_g}, H_{b_g})$ are independent. It is also true that the random variables

$$(H_{\lambda_1}, \ldots, H_{\lambda_r}) \bmod G \quad \text{and} \quad (H_{a_1}, H_{b_1}, \ldots, H_{a_g}, H_{b_g}) \bmod G$$

  with values in G^r/G and G^{2g}/G, where G acts by conjugation, are independent conditional on $H_C \bmod G$, that is, conditional on the conjugacy class of H_C.

On a surface of genus g, the probabilistic backbone of the holonomy process can thus be described as consisting of a segment of a Brownian motion on G of length $T|M|$ and $2g$ independent Haar distributed random variables on G, jointly conditioned on the final point of the Brownian motion being equal to the products of the g commutators of the uniform random variables taken in pairs.

The case where M is a sphere is special, in the sense that it involves no uniform random variables, but only a Brownian bridge on G going from 1_G to 1_G in a time equal to T times the total area of the sphere.

1.5 Wilson Loop Expectations

A different approach to the description of the distribution of the Yang–Mills holonomy process consists in identifying a natural class of scalar, gauge-invariant, functionals of this process, the distribution of which is hoped to contain as much information as possible. The most natural class of such functionals is that of *Wilson loop functionals*, which are indeed the most important scalar observables of the theory. A Wilson loop functional is constructed by choosing a certain number of loops $\ell_1, \ldots, \ell_n$ on M, then the same number of conjugation-invariant functions $\chi_1, \ldots, \chi_n : G \to \mathbb{C}$ and by forming the product

$$\chi_1(H_{\ell_1}) \ldots \chi_n(H_{\ell_n}) \tag{13}$$

When G is a group of matrices, the simplest choice of conjugation-invariant function is the trace. The *Wilson loop expectations*, which play in this theory the role of n-point functions, are the numbers

$$\mathbb{E}[\mathrm{Tr}(H_{\ell_1}) \ldots \mathrm{Tr}(H_{\ell_n})] \tag{14}$$

the computation of which is a seemingly endless subject of reflection. We will discuss in the next section a few concrete examples of computation of such numbers. For the time being, let us say a word about the amount of information that they carry.

Suppose we know the collection of all the numbers (14), or more generally the expectation of all functionals of the form (13). Then we know the joint distribution of all random variables of the form $\chi(H_\ell)$ where ℓ is a loop and $\chi : G \to \mathbb{C}$ is an invariant function. Since G is compact, invariant functions separate conjugacy classes and we know, in fact, the joint distribution of the conjugacy classes of all variables H_ℓ. This is certainly an important piece of information. However, the form of the action of the group of gauge transformations on the collection of holonomies, as given by (11), indicates that this action preserves more than just the individual conjugacy classes of the holonomies. Indeed, if $\ell_1, \ldots, \ell_n$ are based at the same point, then it is the orbit of $(H_{\ell_1}, \ldots, H_{\ell_n})$ under the operation of *simultaneous conjugation*

$$(h_1, \ldots, h_n) \mapsto (gh_1 g^{-1}, \ldots, gh_n g^{-1})$$

that is gauge-invariant. To grasp the geometric meaning of this invariance, it is useful to take a concrete example for G, say $G = \mathrm{SU}(N)$ or even $G = \mathrm{SO}(3)$. In these groups, knowing the individual conjugacy classes of a collection of elements amounts to knowing their eigenvalues, that is, in the case of SO(3), the angles of the rotations. On the other hand, to know the orbit of these elements under simultaneous conjugation requires the additional knowledge of the relative positions of their eigenspaces, or for rotations, the relative positions of their axes.

The main question is then the following. Is it the case that the Wilson loop expectations describe not only the individual conjugacy classes of the G-valued random variables that constitute the Yang–Mills process, but also the simultaneous conjugacy class of all variables associated with the loops based at some point m of M? In more precise terms, is it true that the algebra of functions on $\mathscr{A}/\mathscr{G}$ generated by Wilson loop functionals separates points? If not, it cannot be said that the Wilson loop functionals constitute a complete set of gauge-invariant scalar observables.

The answer turns out to depend entirely on the group G, and it does not seem to be known in all cases, even for compact Lie groups.[20] The property that G must have for the answer to be positive is the following.[21]

Definition 1 *(Property W)* We say that a group G has the property W if for any $n \geqslant 2$ and any two collections $x_1, \ldots, x_n$ and $x_1', \ldots, x_n'$ of elements of G, the assumption that every word in $x_1, \ldots, x_n$ and their inverses is conjugated to the same word in $x_1', \ldots, x_n'$ and their inverses implies the existence of an element y of G such that $x_1' = y x_1 y^{-1}, \ldots, x_n' = y x_n y^{-1}$.

Since this long definition is maybe not very pleasant to read, let us word it differently. We are comparing two relations between n-tuples $(x_1, \ldots, x_n)$ and $(x_1', \ldots, x_n')$ of elements of G. The first is the relation of simultaneous conjugation

$$\exists y \in G, \ x_1' = y x_1 y^{-1}, \ldots, x_n' = y x_n y^{-1} \tag{SC}$$

The second could be called *lexical conjugation* and holds exactly when

$$\text{every word in } x_1, \ldots, x_n \text{ is conjugated to the same word in } x_1', \ldots, x_n' \tag{LC}$$

where a word in a certain set of letters can involve these letters and their inverses. We also considered a third property of individual conjugation

$$\exists y_1, \ldots, y_n \in G, \ x_1' = y_1 x_1 y_1^{-1}, \ldots, x_n' = y_n x_n y_n^{-1} \tag{IC}$$

In any group, one has the chain of implications

$$(SC) \Rightarrow (LC) \Rightarrow (IC)$$

Unless the group G has very special properties (for instance that of being abelian), the second implication is not an equivalence, and the property (IC) is much weaker than the property (LC). For the group G to have the property W means that the properties (SC) and (LC) are equivalent. The proof of the following result can be found in [22], see also [10, 39].

---

[20]It would be more prudent to say that it is not known to the author.

[21]The name of Property W is by no means standard.

Theorem 1.9 *Any Cartesian product of special orthogonal, orthogonal, special unitary, unitary and symplectic groups has the property W.*

It is known that some non-compact groups fail to have the property W. However, it seems not be known whether this equivalence holds, for instance, for spin groups.

2 Computation of Wilson Loop Expectations

In this section, we will give a few concrete examples of computations with the Yang–Mills holonomy process, with an eye to its so-called *large N limit*, that is, its behaviour when the group G is taken to be U(N) with an appropriately scaled invariant product on its Lie algebra, and N tends to infinity.[22]

The basis of virtually any computation in 2-dimensional Yang–Mills theory is the Driver–Sengupta formula (DS). This formula can be combined with an expression of the heat kernel on G, for example its Fourier expansion, and lead to very concrete calculations. It is also possible to use a more dynamical, either analytic or probabilistic approach to the heat kernel, by seeing it as the solution of the heat equation or, almost equivalently, as the density of the distribution of the Brownian motion on G. We will illustrate these possibilities on a few examples in the simplest case where M is the plane, and then turn to the much more complicated case where M is the 2-dimensional sphere. For the sake of simplicity, we will assume in this section that the coupling constant T that appears in (DS) is equal to 1.

2.1 The Brownian Motion on the Unitary Group

In order to be as concrete as possible, and because we are interested in the large N limit, we will in this section choose $G = $ U(N), the unitary group of rank N. As indicated earlier (see Footnote 4), we endow the Lie algebra of U(N), which is the space $\mathfrak{u}(N)$ of $N \times N$ skew-Hermitian matrices, with the scalar product $\langle X, Y \rangle = N\mathrm{Tr}(X^*Y)$. In the Euclidian space $(\mathfrak{u}(N), \langle \cdot, \cdot \rangle)$, we consider a linear Brownian motion $(K_t)_{t \geqslant 0}$, use it to form the stochastic differential equation

$$dU_t = U_t \, \mathrm{d}K_t - \frac{1}{2} U_t \, \mathrm{d}t \, , \quad U_0 = I_N \tag{15}$$

and call the unique solution to this equation the Brownian motion on U(N).

---

[22]The notion of large N limit also applies to the cases where $G = $ SO(N) and $G = $ Sp(N), the real and quaternionic analogues of U(N) or SU(N). As far as we understand today, there is no essential difference between the three cases. More precisely, the computations for finite N are similar in the three cases, if generally a bit more complicated in the orthogonal case and even more so in the symplectic case, and the large N limits are identical.

Using the notation Tr for the usual trace of a $N \times N$ matrix and $\mathrm{tr} = \frac{1}{N}\mathrm{Tr}$ for its normalised trace, the usual rules of stochastic calculus take, in this matricial context, the following nice form: for all $N \times N$ matrix A, measurable with respect to $\sigma(K_s : s \leqslant t)$, we have

$$\mathrm{d}K_t A \, \mathrm{d}K_t = -\mathrm{tr}(A) \, \mathrm{d}t \quad \text{and} \quad \mathrm{d}K_t \, \mathrm{tr}(A \, \mathrm{d}K_t) = -\frac{1}{N^2} A \, \mathrm{d}t \qquad (16)$$

This relation can be used to check that $\mathrm{d}(U_t U_t^*) = 0$, so that the trajectories of the process B stay almost surely, as expected, in U(N).

The density of the distribution of U_t with respect to the normalised Haar measure on U(N) is the function p_t appearing in the Driver–Sengupta formula, and that we described in Section 1.4.2.

It will be useful to know the Fourier series of this function $p_t : \mathrm{U}(N) \to \mathbb{R}$. To describe it, let us introduce the set $\widehat{\mathrm{U}}(N)$ of equivalence classes of irreducible representations (or *irreps*) of U(N). For every $\alpha \in \widehat{\mathrm{U}}(N)$, let us denote by d_α the degree of α, that is, the dimension of the space on which U(N) acts through α. Let us also denote by $\chi_\alpha : \mathrm{U}(N) \to \mathbb{C}$ the character of α, and by $c_2(\alpha)$ the quadratic Casimir number of α, that is, the non-negative real number such that

$$\Delta \chi_\alpha = -c_2(\alpha)\chi_\alpha$$

The Fourier series of the heat kernel is then

$$p_t = \sum_{\alpha \in \widehat{\mathrm{U}}(N)} e^{-\frac{c_2(\alpha)t}{2}} d_\alpha \chi_\alpha \qquad (17)$$

and there is nothing specific to U(N) in this formula.

It is however possible, in the case of U(N), to write explicitly each of its ingredients. Indeed, the set of irreps of U(N) is conveniently labelled by non-increasing sequences of N relative integers $\lambda = (\lambda_1 \geqslant \ldots \geqslant \lambda_N)$, called dominant weights. The dimension and quadratic Casimir number of the irrep with highest weight λ are given by the formulas

$$d_\lambda = \prod_{1 \leqslant i < j \leqslant N} \frac{\lambda_i - \lambda_j + j - i}{j - i} \quad \text{and} \quad N c_2(\lambda) = \sum_{1 \leqslant i \leqslant N} \lambda_i^2 + \sum_{1 \leqslant i < j \leqslant N} (\lambda_i - \lambda_j)$$

$$(18)$$

The character of this representation is given, up to a power of the determinant, by a Schur function, but we will not need its explicit formula.

We are now equipped to make some computations with the Yang–Mills holonomy process.

2.2 The Simple Loop on the Plane

2.2.1 Using Harmonic Analysis

Let us consider, on the plane, a loop ℓ that is a simple loop going once around a domain of area t (see, if needed, Figure 2). The partition function of the Yang–Mills model on the plane is equal to 1 and the Driver–Sengupta formula (DS) tells us that for every continuous test function $f : U(N) \to \mathbb{C}$, we have

$$\mathbb{E}[f(H_\ell)] = \int_{U(N)} f(x) p_t(x) \, dx$$

In other words, H_ℓ has the same distribution as U_t, the value at time t of the Brownian motion on $U(N)$ defined in the previous section.

Using the Fourier expansion (17) and the classical orthogonality relations between characters, we find, for every irrep α of $U(N)$ acting on the vector space V_α, the equality

$$\mathbb{E}[\alpha(H_\ell)] = e^{-\frac{c_2(\alpha)t}{2}} \, \mathrm{id}_{V_\alpha}$$

which holds in $\mathrm{End}(V_\alpha)$. In particular, since the usual trace is, on $U(N)$, the character of the natural representation, which has highest weight $(1, 0, \ldots, 0)$, dimension N and quadratic Casimir 1, we find

$$\mathbb{E}[H_\ell] = e^{-\frac{t}{2}} I_N \quad \text{and} \quad \mathbb{E}[\mathrm{tr}(H_\ell)] = e^{-\frac{t}{2}} \tag{19}$$
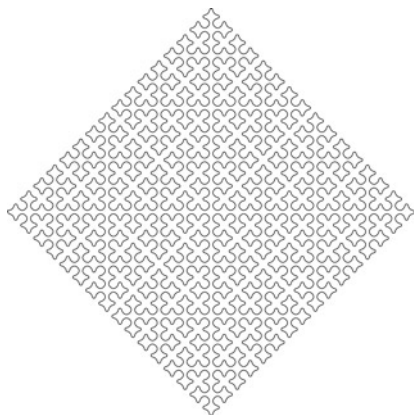
Fig. 2 A simple loop on the plane

Suppose now that we want to compute the expectation of $\text{tr}(H_\ell^2)$, which is also the expectation of $\text{tr}(H_{\ell^2})$, where ℓ^2 is the loop ℓ gone along twice. From the Driver–Sengupta formula and the Fourier expansion of the heat kernel, we get the expression

$$\mathbb{E}[\text{tr}(H_\ell^2)] = \sum_{\lambda \in \widehat{U}(N)} e^{-\frac{c_2(\lambda)t}{2}} d_\lambda \int_{U(N)} \text{tr}(x^2) \chi_\lambda(x)\, \mathrm{d}x$$

In order to go further, we need to know that, at least when $N \geqslant 2$,

$$\text{tr}(x^2) = \chi_{(2,0,\dots,0)}(x) - \chi_{(1,1,0\dots,0)}(x)$$

Using again the orthogonality of characters, we find, after some reordering of the terms,

$$\mathbb{E}[\text{tr}(H_\ell^2)] = e^{-t}\left(\cosh\frac{t}{N} - N \sinh\frac{t}{N} \right) \tag{20}$$

It is possible to go further down this road, by systematically writing the function $x \mapsto \text{tr}(x^n)$ as a linear combination of characters. This is what Philippe Biane did to determine the large N limit of the non-commutative distribution of the Brownian motion on the unitary group. The simplest non-trivial case is the large N limit of (20):

$$\lim_{N \to \infty} \mathbb{E}[\text{tr}(H_\ell^2)] = e^{-t}(1 - t) \tag{21}$$

The general formula is nice enough, at least in the limit when N tends to infinity, to be quoted explicitly. It was discovered independently by Philippe Biane and Eric Rains, who formulated it in terms of the Brownian motion on $U(N)$ rather than the Yang–Mills holonomy process.

Theorem 2.1 (Biane [2], Rains [37]) *With the current notation, and for every integer $n \geqslant 1$,*

$$\lim_{N \to \infty} \mathbb{E}[\text{tr}(H_\ell^n)] = e^{-\frac{nt}{2}} \sum_{k=0}^{n-1} \frac{(-t)^k}{k!} n^{k-1} \binom{n}{k+1} \tag{22}$$

It must be said that this result already appeared, without proof, in Isadore Singer's seminal paper on the large N limit of the Yang–Mills holonomy field [42].[23]

---

[23] Singer and Rains recognise, in the right-hand side of (22), modified Laguerre polynomials of the first kind. As far as I know, a structural explanation for the appearance of these polynomials in this context has yet to be given.

One of Biane's aims in [2] was to prove the following theorem concerning the limit as N tends to infinity of the Brownian motion on $\mathrm{U}(N)$ as a stochastic process. This convergence result is stated in the language of free probability, a theory presented in detail in the book of Alexandru Nica and Roland Speicher [34].

Theorem 2.2 (Biane [2]) *As N tends to infinity, the Brownian motion on $\mathrm{U}(N)$ converges in non-commutative distribution, as a process, towards a unitary non-commutative process $(u_t)_{t \geqslant 0}$ with free stationary multiplicative increments such that for all integer $n \geqslant 0$ and all real $t \geqslant 0$, the expectation of u_t^n and that of $(u_t^*)^n$ are given by the right-hand side of (22).*

2.2.2 Using Stochastic Calculus

Let us illustrate, on the same example of a simple loop on the plane, the dynamical approach to the same computations, based on the use of Itō's formula. The general principle of these computations is to see the quantities such as the left-hand sides of (19) and (20) as functions of t, and to write a differential equation that they satisfy. Recall that t, in our current notation, is the area of the disk enclosed by the simple loop ℓ. A variation of t can thus be described, in geometrical terms, as a variation of the area of the unique face enclosed by ℓ.

As a first example, let us use (15) and Itō's formula to find

$$\frac{d}{dt}\mathbb{E}[\mathrm{tr}(H_\ell)] = \frac{d}{dt}\mathbb{E}[\mathrm{tr}(U_t)] = -\frac{1}{2}\mathbb{E}[\mathrm{tr}(U_t)]$$

which, together with the information $\mathbb{E}[\mathrm{tr}(U_0)] = 1$, yield immediately (19).

Let us apply the same strategy to the computation of $\mathbb{E}[\mathrm{tr}(H_\ell^2)] = \mathbb{E}[\mathrm{tr}(U_t^2)]$. The computation is more interesting and involves the first of the two rules (16). We find

$$\frac{d}{dt}\mathbb{E}[\mathrm{tr}(U_t^2)] = -\mathbb{E}[\mathrm{tr}(U_t^2)] - \mathbb{E}[\mathrm{tr}(U_t)^2] \tag{23}$$

and see a function of t pop up that we were initially not interested in, namely $\mathbb{E}[\mathrm{tr}(U_t)^2]$. The only way out left to us is retreat forwards and we compute the derivative with respect to t of this new function, using now the second rule of (16):

$$\frac{d}{dt}\mathbb{E}[\mathrm{tr}(U_t)^2] = -\frac{1}{N^2}\mathbb{E}[\mathrm{tr}(U_t^2)] - \mathbb{E}[\mathrm{tr}(U_t)^2] \tag{24}$$

All's well that ends well: (23) and (24) form a closed system of ordinary differential equations that is easily solved and from which we recover, in particular, (20). As a bonus, we get

$$\mathbb{E}[\mathrm{tr}(H_\ell)^2] = e^{-t}\left(\cosh\frac{t}{N} - \frac{1}{N}\sinh\frac{t}{N}\right) \tag{25}$$

The only change with respect to (20) is the change from N to $\frac{1}{N}$ in front of the hyperbolic sine, with the effect that

$$\lim_{N\to\infty} \mathbb{E}[\operatorname{tr}(H_\ell)^2] = e^{-t} = \lim_{N\to\infty} \mathbb{E}[\operatorname{tr}(H_\ell)]^2 \tag{26}$$

This is an instance of a general factorisation property which was observed, among others, by Feng Xu [47], and which is a consequence of the concentration, in the limit where N tends to infinity, of the spectra of the random matrices that we are considering.

2.3 Yin . . .

Let us consider a slightly more complicated loop depicted on Figure 3. This loop goes once around a domain of area $s + t$ and then once around a smaller domain of area t contained in the first one.

Let us apply the Driver–Sengupta formula in this case. We denote a generic element of the configuration space $U(N)^2$ by (x_a, x_b), in relation with our labelling by a and b of the two edges of the graph formed by ℓ. Thus, for every continuous test function $f : U(N) \to \mathbb{C}$, we have

$$\mathbb{E}[f(H_\ell)] = \int_{U(N)^2} f(x_b x_a) p_s(x_b^{-1} x_a) p_t(x_b) \, dx_a \, dx_b$$

Note that, according to (10), the discrete holonomy map is order-reversing, so that the loop $\ell = ab$ gives rise to the map $h_\ell(x_a, x_b) = x_b x_a$.

The change of variables $(y, z) = (x_b^{-1} x_a, x_b)$ preserves the Haar measure on $U(N)^2$ and we have

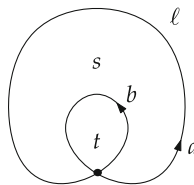$$\mathbb{E}[f(H_\ell)] = \int_{U(N)^2} f(z^2 y) p_s(y) p_t(z) \, dy \, dz \tag{27}$$

Fig. 3 The loop ℓ goes first once along the larger circle (the edge a) and then once along the smaller circle (the edge b). The loop ab is equivalent to the concatenation of ab^{-1}, b and b. The loops ab^{-1} and b are essentially simple loops surrounding disjoint domains

This corresponds to the fact, explained in the caption of Figure 3, that the loop ℓ can be written as $\ell_1 \ell_2 \ell_2$, where ℓ_1 goes around the moon-shaped domain sitting between the two disks, and ℓ_2 goes around the small circle of area t. These loops enclose disjoint domains, and although ℓ_1 is not strictly speaking self-intersection free, they are essentially simple, in the sense that they can be approximated by simple loops.

From this graphical decomposition of ℓ, or from (27), we infer that H_ℓ has the distribution of $V_t^2 U_s$, where U and V are independent Brownian motions on U(N).[24] Using the independence, the fact that the expectation of U_s is $e^{-\frac{s}{2}} I_N$ (see (19)), and (20), we find

$$\mathbb{E}[\mathrm{tr}(H_\ell)] = e^{-\frac{s}{2}-t} \left(\cosh \frac{t}{N} - N \sinh \frac{t}{N} \right) \tag{28}$$

and, letting N tend to infinity,

$$\lim_{N\to\infty} \mathbb{E}[\mathrm{tr}(H_\ell)] = e^{-\frac{s}{2}-t}(1-t) \tag{29}$$

We succeeded in computing the expectation of $\mathrm{tr}(H_\ell)$, but we did so by taking advantage of the favourable circumstances, namely the fact that the word $V_t^2 U_s$ is a very simple one, with two independent Brownian motions appearing one after the other (and not, for example, as $U_s V_t U_s V_t$), and the fact that the expectation of U_s is a very simple matrix.

A more systematic approach is possible, by looking at $\mathbb{E}[\mathrm{tr}(V_t^2 U_s)]$ as a function of s and t and by using Itō's formula to compute its partial derivatives. One finds

$$\partial_s \mathbb{E}[\mathrm{tr}(V_t^2 U_s)] = -\frac{1}{2}\mathbb{E}[\mathrm{tr}(V_t^2 U_s)]$$

$$\partial_t \mathbb{E}[\mathrm{tr}(V_t^2 U_s)] = -\mathbb{E}[\mathrm{tr}(V_t^2 U_s)] - \mathbb{E}[\mathrm{tr}(V_t)\mathrm{tr}(V_t U_s)]$$

Once again, a function appears that we were not considering at first. Let us apply the same treatment to this new function:

$$\partial_s \mathbb{E}[\mathrm{tr}(V_t)\mathrm{tr}(V_t U_s)] = -\frac{1}{2}\mathbb{E}[\mathrm{tr}(V_t)\mathrm{tr}(V_t U_s)]$$

$$\partial_t \mathbb{E}[\mathrm{tr}(V_t)\mathrm{tr}(V_t U_s)] = -\mathbb{E}[\mathrm{tr}(V_t)\mathrm{tr}(V_t U_s)] - \frac{1}{N^2}\mathbb{E}[\mathrm{tr}(V_t^2 U_s)]$$

It is possible to solve this system and to recover (28).

---

[24]Thanks to the independence of the multiplicative increments of the Brownian motion, this distribution is of course also that of $V_t^2(V_t^{-1}V_{t+s}) = V_t V_{s+t}$. Reasoning in this way amounts to undo the change of variables that we did to obtain (27).

An interesting observation is the fact that the linear combination $2\partial_s - \partial_t$ of partial derivatives is particularly simple:

$$(2\partial_s - \partial_t)\mathbb{E}[\mathrm{tr}(V_t^2 U_s)] = \mathbb{E}[\mathrm{tr}(V_t)\mathrm{tr}(V_t U_s)] \tag{30}$$

$$\text{and } (2\partial_s - \partial_t)\mathbb{E}[\mathrm{tr}(V_t)\mathrm{tr}(V_t U_s)] = \frac{1}{N^2}\mathbb{E}[\mathrm{tr}(V_t^2 U_s)] \tag{31}$$

These are instances of the Makeenko–Migdal equations that we will discuss in greater detail in the next section. Before that, let us study another example.

2.4   ... And Yang

Let us now consider the eight-shaped loop drawn on Figure 4. The Driver–Sengupta formula yields, with the by now usual notation, and taking the inversion of the order into account,

$$\mathbb{E}[f(H_\ell)] = \int_{\mathrm{U}(N)^6} f(x_f x_e x_d x_c x_b x_a) p_s(x_a x_c x_e) p_t(x_f x_b x_d) p_u(x_c^{-1} x_f) p_v(x_a^{-1} x_d) \, \mathrm{d}x$$

The appropriate change of variables is dictated by the geometry of the loop, more precisely by a decomposition in product of lassos, one of which is given in the caption of Figure 4. Accordingly, let us set

$$(y, z, g, h, e, f) = (x_c x_e x_a, x_f x_b x_d, x_c x_f^{-1}, x_d^{-1} x_a, x_e, x_f)$$
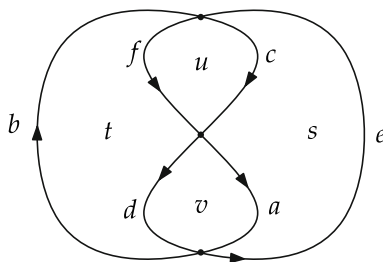
Fig. 4   An eight-shaped loop on the plane. The letters s, t, u, v in the faces indicate the areas of the faces. The other letters label the edges of the graph. The loop can be decomposed, as we did for the heart-shaped loop, as a product of lassos enclosing pairwise disjoint domains: $abcdef = (ad^{-1})(dbf)(f^{-1}c)(da^{-1})(aec)(c^{-1}f)$. Here, by a lasso, we mean a loop of the form clc^{-1}, where c is a path starting from the starting point of our loop and l is a simple loop. In this particular case, the path c is always the constant path

This change of variables preserves the Haar measure on $U(N)^6$.[25] Thus, we find

$$\mathbb{E}[f(H_\ell)] = \int_{U(N)^4} f(g^{-1}yh^{-1}gzh) p_s(y) p_t(z) p_u(g) p_v(h) \, dg \, dh \, dy \, dz$$

after integrating with respect to e and f which do not appear in the integrand. Thus, considering four independent Brownian motions G, H, Z, Y on $U(N)$, we find the equality in distribution

$$H_\ell \overset{\text{dist.}}{=} G_u^{-1} Y_s H_v^{-1} G_u Z_t H_v \tag{32}$$

The quantity $\mathbb{E}[\text{tr}(H_\ell)]$ appears now as a function of the four real parameters s, t, u, v and we can use stochastic calculus to differentiate it with respect to each of them. In fact, using the first assertion of (19), which in the language of Brownian motion reads $\mathbb{E}[Y_s] = e^{-\frac{s}{2}} I_N$ and $\mathbb{E}[Z_t] = e^{-\frac{t}{2}} I_N$, we can simplify the problem to

$$\mathbb{E}[\text{tr}(H_\ell)] = e^{-\frac{s+t}{2}} \mathbb{E}[\text{tr}(G_u^{-1} H_v^{-1} G_u H_v)]$$

The expectation in the right-hand side of this equality is a symmetric function of u and v. Using stochastic calculus, we find

$$\partial_u \mathbb{E}[\text{tr}(G_u^{-1} H_v^{-1} G_u H_v)] = -\mathbb{E}[\text{tr}(G_u^{-1} H_v^{-1} G_u H_v)] + \mathbb{E}[\text{tr}(H_v^{-1}) \text{tr}(H_v)] \tag{33}$$

The new function $\mathbb{E}[\text{tr}(H_v^{-1}) \text{tr}(H_v)]$ of v can in turn be computed using Itō's formula, since it is equal to 1 when $v = 0$ and satisfies the differential equation

$$\partial_v \mathbb{E}[\text{tr}(H_v^{-1}) \text{tr}(H_v)] = -\mathbb{E}[\text{tr}(H_v^{-1}) \text{tr}(H_v)] + \frac{1}{N^2}$$

---

[25]This is because the normalised Haar measure on $U(N)^n$, or on G^n for any compact topological group G, is pushed forward onto itself by each of the elementary maps

- $(x_1, x_2, \ldots, x_n) \mapsto (x_1^{-1}, x_2, \ldots, x_n)$
- $(x_1, x_2, \ldots, x_n) \mapsto (x_1 x_2, x_2, \ldots, x_n)$
- $(x_1, \ldots, x_n) \mapsto (x_{\sigma(1)}, \ldots, x_{\sigma(n)})$, where σ is any permutation of $\{1, \ldots, n\}$

and it is not difficult to check that our change of variables can be obtained as a composition of these maps.

Interestingly, these elementary operations are exactly the Nielsen transformations, which generate the group of automorphisms of the free group of rank n (see [30]). Thus, the random homomorphism from the free group $\mathbb{F}_n$ to a compact topological group G constructed by picking a basis of $\mathbb{F}_n$ and sending this basis to a uniformly chosen element of G^n does not depend, in distribution, on the basis of $\mathbb{F}_n$ used to construct it. In particular, the distribution of the image of every element of the free group is intrinsically defined, and one may for instance wonder, for specific or for general G, which elements of $\mathbb{F}_n$ are sent to a uniformly distributed element of G. I am grateful to the referee for pointing out to me that this problem was solved for finite groups in [35].

which is solved in

$$\mathbb{E}[\mathrm{tr}(H_v^{-1})\mathrm{tr}(H_v)] = \frac{1}{N^2}(1 - e^{-v}) + e^{-v} \tag{34}$$

Replacing in (33) and solving, we find finally

$$\mathbb{E}[\mathrm{tr}(H_\ell)] = e^{-\frac{s+t}{2}} \left(e^{-u} + e^{-v} - e^{-(u+v)} + \frac{1}{N^2}(1 - e^{-u})(1 - e^{-v}) \right) \tag{35}$$

and, letting N tend to infinity,

$$\lim_{N \to \infty} \mathbb{E}[\mathrm{tr}(H_\ell)] = e^{-\frac{s+t}{2}} \left(e^{-u} + e^{-v} - e^{-(u+v)} \right) \tag{36}$$

We did these computations without taking great care of a possible geometric meaning of the successive steps. Anticipating our discussion of the Makeenko–Migdal equations, it is interesting to check that

$$(\partial_u + \partial_v - \partial_s - \partial_t)\mathbb{E}[\mathrm{tr}(H_\ell)] = e^{-\frac{s+t}{2}} \left(e^{-(u+v)} + \frac{1}{N^2}(1 - e^{-(u+v)}) \right) = \mathbb{E}[\mathrm{tr}(H_{\ell'})\mathrm{tr}(H_{\ell''})] \tag{37}$$

where ℓ' and ℓ'' are the loops drawn on Figure 5.

Perhaps even more interesting than the fact that (37) holds, which after all is a consequence of Theorem 3.1, is the observation that (37) does not seem to be easily guessed from (32) and Itō's formula. More precisely, Itō's formula allows us to give an expression of the left-hand side of (37) and it is not obvious that this expression coincides with the right-hand side of (37). We take this as a sign that the Makeenko–Migdal equations give an information that is practically non-trivial.
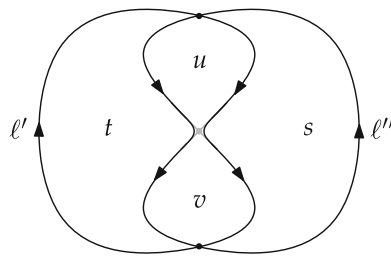

Fig. 5 The loops ℓ' and ℓ'' are obtained from ℓ by an operation that will feature prominently in Section 3

2.5 The Case of the Sphere: A Not So Simple Loop

Computations involving the Yang–Mills holonomy process on the sphere, although in principle based on the same formulas as in the case of the plane, are in general much more complicated. This can be explained by the fact that, as we indicated in Section 1.4.5, the stochastic core of the Yang–Mills holonomy process on a sphere is a Brownian *bridge* on $U(N)$, or on the compact Lie group G, instead of a Brownian motion.

In this section, we are going to illustrate some of the difficulties that one meets when working on a sphere. The first is that the partition function is not equal to 1 anymore. Instead, according to (1.5), it is given, on a sphere of total area T, by

$$Z_{S^2}(T) = p_T(I_N) = \|p_{\frac{T}{2}}\|^2_{L^2(U(N))} = \sum_{\alpha \in \widehat{U}(N)} e^{-\frac{T}{2}c_2(\alpha)} d_\alpha^2$$

This is also an expression in which nothing is specific to $U(N)$: it is valid for any compact Lie group.[26]

The most basic question about the Yang–Mills holonomy process on the sphere is the analogue to the question that we treated in Section 2.2, namely to compute the expectation of the normalised trace of the holonomy along a simple loop ℓ enclosing a domain of area t. The Driver–Sengupta formula yields the following expression for this expectation:

$$\mathbb{E}[\text{tr}(H_\ell)] = \frac{1}{Z_{S^2}(T)} \int_{U(N)} \text{tr}(x) p_t(x) p_{T-t}(x^{-1}) \, dx \tag{38}$$

Using the Fourier expansion of the heat kernel, one finds

$$\mathbb{E}[\text{tr}(H_\ell)] = \frac{1}{Z_{S^2}(T)} \sum_{\lambda,\mu \in \widehat{U}(N)} e^{-c_2(\lambda)\frac{t}{2} - c_2(\mu)\frac{T-t}{2}} d_\lambda d_\mu \int_{U(N)} \text{tr}(x)\chi_\lambda(x)\chi_\mu(x^{-1}) \, dx$$

The integral can be computed thanks to Pieri's rule: it is equal to 0 unless μ is obtained from λ by adding 1 to exactly one component, in which case it is equal to 1. We write $\lambda \nearrow \mu$ when this happens. Thus,

$$\mathbb{E}[\text{tr}(H_\ell)] = \frac{1}{Z_{S^2}(T)} \sum_{\lambda \in \widehat{U}(N)} e^{-c_2(\lambda)\frac{T}{2}} d_\lambda^2 \underbrace{\left[\sum_{\substack{\mu \in \widehat{U}(N) \\ \lambda \nearrow \mu}} e^{-(c_2(\mu) - c_2(\lambda))\frac{T-t}{2}} \frac{d_\mu}{d_\lambda} \right]}_{f_1(\lambda)} \tag{39}$$

---

[26]Note that T, which used to denote the coupling constant in (1.5), now denotes the total area of our surface. This is not a problem because the only meaningful quantity is the product of the coupling constant by the total area of the surface.

It seems difficult to give an expression of $\mathbb{E}[\text{tr}(H_\ell)]$ much simpler than (38) or (39) which, as is hardly necessary to emphasise, is much more complicated than the one that we obtained in the case of the plane.[27]

It is, however, possible to analyse the limit of this quantity as N tends to infinity. A first step in this direction is based on the realisation that Pieri's rule is simple, and the quantity between square brackets, which we denote by $f_1(\lambda)$ is a finite sum and can be written explicitly using (18):

$$f_1(\lambda) = e^{-\frac{T-t}{2}} \sum_{i=1}^{N} \mathbb{1}_{\{i=1 \text{ or } \lambda_{i-1} > \lambda_i\}} e^{-(T-t)\left(\lambda_i + \frac{N-2i+1}{2}\right)} \prod_{\substack{1 \leqslant j \leqslant N \\ j \neq i}} \left(1 + \frac{1}{\lambda_i - \lambda_j + j - i}\right)$$

This suggests to associate with the highest weight λ the decreasing sequence $l = (l_1 > \ldots > l_N)$ of half-integers defined by

$$l_i = \lambda_i + \frac{N - 2i + 1}{2}$$

so that

$$f_1(\lambda) = e^{-\frac{T-t}{2}} \sum_{i=1}^{N} \mathbb{1}_{\{i=1 \text{ or } \lambda_{i-1} > \lambda_i\}} e^{-\frac{T-t}{N} l_i} \prod_{\substack{1 \leqslant j \leqslant N \\ j \neq i}} \left(1 + \frac{1}{l_i - l_j}\right)$$

Let us now introduce the probability measure $\pi_{N,T}$ on $\widehat{U}(N)$ such that for every highest weight λ, one has

$$\pi_{N,T}(\{\lambda\}) \propto e^{-c_2(\lambda)\frac{T}{2}} d_\lambda^2$$

Then (39) can be written more compactly as

$$\mathbb{E}[\text{tr}(H_\ell)] = \int_{\widehat{U}(N)} f_1(\lambda) \, d\pi_{N,T}(\lambda) \tag{40}$$

Moreover, there exists for each integer $n \geqslant 2$ a function f_n on $\widehat{U}(N)$, not very different from f_1, and the integral of which against $\pi_{N,T}$ yields $\mathbb{E}[\text{tr}(H_\ell^n)]$.

We would like to express that, as N tends to infinity, the measure $\pi_{N,T}$ concentrates on a few highest weights, characterised by a certain limiting shape. One unpleasant feature of (40) in this respect is that the set on which the integral is taken, namely $\widehat{U}(N)$, depends on N. It is thus uneasy to formulate a concentration

---

[27]Let us drive the point home: (39), once made fully explicit using (18), is the exact analogue of the $e^{-\frac{t}{2}}$ that we see in the second assertion of (19).
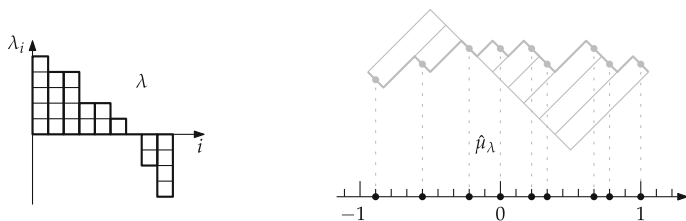
Fig. 6 With $N = 9$, the highest weight $\lambda = (5, 4, 4, 2, 2, 1, 0, -2, -4)$ drawn in the style of a Young diagram, and its empirical measure. Each dot represents $\frac{1}{9}$ of mass and any two dots are distant by a multiple of $\frac{1}{9}$

result. One classical and efficient way around this problem is to associate with each highest weight λ its *empirical measure* (Figure 6)

$$\hat{\mu}_\lambda = \frac{1}{N} \sum_{i=1}^{N} \delta_{\frac{l_i}{N}} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\frac{1}{N}(\lambda_i + \frac{N-2i+1}{2})}$$

Pushing the probability measure $\pi_{N,T}$ forward by the map $\lambda \mapsto \hat{\mu}_\lambda$ yields a probability measure, which we denote by $\Pi_{N,T}$, on the set of probability measures on the real line. It is possible to predict the behaviour of this probability as N tends to infinity by writing $c_2(\lambda)$ and d_λ in terms of the empirical measure of λ. Up to some inessential terms (see [25, Eq. (24)] for complete expressions), one finds

$$c_2(\lambda) \simeq N^2 \int_{\mathbb{R}} x^2 \, d\hat{\mu}_\lambda(x) \quad \text{and}$$

$$d_\lambda^2 \simeq \exp\left[-N^2 \int_{\{(x,y)\in\mathbb{R}^2, x\neq y\}} -\log|x-y| \, d\hat{\mu}_\lambda(x) \, d\hat{\mu}_\lambda(y) \right]$$

Introducing, for every probability measure μ, the quantity

$$\mathcal{J}_T(\mu) = \int_{\{(x,y)\in\mathbb{R}^2, x\neq y\}} -\log|x-y| \, d\mu(x) \, d\mu(y) + \frac{T}{2} \int_{\mathbb{R}} x^2 \, d\mu(x)$$

we see that the probability measure $\Pi_{N,T}$ assigns to any probability measure μ that is the empirical measure of a highest weight a mass proportional to

$$\Pi_{N,T}(\{\mu\}) \propto \exp(-N^2 \mathcal{J}_T(\mu))$$

In the large N limit, it seems plausible that $\Pi_{N,T}$ will concentrate on the minimisers, or even better, on the unique minimiser of the functional $\mathcal{J}_T$. This turns out to be true, with a little twist that we will explain and contributes to making the story much

more interesting than it already is. Let us summarise the main results on which one
can ground a rigorous analysis of the situation.

- Minimising the functional $\mathcal{J}_T$ on the space of all probability measures on $\mathbb{R}$ is
  one of the simplest examples of a rich and well-developed theory which is, for
  example, exposed in the book of Edward Saff and Vilmos Totik [38]. This is also
  a very common problem in random matrix theory. Indeed, the unique minimiser
  of $\mathcal{J}_T$ is Wigner's semi-circular distribution with variance $\frac{1}{T}$:

$$d\sigma_{1/T}(x) = \frac{T}{2\pi} \sqrt{\frac{4}{T} - x^2} \, \mathbb{1}_{\left[-\frac{2}{\sqrt{T}}, \frac{2}{\sqrt{T}}\right]}(t) \, dt \qquad (41)$$

- The fact that the measure $\Pi_{N,T}$ concentrates, as N tends to infinity, to the
  minimiser of $\mathcal{J}_T$ is a special case of a principle of large deviations proved by
  Alice Guionnet and Mylène Maïda in [16]. However, the minimiser of $\mathcal{J}_T$ that
  one must consider is not the absolute minimiser on the set of all probability
  measures on $\mathbb{R}$. Indeed, for all $N \geqslant 1$, the measure $\Pi_{N,T}$ is supported by the set
  of empirical measures of highest weights of $U(N)$, which form a rather special
  set of probability measures. A distinctive feature of these measures is that they
  are atomic, with atoms of mass $\frac{1}{N}$ spaced by integer multiples of $\frac{1}{N}$. Weak limits,
  as N tends to infinity, of such measures can only be absolutely continuous with
  respect to the Lebesgue measure on $\mathbb{R}$, with a density not exceeding 1: a class
  of probability measures that we will denote by $\mathcal{L}(\mathbb{R})$. The result of Guionnet and
  Maïda asserts that the measure $\Pi_{N,T}$ concentrates exponentially fast, as N tends
  to infinity, around the unique minimiser μ_T^* of $\mathcal{J}_T$ on the closed set $\mathcal{L}(\mathbb{R})$.
- The problem of minimising $\mathcal{J}_T$ under the constraint of having a density not
  exceeding 1 is a problem that is, in principle, just as well understood as the
  unconstrained problem. The book [38] contains results ensuring the existence and
  uniqueness of the minimiser, and others allowing one to determine its support. In
  fact, the measure $\sigma_{1/T}$ given by (41), and which is the absolute minimiser of $\mathcal{J}_T$,
  is absolutely continuous with a maximal density of $\sqrt{T}/\pi$, so that it belongs to
  $\mathcal{L}(\mathbb{R})$ provided $T \leqslant \pi^2$. For $T > \pi^2$, the constraint becomes truly restrictive,
  and one must make do with a probability measure which is, in $\mathcal{L}(\mathbb{R})$, the best
  available substitute for $\sigma_{1/T}$. The actual determination of this minimiser μ_T^* is,
  depending on one's background, a more or less elementary exercise in Riemann–
  Hilbert theory, and involves manipulating elliptic functions. The density of μ_T^*
  for $T > \pi^2$ is represented on Figure 7. An exact expression of this density can
  be found in [25, Eq. (37)].

Having established the exponential concentration, as N tends to infinity, of the
measure $\Pi_{N,T}$ around μ_T^*, it is possible to come back to our initial problem of
computing $\mathbb{E}[\mathrm{tr}(H_\ell)]$. After noticing that $f_1(\lambda)$ can be expressed as a functional
$F_1(\hat{\mu}_\lambda)$ of the empirical measure of λ, it can be guessed that $\mathbb{E}[\mathrm{tr}(H_\ell)]$ is related
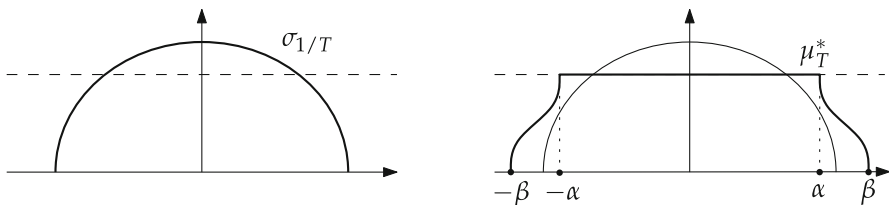to $F_1(\mu_T^*)$. Antoine Dahlqvist and James Norris were the first to rigorously and

Fig. 7 For $T > \pi^2$, the absolute minimiser of the functional $\mathcal{J}_T$ does not belong to the class of probabilities on $\mathbb{R}$ with a density not exceeding 1. The minimiser within this class is represented on the right. Its density is identically equal to 1 on an interval in the middle of its support, and given by elliptic functions outside this interval

successfully pursue this line of reasoning, and to obtain the following remarkably elegant result.

Theorem 2.3 (Dahlqvist–Norris [5]) *Let ρ_T denote the density of the minimiser μ_T^*. Then, for all integer $n \geqslant 0$, one has*

$$\lim_{N \to \infty} \mathbb{E}[\mathrm{tr}(H_\ell^n)] = \lim_{N \to \infty} \mathbb{E}[\mathrm{tr}(H_\ell^{-n})] = \frac{1}{n\pi} \int_{\mathbb{R}} \cosh\left(\frac{nx}{2}(T-2t)\right) \sin(n\pi\rho_T(x)) \, \mathrm{d}x \tag{42}$$

To conclude this long discussion of the simple loop on the sphere, let us mention another result for the statement of which we have all the concepts at hand. Our description of the behaviour of the measure $\Pi_{N,T}$ suggests that the partition function itself is dominated by the contribution of the highest weights that have an empirical measure close to μ_T^*. This is indeed true, and the fact that the shape of μ_T^* changes suddenly when T crosses the critical value π^2 gives rise to a phase transition, in this case of third order, first discovered by Douglas and Kazakov, and named after them. It was first proved rigorously, in a slightly different but equivalent language, by Karl Liechty and Dong Wang in [27], and by Mylène Maïda and the author in [25].

Theorem 2.4 (Douglas–Kazakov phase transition) *The free energy of the Yang–Mills model on a sphere of total area T is given by*

$$F(T) = \lim_{N \to \infty} \frac{1}{N^2} \log Z_{S^2}(T) = \frac{T}{24} + \frac{3}{2} - \mathcal{J}_T(\mu_T^*)$$

The function F is of class C^2 on $(0, \infty)$ and smooth on $(0, \infty) \setminus \{\pi^2\}$. The third derivative of F admits a jump discontinuity at π^2.

This phase transition is not one that is easily detected numerically, as Figure 8 shows.
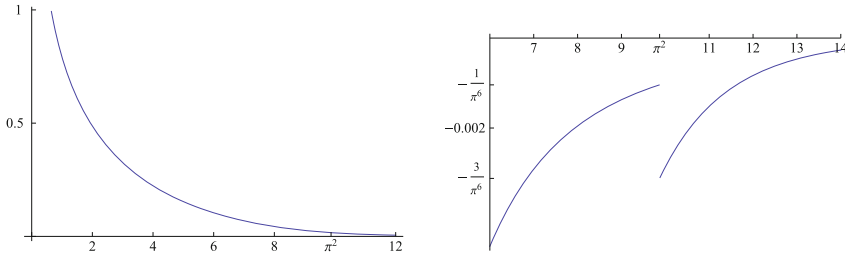
Fig. 8 The graphs of $T \mapsto F(T)$ (on the left) and of $T \mapsto F^{(3)}(T)$ near $T = \pi^2$ (on the right)

3 The Makeenko–Migdal Equations

3.1 First Approach

It is now time that we discuss the equations discovered by Yuri Makeenko and Alexander Migdal and which give their title to these notes. These equations are a powerful tool for the study of the Wilson loop expectations of which we gave a few examples in the previous section. They are related to the approach that we called dynamical, in which an expectation of the form $\mathbb{E}[\mathrm{tr}(H_\ell)]$, where ℓ is some nice loop on a surface M, is seen as a function of the areas of the faces cut by ℓ on the surface M. The Makeenko–Migdal equations give a remarkably elegant expression of the alternated sum of the derivatives of $\mathbb{E}[\mathrm{tr}(H_\ell)]$ with respect to the areas of the four faces that surround a generic point of self-intersection of ℓ. This expression is of the form $\mathbb{E}[\mathrm{tr}(H_{\ell'})\mathrm{tr}(H_{\ell''})]$, where ℓ' and ℓ'' are two loops obtained from ℓ by a very simple operation at this point of self-intersection ℓ. This operation consists in taking the two incoming strands of ℓ at this point and connecting them with the two outgoing strands in the 'other' way, the way that is not realised by ℓ, see Figure 9.

On this figure, we see four faces around the self-intersection point, which need not be pairwise distinct. We denote their areas by t_1, t_2, t_3, t_4 as indicated on Figure 9. The Makeenko–Migdal equation in this case reads

$$\left(\frac{\partial}{\partial t_1} - \frac{\partial}{\partial t_2} + \frac{\partial}{\partial t_3} - \frac{\partial}{\partial t_4} \right) \mathbb{E}[\mathrm{tr}(H_\ell)] = \mathbb{E}[\,\mathrm{tr}(H_{\ell'})\mathrm{tr}(H_{\ell''})] \qquad \text{(MM)}$$

The relation (30), that we derived earlier in an elementary way, is an instance of this equation.

The relation (MM) would become particularly useful if we could combine it with a result saying that $\mathbb{E}[\mathrm{tr}(H_{\ell'})\mathrm{tr}(H_{\ell''})] = \mathbb{E}[\mathrm{tr}(H_{\ell'})]\mathbb{E}[\mathrm{tr}(H_{\ell''})]$. A crucial fact is that this equality, which is of course false in general, becomes true in the large N limit in all cases where this limit has been studied, that is, on the plane and on the sphere. It corresponds to a concentration phenomenon, namely to the fact that the complex-valued random variable $\mathrm{tr}(H_\ell)$ converges, in the large N limit, to a deterministic
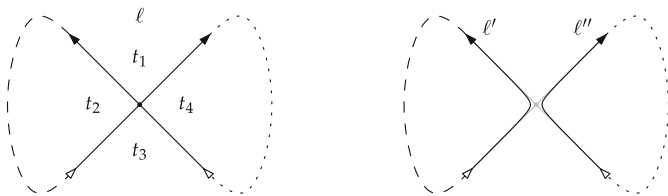
Fig. 9 On the left, we see a loop ℓ around a generic self-intersection point. The dotted and dashed part of ℓ can be arbitrarily complicated, and can meet many times outside the small region of the surface that we are focusing on. It is nevertheless true that after escaping this small region through the North-East corner (resp. North-West corner), the first time ℓ comes back is through the South-East corner (resp. South-West corner). This is why the 'desingularisation' operation illustrated on the right produces exactly two loops, that we call ℓ' and ℓ''

complex, indeed real number $\Phi(\ell)$. This behaviour is expected to occur on any compact surface, and the function $\Phi : \mathscr{L}(M) \to \mathbb{R}$, whose existence has so far been proved when M is the plane or the sphere, is called the *master field*.

In the large N limit, the Makeenko–Migdal equation (MM) becomes a kind of differential equation satisfied by this master field Φ:

$$\left(\frac{\partial}{\partial t_1} - \frac{\partial}{\partial t_2} + \frac{\partial}{\partial t_3} - \frac{\partial}{\partial t_4} \right) \Phi(\ell) = \Phi(\ell')\Phi(\ell'') \qquad (\text{MM}_\infty)$$

On the plane, we will see that this equation, together with the very simple equation (19), essentially characterises the function Φ.

3.2 Makeenko and Migdal's Proof

Makeenko and Migdal discovered the relation (MM), and the extensions that we will describe later, by doing a very clever integration by parts in the functional integral with respect to the Yang–Mills measure (see (3)) that defines a Wilson loop expectation:

$$\mathbb{E}[\text{tr}(H_\ell)] = \frac{1}{Z} \int_{\mathscr{A}} \text{tr}(\text{hol}(\omega, \ell)) e^{-\frac{1}{2} S_{\text{YM}}(\omega)} \, d\omega$$

or instead, as we will explain, in a closely related integral (see [31] and [9]). That this integration by parts performed in an ill-defined integral yields as a final product a perfectly meaningful formula, makes Makeenko and Migdal's original derivation the more intriguing. It is described in mathematical language in the introduction of [24], but this derivation is so beautiful that we reproduce its description here.

The finite-dimensional prototype of the so-called Schwinger–Dyson equations, obtained by integration by parts in functional integrals, is the fact that for all smooth

function $f : \mathbb{R}^n \to \mathbb{R}$ with bounded differential, and for all $h \in \mathbb{R}^n$, the equality

$$\int_{\mathbb{R}^n} d_x f(h) e^{-\frac{1}{2}\|x\|^2} \, dx = \int_{\mathbb{R}^n} \langle x, h \rangle f(x) e^{-\frac{1}{2}\|x\|^2} \, dx$$

holds. This equality ultimately relies on the invariance by translation of the Lebesgue measure on $\mathbb{R}^n$ and it can be proved by writing

$$0 = \frac{d}{dt}_{|t=0} \int_{\mathbb{R}^n} f(x + th) e^{-\frac{1}{2}\|x+th\|^2} \, dx$$

In our description of the Yang–Mills measure μ_{YM} (see (3)), we mentioned that the measure $d\omega$ on the space $\mathscr{A}$ of connections was meant to be a kind of Lebesgue measure, invariant by translations. This is the key to the derivation of the Schwinger–Dyson equations, as we will now explain. In what follows, we will use the differential geometric language introduced in Section 1.2.

Let $\psi : \mathscr{A} \to \mathbb{R}$ be an observable, that is, a function. In general, we are interested in the integral of ψ with respect to the measure μ_{YM}. The tangent space to the affine space $\mathscr{A}$ is the linear space $\Omega^1(M) \otimes \mathrm{Ad}(P)$. To say that the measure $d\omega$ is translation invariant means that for every element η of this linear space,

$$0 = \frac{d}{dt}_{|t=0} \int_{\mathscr{A}} \psi(\omega + t\eta) e^{-\frac{1}{2} S_{\mathsf{YM}}(\omega+t\eta)} \, d\omega$$

and the Schwinger–Dyson equations follow in their abstract form

$$\int_{\mathscr{A}} d_\omega \psi(\eta) \, d\mu_{\mathsf{YM}}(\omega) = \frac{1}{2} \int_{\mathscr{A}} \psi(\omega) d_\omega S_{\mathsf{YM}}(\eta) \, d\mu_{\mathsf{YM}}(\omega) \qquad (43)$$

The directional differential of the Yang–Mills action is well known (see for example [3]) and most easily expressed using the covariant exterior differential $d^\omega : \Omega^0(M) \otimes \mathrm{Ad}(P) \to \Omega^1(M) \otimes \mathrm{Ad}(P)$ defined by $d^\omega \alpha = d\alpha + [\omega \wedge \alpha]$. It is given by

$$d_\omega S_{\mathsf{YM}}(\eta) = 2 \int_M \langle \eta \wedge d^\omega * \Omega \rangle$$

The problem is now to apply this formula to a well-chosen observable ψ and to differentiate in the right direction.

Given a loop ℓ on M, Makeenko and Migdal applied (43) to the observable defined by choosing a skew-Hermitian matrix $X \in \mathfrak{u}(N)$ and setting, for all $\omega \in \mathscr{A}$,

$$\psi_X(\omega) = \mathrm{Tr}(X \, \mathrm{hol}(\omega, \ell)) \qquad (44)$$

To make this definition perfectly meaningful, one needs to choose a reference point in the fibre of P over the base point of ℓ: we will assume that such a point has been chosen and fixed, and compute holonomies with respect to this point.

Let us choose a parametrisation $\ell : [0, 1] \to M$ of ℓ. The directional derivative of the observable ψ_X in the direction of a 1-form $\eta \in \Omega^1(M) \otimes \mathrm{Ad}(P)$ is given by

$$d_\omega \psi_X(\eta) = -\int_0^1 \mathrm{Tr}\left(X \, \mathrm{hol}(\omega, \ell_{[s,1]}) \eta(\dot{\ell}(s)) \mathrm{hol}(\omega, \ell_{[0,s]})\right) \, ds \tag{45}$$

where we denote by $\ell_{[a,b]}$ the restriction of ℓ to the interval $[a, b]$.[28]

One must now choose the direction of differentiation η. Let us assume that ℓ is a nice loop which around each point of self-intersection looks like the left half of Figure 9. Let us assume that for some $s_0 \in (0, 1)$, we have $\ell(s_0) = \ell(0)$ and $\det(\dot{\ell}(0), \dot{\ell}(s_0)) = 1$. Makeenko and Migdal choose for η a distributional 1-form supported at the self-intersection point $\ell(0)$, which one could write as[29]

$$\forall m \in M, \forall v \in T_m M, \ \eta_m(v) = \delta_{m,\ell(0)} \det(\dot{\ell}(0), v) X$$

with $\det(\dot{\ell}(0), v)$ denoting the determinant of the two vectors $\dot{\ell}(0)$ and v. With this choice of η, the directional derivative of ψ_X is given by

$$d_\omega \psi_X(\eta) = -\mathrm{Tr}\left(X \, \mathrm{hol}(\omega, \ell_{[s_0,1]}) X \, \mathrm{hol}(\omega, \ell_{[0,s_0]})\right) = -\mathrm{Tr}\left(X \, \mathrm{hol}(\omega, \ell') X \, \mathrm{hol}(\omega, \ell'')\right) \tag{46}$$

where ℓ' and ℓ'' are the loops defined on the right of Figure 9. Recall that $\mathfrak{u}(N)$ is endowed with the invariant scalar product $\langle X, Y \rangle = -N\mathrm{Tr}(XY)$. The directional derivative of the Yang–Mills action is thus given by

$$d_\omega S_{\mathsf{YM}}(\eta) = -2\langle X, (d^\omega * \Omega)(\dot{\ell}(0)) \rangle = -2N\mathrm{Tr}\left(X d^\omega * \Omega(\dot{\ell}(0))\right)$$

or so it seems from a naive computation. We shall soon see that this expression needs to be reconsidered. For the time being, our Schwinger–Dyson equation reads

$$\int_{\mathscr{A}} \mathrm{Tr}\left(X \, \mathrm{hol}(\omega, \ell') X \, \mathrm{hol}(\omega, \ell'')\right) \, d\mu_{\mathsf{YM}}(\omega) = N \int_{\mathscr{A}} \mathrm{Tr}(X \, \mathrm{hol}(\omega, \ell)) \mathrm{Tr}(X \, d^\omega * \Omega(\dot{\ell}(0))) \, d\mu_{\mathsf{YM}}(\omega) \tag{SD$_X$}$$

---

[28] At first glance, (45) may seem to require the choice of a point in $P_{\ell(s)}$ for each s, but in fact it does not, for the way in which the two holonomies and the term $\eta(\dot{\ell}(s))$ would depend on the choice of this point cancel exactly.

[29] It may seem that we are progressively letting go of the intrinsic character of our construction, but the interested reader can check that everything is still geometrically meaningful at this point.

Let us add the equalities (SD_X) obtained by letting X take all the values $X_1, \ldots, X_{N^2}$ of an orthonormal basis of $\mathfrak{u}(N)$. With the scalar product which we chose, the relations[30]

$$\sum_{k=1}^{N^2} \mathrm{Tr}(X_k A X_k B) = -\frac{1}{N}\mathrm{Tr}(A)\mathrm{Tr}(B) \text{ and } \sum_{k=1}^{N^2} \mathrm{Tr}(X_k A)\mathrm{Tr}(X_k B) = -\frac{1}{N}\mathrm{Tr}(AB)$$

(47)

hold for any two matrices A and B, so that we find

$$\int_{\mathscr{A}} \mathrm{tr}\left(\mathrm{hol}(\omega, \ell')\right)\mathrm{tr}(\mathrm{hol}(\omega, \ell'')) \; \mathrm{d}\mu_{\mathsf{YM}}(\omega) = \int_{\mathscr{A}} \mathrm{tr}\left(\mathrm{hol}(\omega, \ell)d^{\omega} * \Omega(\dot{\ell}(0))\right) \; \mathrm{d}\mu_{\mathsf{YM}}(\omega).$$

The left-hand side of this equation is the right-hand side of (MM). The last and most delicate heuristic step is to interpret the right-hand side of this equation. For this, we must understand the term $d^{\omega} * \Omega(\dot{\ell}(0))$ and we do this by combining two facts: the fact that d^{ω} acts by differentiation in the horizontal direction and the fact that $*\Omega$ computes the holonomy along infinitesimal rectangles. We must also remember that this term comes from the computation of the exterior product of the distributional form η with the form $d^{\omega} * \Omega$. It turns out that, instead of a derivative in the horizontal direction with respect to s at $s = 0$, we should think of the difference between the values at 0^+ and at 0^-, which we denote by $\Delta_{|s=0}$.

With all this preparation and, it must be said, a small leap of faith, the right-hand side of the Schwinger–Dyson equation can finally be drawn as follows:

$$\Delta_{|s=0}\frac{d}{d\epsilon}_{|\epsilon=0} \int_{\mathscr{A}} \mathrm{tr}\,\mathrm{hol}\left(\omega, \; \text{⬡}\right) \mathrm{d}\mu_{YM}(\omega)$$

$$= \frac{d}{d\epsilon}_{|\epsilon=0} \int_{\mathscr{A}} \mathrm{tr}\,\mathrm{hol}\left(\omega, \; \text{⬡}\right) \mathrm{d}\mu_{YM}(\omega)$$

$$- \frac{d}{d\epsilon}_{|\epsilon=0} \int_{\mathscr{A}} \mathrm{tr}\,\mathrm{hol}\left(\omega, \; \text{⬡}\right) \mathrm{d}\mu_{YM}(\omega)$$

This is indeed the left-hand side of the Makeenko–Migdal equation (MM).

---

[30]These relations are strictly equivalent to (16). They are, in one form or the other, *the* fundamental fact of all this story.

3.3 The Equations, Their Merits and Demerits

The strategy of proof described in the previous section can be used, and was used by Makeenko and Migdal, to derive equations slightly more general than (MM). Let us indeed consider a collection $\ell_1, \ldots, \ell_n$ of loops on the surface M. We assume that these loops are nice and in generic position, in the sense that every crossing between two portions of these loops, be they two portions of the same loop or portions of two different loops, is a simple transverse intersection. Around such a crossing, we see, as before, four faces of the graph cut on M by $\ell_1, \ldots, \ell_n$, and we label the areas of these faces t_1, t_2, t_3, t_4 as indicated on Figures 9 and 10. The Makeenko–Migdal equations express the alternated sum of the derivatives with respect to t_1, t_2, t_3, t_4 of $\mathbb{E}[\mathrm{tr}(H_{\ell_1}) \ldots \mathrm{tr}(H_{\ell_n})]$. The equations come in two variants, depending on whether the crossing is between two strands of the same loop (let us call this the case I) or between strands of two distinct loops (the case II). In the case II, where the crossing is between strands of two distinct loops, say ℓ_1 and ℓ_2, the same desingularisation operation explained at the beginning of Section 3.1 gives rise to one new loop ℓ_{12}, as explained in Figure 10.

Calling, in all cases, ℓ_1 the loop containing the South-West – North-East strand, one should replace the observable ψ_X defined in (44) by

$$\psi_X(\omega) = \mathrm{Tr}(X\mathrm{hol}(\omega, \ell_1))\mathrm{Tr}(\mathrm{hol}(\omega, \ell_2)) \ldots \mathrm{Tr}(\mathrm{hol}(\omega, \ell_n))$$

Then the directional derivative of ψ_X is given by

$$d_\omega \psi_X(\eta) = \begin{vmatrix} \mathrm{Tr}\big(X \, \mathrm{hol}(\omega, \ell') X \, \mathrm{hol}(\omega, \ell'')\big)\mathrm{Tr}(\mathrm{hol}(\omega, \ell_2)) \ldots \mathrm{Tr}(\mathrm{hol}(\omega, \ell_n)) & \text{(case I)} \\ \mathrm{Tr}(X \, \mathrm{hol}(\omega, \ell_1))\mathrm{Tr}(X \, \mathrm{hol}(\omega, \ell_2))\mathrm{Tr}(\mathrm{hol}(\omega, \ell_3)) \ldots \mathrm{Tr}(\mathrm{hol}(\omega, \ell_n)) & \text{(case II)} \end{vmatrix}$$

Then, the key to the computation is, as always, given by the equations (47). The final result, with the current notation, is the following.

Theorem 3.1 (Makeenko–Migdal equations) *Let $\ell_1, \ldots, \ell_n$ be nice loops on M in generic position. Consider a crossing point of two strands of ℓ_1 (case I) or of one strand of ℓ_1 and one strand of ℓ_2 (case II). Let t_1, t_2, t_3, t_4 denote the areas of the*
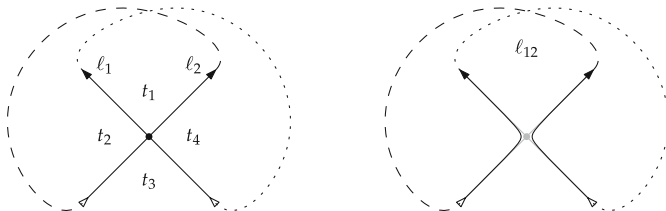

Fig. 10 When performed at a crossing of two distinct loops ℓ_1 and ℓ_2, the operation of reconnecting the incoming and outgoing strands in the other way that is consistent with orientation produces, from ℓ_1 and ℓ_2, one bigger loop that we denote by ℓ_{12}

four faces around this crossing point, as illustrated on Figures 9 and 10. Then, with the notation of these figures,

$$\left(\frac{\partial}{\partial t_1} - \frac{\partial}{\partial t_2} + \frac{\partial}{\partial t_3} - \frac{\partial}{\partial t_4}\right)\mathbb{E}[tr(H_{\ell_1})\dots tr(H_{\ell_n})] = \begin{vmatrix} \mathbb{E}[tr(H_{\ell'})tr(H_{\ell''})tr(H_{\ell_2})\dots tr(H_{\ell_n})] \ \text{(I)} \\ \frac{1}{N^2}\mathbb{E}[tr(H_{\ell_{12}})tr(H_{\ell_3})\dots tr(H_{\ell_n})] \quad \text{(II)} \end{vmatrix}$$

It is understood that if two of the four faces around the crossing under consideration are identical, then the corresponding derivative should be taken twice. Moreover, in the case where $M = \mathbb{R}^2$, any term corresponding to the derivative with respect to the area of the unbounded face should be ignored.

Makeenko and Migdal's original paper on this subject is [31]. The first mathematical proof of the equations was given in [24]. It was rather long and convoluted, and restricted to the case where the surface M is the plane $\mathbb{R}^2$. Three very short and elegant proofs of the equations were then given, still for the case of the plane, by Bruce Driver, Brian Hall and Todd Kemp in [8]. Immediately after, the same team joined by Franck Gabriel proved in [7] that the equations hold on any compact surface. There is little point in reproducing here the content of these beautiful papers. Let us simply emphasise that the fundamental computations remain those summarised in (47).

In addition to their simplicity, the Makeenko–Migdal equations have one major quality which is the fact that the collection of loops appearing in the right-hand side has one crossing less compared with the original collection of loops. Indeed, the operation of desingularisation replaces the crossing where it takes place by a tangential contact which, to the price of an arbitrarily small deformation of the loops, can be suppressed. This suggests the possibility of a recursive computation of Wilson loop expectations. We will explain in the next section that it is indeed possible to use the Makeenko–Migdal equations to set up a recursive computation of the large N limit of Wilson loop expectations.

What the Makeenko–Migdal do not do however, is to give a simple formula for the derivative of a Wilson loop expectation with respect to the area of a single face of the graph traced by a given configuration of loops. Only very special linear combinations of these derivatives are accessible. Of course, unless one is working on the plane, the total area of the surface is prescribed and the best one could hope for is a formula describing the variation of the Wilson loop expectations under an arbitrary variation of the areas of the faces that preserves the total area. However, this is, in general, not given by the Makeenko–Migdal equations, see for example Figure 11.

It is, in fact, not too difficult to understand what information is available in the Makeenko–Migdal equations. Let us consider n loops $\ell_1, \dots, \ell_n$ on our surface M. Let $F_1, \dots, F_r$ denote the faces of the graph traced by these loops. Let us identify a vector $(c_1, \dots, c_r)$ of the vector space $\mathbb{R}^r$ with the linear combination of derivatives

$$c_1 \frac{\partial}{\partial |F_1|} + \dots + c_r \frac{\partial}{\partial |F_r|}$$
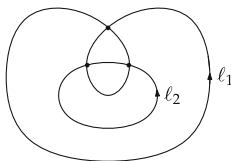
Fig. 11 Consider this configuration of two loops on a sphere. It has five faces and three vertices. Moreover, of the three instances of the Makeenko–Migdal equations, two compute the same linear combination of derivatives. There is no hope that the Makeenko–Migdal equations alone will allow one to compute the corresponding Wilson loop expectation

acting on Wilson loop expectations. Let us define the linear subspace $M \subset \mathbb{R}^r$ generated by the linear combinations given by the Makeenko–Migdal equations applied at each crossing of the loops $\ell_1, \ldots, \ell_n$. This subspace M is of course contained in the hyperplane $\mathbb{R}^r_0$ of equation $c_1 + \ldots + c_r = 0$. Every element of $\mathbb{R}^r$ can naturally be identified with a function on M that is constant on each face of the graph. To each loop ℓ_i, we can associate the unique element n_{ℓ_i} of $\mathbb{R}^r_0$ which, as a function on M, varies by 1 across ℓ_i[31] and is constant across every other loop. This function is a substitute for the winding number of the loop ℓ_i on the surface M.

It is not difficult to check that it is equivalent, for an element of $\mathbb{R}^r$, to be orthogonal, for the simplest scalar product, to the subspace M, or to have a constant jump across every loop, the constant possibly depending on the loop. A more formal statement is the following. We denote by $\mathsf{1}$ the vector $(1, \ldots, 1)$.

Proposition 3.2 *In $\mathbb{R}^r$, one has the equality of linear subspaces*

$$M = \mathrm{Vect}(\mathsf{1}, \mathsf{n}_{\ell_1}, \ldots, \mathsf{n}_{\ell_n})^\perp$$

In particular, $\dim M = \dim \mathbb{R}^r_0 - n$.

The greater the number of loops, the worse the situation. Even with one single loop, we see that all the information about the Wilson loop expectations is not contained in the Makeenko–Migdal equations.

It is time to turn to a case where things improve drastically, namely the large N limit of the Wilson loop expectations.

3.4 The Master Field on Compact Surfaces

We saw in Section 2 that when $G = \mathrm{U}(N)$, Wilson loop expectations tend to take simpler forms in the limit where N tends to infinity (compare for example (20) and (21)). We also observed some instances of a property of factorisation, see for

---

[31] A convention must be chosen regarding the definition of a positive crossing of ℓ_i.

example (26). The factorisation is due to a phenomenon of concentration, with the effect that, as N tends to infinity, and provided one scales the scalar product on $\mathfrak{u}(N)$ correctly (which we did), the Wilson loop functionals, that is, the normalised traces of the random holonomies, become deterministic. The limit is thus a number depending on a loop, and this function is relatively simple, at least when one is working on the plane, because it satisfies, and is essentially determined, by the Makeenko–Migdal equations.

The main theorem of convergence is the following.

Theorem 3.3 (Master field) *Let M be either the plane $\mathbb{R}^2$ or the sphere S^2. For each $N \geqslant 1$, let $(H_{N,\ell})_{\ell \in \mathscr{L}(M)}$ be the Yang–Mills holonomy process on M with structure group $G = \mathrm{U}(N)$, and with scalar product $\langle X, Y \rangle = N\mathrm{Tr}(X^*Y)$ on $\mathfrak{u}(N)$. Then for every loop $\ell \in \mathscr{L}(M)$, the convergence of complex-valued random variables*

$$\mathrm{tr}(H_{N,\ell}) \xrightarrow[N \to \infty]{P} \Phi(\ell) \tag{48}$$

holds in probability, towards a deterministic real limit.

This theorem was proved in [24] in the case of the plane, and in [5] in the case of the sphere, see also [17]. In the case of the plane, which is simpler, it is also known that the convergence occurs quickly, in the sense that the series $\sum_{N \geqslant 1} \mathrm{Var}(\mathrm{tr}(H_{N,\ell}))$ converges. Thus, the convergence (48) holds almost surely. The conclusion is also known to be true if one replaces the unitary group by the special unitary group, the special orthogonal group, or the symplectic group.

It is expected that Theorem 3.3 is true on any compact surface, but a proof of this fact still has to be given.

In any case, when this theorem holds, the aforementioned asymptotic factorisation takes place, in the sense that for all loops $\ell_1, \ldots, \ell_n$,

$$\lim_{N \to \infty} \mathbb{E}[\mathrm{tr}(H_{\ell_1}) \ldots \mathrm{tr}(H_{\ell_n})] = \lim_{N \to \infty} \mathbb{E}[\mathrm{tr}(H_{\ell_1})] \ldots \lim_{N \to \infty} \mathbb{E}[\mathrm{tr}(H_{\ell_n})] = \Phi(\ell_1) \ldots \Phi(\ell_n)$$

The function $\Phi : \mathscr{L}(M) \to \mathbb{R}$ which appears in (48) is called the *master field*. This is a continuous function with respect to the convergence of loops with fixed endpoints (see the beginning of Section 1.4.4) and it satisfies, crucially, the Makeenko–Migdal equation (MM_∞), which is all that there is left of the full set of equations stated in Theorem 3.1 as N tends to infinity.

Theorem 3.4 *Assume that M is either the plane $\mathbb{R}^2$ or the sphere S^2. The function $\Phi : \mathscr{L}(M) \to \mathbb{R}$ is the unique function that is continuous, invariant under area-preserving diffeomorphisms, satisfying the Makeenko–Migdal equation (MM_∞) and such that for every simple loop ℓ enclosing a domain of area t, one has, depending on whether M is the plane or a sphere of total area T,*

$$\Phi(\ell) = e^{-\frac{t}{2}} \tag{$M = \mathbb{R}^2$}$$

or

$$\Phi(\ell) = \frac{1}{\pi} \int_{\mathbb{R}} \cosh\left(\frac{x}{2}(T - 2t)\right) \sin(\pi\rho_T(x)) \, \mathrm{d}x \qquad (M = S^2)$$

3.5 A Value of the Master Field on the Plane

As a conclusion to these notes, we give an example of computation of a value of the master field Φ on the plane, and choose an example that is not listed at the end of [24]. We choose the loop ℓ represented on the left half of Figure 12.

Although we did not include this in our description of the function Φ on the plane $\mathbb{R}^2$, it is not difficult to check that the derivative of Φ of any loop with respect to the area of a face adjacent to the unbounded face is equal to $-\frac{1}{2}$ times the value of Φ on this loop. This factor $-\frac{1}{2}$ comes of course from the stochastic differential equation (15) satisfied by the Brownian motion on $\mathrm{U}(N)$.

Given the value of Φ on simple loops and (29), the Makeenko–Migdal equation applied to the vertex of ℓ_0 that is marked in Figure 12 yields

$$(2\partial_s - \partial_{t_2})\Phi(\ell_0) = (-1 - \partial_{t_2})\Phi(\ell_0) = e^{-\frac{s}{2} - t_1 - t_2}(1 - t_1)$$

which is solved in

$$\Phi(\ell_0) = e^{-\frac{s}{2} - t_1 - t_2}(1 - t_1)(1 - t_2)$$

If we can determine $\partial_u\Phi(\ell)$ explicitly, we are done, since $\Phi(\ell_0)$ is exactly the value of $\Phi(\ell)$ at $u = 0$. Applying the Makeenko–Migdal equations at the three marked vertices in Figure 12 yields the derivatives $(\partial_{s_1} + \partial_{s_2} - \partial_{t_2})\Phi(\ell)$, $(\partial_{s_1} + \partial_{s_2} - \partial_{t_1})\Phi(\ell)$, and $(\partial_{t_1} + \partial_{t_2} - \partial_{s_2} - \partial_u)\Phi(\ell)$. Adding the three expressions and using the fact that $\partial_{s_1}\Phi(\ell) = \partial_{s_2}\Phi(\ell) = -\frac{1}{2}\Phi(\ell)$, we find
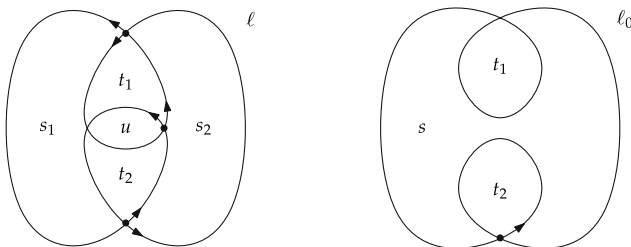

Fig. 12 We are interested in computing $\Phi(\ell)$. The strategy is to use the Makeenko–Migdal equations to compute $\partial_u\Phi(\ell)$. As $u = 0$, the two inner windings of ℓ disentangle, and ℓ becomes identical to ℓ_0. This loop ℓ_0 is similar to the loop that we studied in Section 2.3, and becomes exactly this loop when $t_2 = 0$. Our first task is thus to compute $\partial_{t_2}\Phi(\ell_0)$

$$\left(-\frac{3}{2} - \partial_u\right)\Phi(\ell) = e^{-\frac{s_1+s_2}{2} - t_1 - t_2 - \frac{3u}{2}}(3 - t_1 - t_2 - u)$$

and finally

$$\Phi(\ell) = e^{-\frac{s_1+s_2}{2} - (t_1+t_2) - \frac{3u}{2}}\left(\frac{u^2}{2} + (t_1 + t_2 - 3)u + (1 - t_1)(1 - t_2)\right)$$

(49)

Evaluating this expression with $s_1 = s_2 = t_1 = t_2 = 0$ yields the large N limit of the third moment of the unitary Brownian motion at time u, as expressed by (22) with $n = 3$. This is consistent with the fact that shrinking all faces but the face of area u reduces ℓ to a loop winding three times around a simple domain of area u.

References

1. M.F. Atiyah, R. Bott, The Yang-Mills equations over Riemann surfaces. Philos. Trans. R. Soc. Lond. A **308**, 523–615 (1983). https://doi.org/10.1098/rsta.1983.0017
2. P. Biane, Free Brownian motion, free stochastic calculus and random matrices, in *Free Probability Theory (Waterloo, ON, 1995)*. Fields Institute Communications, vol. 12 (American Mathematical Society, Providence, 1997), pp. 1–19
3. D. Bleecker, *Gauge Theory and Variational Principles*. Global Analysis Pure and Applied Series A, vol. 1 (Addison-Wesley, Reading, 1981), xviii+179
4. I. Chevyrev, Yang-mills measure on the two-dimensional torus as a random distribution (2018). http://arxiv.org/abs/arXiv:1808.09196
5. A. Dahlqvist, J.R. Norris, Yang-mills measure and the master field on the sphere (2017). http://arxiv.org/abs/arXiv:1703.10578
6. B.K. Driver, YM$_2$: continuum expectations, lattice convergence, and lassos. Commun. Math. Phys. **123**(4), 575–616 (1989)
7. B.K. Driver, F. Gabriel, B.C. Hall, T. Kemp, The Makeenko-Migdal equation for Yang-Mills theory on compact surfaces. Commun. Math. Phys. **352**(3), 967–978 (2017). https://doi.org/10.1007/s00220-017-2857-2
8. B.K. Driver, B.C. Hall, T. Kemp, Three proofs of the Makeenko-Migdal equation for Yang-Mills theory on the plane. Commun. Math. Phys. **351**(2), 741–774 (2017). https://doi.org/10.1007/s00220-016-2793-6
9. A. Dubin, Y. Makeenko, Loop equations and nonperturbative QCD, in *At the Frontier of Particle Physics*, ed. by S. Misha. Handbook of QCD. Boris Ioffe Festschrift, vol. 4 (World Scientific, Singapore, 2002), pp. 2479–2525
10. B. Durhuus, On the structure of gauge invariant classical observables in lattice gauge theories. Lett. Math. Phys. **4**(6), 515–522 (1980). https://doi.org/10.1007/BF00943439

11. W.M. Goldman, The symplectic nature of fundamental groups of surfaces. Adv. Math. **54**, 200–225 (1984). https://doi.org/10.1016/0001-8708(84)90040-9
12. W.M. Goldman, The symplectic geometry of affine connections on surfaces. J. Reine Angew. Math. **407**, 126–159 (1990). https://doi.org/10.1515/crll.1990.407.126
13. L. Gross, A Poincaré lemma for connection forms. J. Funct. Anal. **63**(1), 1–46 (1985)
14. L. Gross, The Maxwell equations for Yang-Mills theory, in *Mathematical Quantum Field Theory and Related Topics (Montreal, PQ, 1987)*. CMS Conference Proceedings, vol. 9 (American Mathematical Society, Providence, 1988), pp. 193–203
15. L. Gross, C. King, A. Sengupta, Two-dimensional Yang-Mills theory via stochastic differential equations. Ann. Phys. **194**(1), 65–112 (1989)
16. A. Guionnet, M. Maïda, Character expansion method for the first order asymptotics of a matrix integral. Probab. Theory Related Fields **132**(4), 539–578 (2005). https://doi.org/10.1007/s00440-004-0403-6
17. B.C. Hall, The large-N limit for two-dimensional Yang-Mills theory. Comm. Math. Phys. **363**(3), 789–828 (2018). https://doi.org/10.1007/s00220-018-3262-1
18. N.J. Hitchin, The Yang-Mills equations and the topology of 4-manifolds (after Simon K. Donaldson), in *Bourbaki seminar, Vol. 1982/83*. Astérisque, vol. 105 (Soc. Math. France, Paris, 1983), pp. 167–178
19. N.J. Hitchin, The self-duality equations on a Riemann surface. Proc. Lond. Math. Soc. (3) **55**(1), 59–126 (1987). https://doi.org/10.1112/plms/s3-55.1.59
20. C. King, A. Sengupta, The semiclassical limit of the two-dimensional quantum Yang-Mills model. J. Math. Phys. **35**(10), 5354–5361 (1994). https://doi.org/10.1063/1.530756
21. S. Kobayashi, K. Nomizu, *Foundations of Differential Geometry. Vol. I.* Wiley Classics Library (Wiley, New York, 1996), xii+329. Reprint of the 1963 original, A Wiley-Interscience Publication
22. T. Lévy, Wilson loops in the light of spin networks. J. Geom. Phys. **52**(4), 382–397 (2004). https://doi.org/10.1016/j.geomphys.2004.04.003
23. T. Lévy, Two-dimensional Markovian holonomy fields. Astérisque **329**, 172 (2010)
24. T. Lévy, The master field on the plane. Astérisque **388**, ix+201 (2017)
25. T. Lévy, M. Maïda, On the Douglas-Kazakov phase transition. Weighted potential theory under constraint for probabilists, in *Modélisation Aléatoire et Statistique—Journées MAS 2014*. ESAIM Proceedings Surveys, vol. 51 (EDP Sci., Les Ulis, 2015), pp. 89–121. https://doi.org/10.1051/proc/201551006
26. T. Lévy, A. Sengupta, Four chapters on low-dimensional gauge theories, in *Stochastic Geometric Mechanics. CIB, Lausanne, Switzerland, January–June 2015* (Springer, Cham, 2017), pp. 115–167
27. K. Liechty, D. Wang, Nonintersecting Brownian motions on the unit circle. Ann. Probab. **44**(2), 1134–1211 (2016). https://doi.org/10.1214/14-AOP998
28. K. Liu, Heat kernel and moduli space. Math. Res. Lett. **3**(6), 743–762 (1996). https://doi.org/10.4310/MRL.1996.v3.n6.a3
29. K. Liu, Heat kernel and moduli spaces. II. Math. Res. Lett. **4**(4), 569–588 (1997). https://doi.org/10.4310/MRL.1997.v4.n4.a12
30. R.C. Lyndon, P.E. Schupp, *Combinatorial Group Theory*. Classics in Mathematics (Springer, Berlin, 2001), xiv+339. Reprint of the 1977 edition
31. Y. Makeenko, A.A. Migdal, Exact equation for the loop average in multicolor QCD. Phys. Lett. B **88B**, 135 (1979)
32. A.A. Migdal, Recursion equations in gauge field theories. Sov. Phys. JETP **42**(3), 413–418 (1975)
33. S. Morita, *Geometry of Differential Forms*. Translations of Mathematical Monographs, vol. 201 (American Mathematical Society, Providence, 2001), xxiv+321. Translated from the two-volume Japanese original (1997, 1998) by Teruko Nagase and Katsumi Nomizu, Iwanami Series in Modern Mathematics

34. A. Nica, R. Speicher, *Lectures on the Combinatorics of Free Probability*. London Mathematical Society Lecture Note Series, vol. 335 (Cambridge University Press, Cambridge, 2006), xvi+417. https://doi.org/10.1017/CBO9780511735127

35. D. Puder, O. Parzanchevski, Measure preserving words are primitive. J. Amer. Math. Soc. **28**(1), 63–97 (2015). https://doi.org/10.1090/S0894-0347-2014-00796-7

36. J. Råde, On the Yang-Mills heat equation in two and three dimensions. J. Reine Angew. Math. **431**, 123–163 (1992). https://doi.org/10.1515/crll.1992.431.123

37. E.M. Rains, Combinatorial properties of Brownian motion on the compact classical groups. J. Theoret. Probab. **10**(3), 659–679 (1997). https://doi.org/10.1023/A:1022601711176

38. E.B. Saff, V. Totik, *Logarithmic Potentials with External Fields*. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 316 (Springer, Berlin, 1997), xvi+505. Appendix B by Thomas Bloom. https://doi.org/10.1007/978-3-662-03329-6

39. A. Sengupta, Gauge invariant functions of connections. Proc. Amer. Math. Soc. **121**(3), 897–905 (1994). https://doi.org/10.2307/2160291

40. A.N. Sengupta, Gauge theory on compact surfaces. Mem. Amer. Math. Soc. **126**(600), viii+85 (1997)

41. H. Shen, Stochastic quantization of an abelian gauge theory (2018). http://arxiv.org/abs/arXiv:1801.04596

42. I.M. Singer, On the master field in two dimensions, in *Functional Analysis on the Eve of the 21st Century, Vol. 1 (New Brunswick, NJ, 1993)*. Progress in Mathematics, vol. 131 (Birkhäuser Boston, Boston, 1995), pp. 263–281

43. M. Spivak, *A Comprehensive Introduction to Differential Geometry. Vol. I*, 2nd edn. (Publish or Perish, Inc., Wilmington, 1979), xiv+668.

44. K.K. Uhlenbeck, Connections with L^p bounds on curvature. Commun. Math. Phys. **83**(1), 31–42 (1982)

45. E. Witten, On quantum gauge theories in two dimensions. Commun. Math. Phys. **141**(1), 153–209 (1991). https://doi.org/10.1007/BF02100009

46. E. Witten, Two dimensional gauge theories revisited. J. Geom. Phys. **9**(4), 303–368 (1992). https://doi.org/10.1016/0393-0440(92)90034-X

47. F. Xu, A random matrix model from two-dimensional Yang-Mills theory. Commun. Math. Phys. **190**(2), 287–307 (1997)

48. C.N. Yang, R.L. Mills, Conservation of isotopic spin and isotopic gauge invariance. Phys. Rev. II. Ser. **96**, 191–195 (1954)