Limit shapes, real and imagined

Andrei Okounkov

1 Introduction

1.1

We are surrounded by random surfaces. In fact, every shape around us is random on a sufficiently small scale. The definite shapes that we perceive are manifestations of the *law of large numbers* that makes averages much larger than the fluctuations around them. Same law is at work e.g. with a balloon floating through the air. Our eyes trace out a smooth trajectory while air molecules hit the balloon randomly from all directions.

Can mathematical physics link the microscopic dynamics to the resulting macroscopic shape, say, for objects of inorganic natural origin ? The endless variety of snowflake shapes produced by commonplace water under ordinary conditions illustrates how challenging this question is.

Yet, at or near *equilibrium* a successful theory explaining the macroscopic shape from the microscopic laws may be developed. It requires deep and powerful mathematics that we can only begin to explore it the course of 3 lectures. In turn, it impacts areas of mathematics that may seem very distant from equilibrium crystals or liquid droplets.

Mathematics and physics are full of random geometric objects, especially when surveyed with a trained eye. In dealing with them, the experience and intuition acquired in the study of equilibrium crystals can be very valuable.

1.2

The main goal of this lecture is to provide an accessible introduction to an area of mathematical physics that I, personally, find captivating. The flipside of accessibility is that we won't be able to get much beyond the first principles. Fortunately, there is extensive literature on most of the topics discussed in these lectures. For example, mathematical results on equilibrium shapes of Ising crystals are covered in the books [6, 11], where great many further references may also be found. For an introduction to the gauge theory problems discussed in the 3rd lecture, see for example [10, 13, 14, 52].

The bulk of these lectures is about computations at zero temperature, reflecting both their introductory nature and the author's expertise. Needless to say, dealing with positive temperature requires an entirely different level of mathematical sophistication.

1.3

The plan of these lectures is as follows. Any discussion of probability has to start by talking about a random walk. We gladly conform to this custom in the opening section of the first lecture. We then discuss the 2-dimensional Ising crystal, and especially about its zero-temperature limit, which is the random walk again.

The second lecture is devoted to zero-temperature interfaces in the 3-dimensional Ising model, also known as stepped surfaces. The limit shapes for these interfaces are diverse and beautiful. For example, in many case they turn out to be thinly disguised algebraic curves.

The third and final lecture discusses an application of limit shape ideas to gauge theory. Again, we limit ourselves to computation at zero temperature, or in the presence of supersymmetry, which may be interpreted as statistical mechanics of a gas of instantons. By the end of the lecture we introduce, following Nekrasov, the instanton partition function.

After explaining a few key points about Nekrasov's conjecture [31] and its proof, we conclude with a brief sampling of directions in which the field has been developing.

1.4

I am very grateful to AMS for the honor to give these lectures, and on this occasion I would like to thank many people who very much influenced my thinking about the subject. I have been very fortunate to be able to learn from G. Olshanski, S. Shlosman, and other colleagues from Dobrushin's laboratory on one side of the Moscow–St. Petersburg railway, and from S. Kerov and A. Vershik — on the other. In particular, the idea that the principles of statistical mechanics can and should be applied in areas of mathematics seemingly very distant from probability or mathematical physics will be forever associated in my mind with A. Vershik. I thank Richard Kenyon and Nikita Nekrasov for the joy of collaboration on the topics discussed in these lectures, and I want to conclude with special words of gratitude to S. Shlosman for helping me improve these notes.

2 The simple random walk

2.1

This must be a very familiar concept. We consider the simple random walk in 1 dimension: a particle starts at the origin $0 \in \mathbb{Z}$ at time zero and with each tick of the clock jumps by ± 1 with probability $\frac{1}{2}$. For example, its position $X \in \mathbb{Z}$ at time T may represent your total gain after trying to guess the outcome of T independent flips of a fair coin. Plotted in space-time, the trajectory represents a random zig-zag curve that may look like the curve in Figure 1.

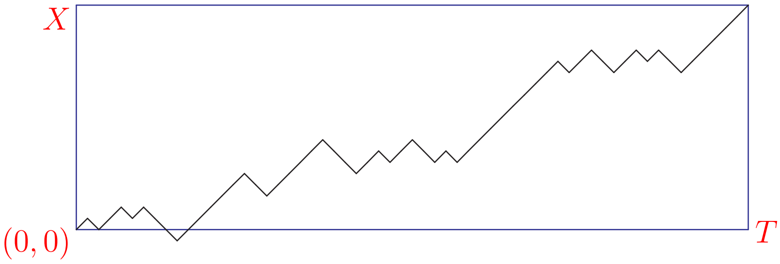

Figure 1: A simple random walk from 0 to X in T steps.

It is well-known that for large times T this random zig-zag curve, rescaled by T in the horizontal and $\sqrt{T}$ vertical direction, converges to remarkable random continuous function — the Wiener process or the 1-dimensional Brownian motion. We, however, are interested in a different, and much simpler, asymptotic regime when the result is elementary and *nonrandom*.

2.2

Fix T and X and look at a random zig-zag curve from $(0,0)$ to (T,X). A probabilist would say that we condition the random walk to reach X in T

steps. Now let $X, T \to \infty$ so that the average velocity $V = X/T$ remains constant. Alternatively, one may keep X and T fixed, while modifying the law of the random walk: it will now make steps

$$(\delta, \pm\delta)$$

in space-time, where the mesh size δ is small.

Let $Z(t)$, $t \in [0, T]$ denote our random zig-zag curve. What happens to it as $\delta \to 0$? It converges to a nonrandom curve, namely a straight line, see Figure 2. This our first example of a law of large numbers.

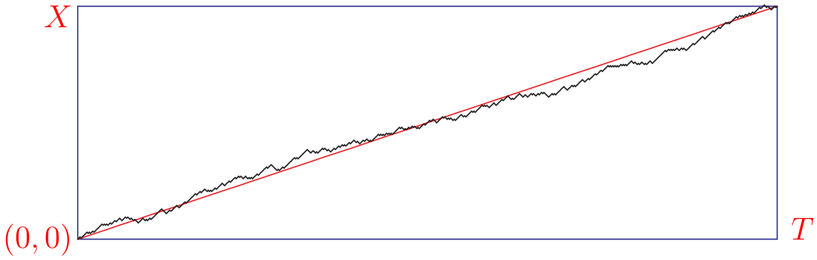

Figure 2: Random walk with small steps converges to a straight line

More formally, any neighborhood $\mathcal{U}$ of the straight line L

$$x = Vt\,, \quad t \in [0, T]\,, \quad V = X/T\,,$$

in the space $C[0, T]$ of continuous functions, has the property that

$$\text{Prob}\,\{Z(t) \in \mathcal{U}\} \to 1\,, \quad \delta \to 0\,.$$

The proof, which may be found in practically any probability textbook, proceeds as follows.

2.3

Since $Z(t)$ is Lipschitz with constant 1, that is,

$$|Z(t_1) - Z(t_2)| \leq |t_1 - t_2|\,, \quad \forall t_1, t_2 \in [0, T],$$

it suffices to prove that for any $\varepsilon > 0$

$$\text{Prob}\,\{\max_i |Z(t_i) - Vt_i| > \varepsilon\} \to 0\,, \quad \delta \to 0\,,$$

for any finite set $\{t_i\} \subset [0,T]$.

Obviously,

$$\operatorname{Prob}\{\max_i |Z(t_i) - Vt_i| > \varepsilon\} \le \sum_i \operatorname{Prob}\{|Z(t_i) - Vt_i| > \varepsilon\},$$

and hence it is enough to show that every term in this sum goes to 0 as $\delta \to 0$. This is a question of counting random walks.

2.4

There are

$$\binom{T}{(X+T)/2}$$

walks reaching X in T steps because such walk in uniquely determined by the choice of $(X+T)/2$ up-steps out of T possible. From Stirling formula, we have

$$\frac{1}{T} \ln \binom{T}{pT} \to S(p), \tag{1}$$

where the function

$$S(p) = -p \ln p - (1-p) \ln(1-p) \tag{2}$$

is known as the *Shannon entropy.*

The probability that $Z(t) = x$ for some $t \in [0,T]$ equals

$$\binom{\delta^{-1}\, t}{\delta^{-1}(x+t)/2} \binom{\delta^{-1}\,(T-t)}{\delta^{-1}(X+T-x-t)/2} \binom{\delta^{-1}\, T}{\delta^{-1}(X+T)/2}^{-1}$$

From (1), we conclude that as $\delta \to 0$

$$\delta \ln \operatorname{Prob}\{Z(t) = x\} \to t\, S(p) + t'\, S(p') - T\, S\left(\frac{t\, p + t'\, p'}{T}\right), \tag{3}$$

where

$$t' = T - t\,, \quad p = \frac{x+t}{2t}\,, \quad p' = \frac{X - x + t'}{2t'}\,.$$

Since S is strictly concave, (3) implies $\operatorname{Prob}\{Z(t) = x\}$ is exponentially small unless $p = p'$, which is equivalent to $x = Vt$. This concludes the proof.

2.5

The above argument may be adopted to estimate the probability of any subset of $C[0,T]$. Let f be Lipschitz with constant 1 and $\mathcal{U}$ a small neighborhood of f. To estimate the number of zig-zag curves in $\mathcal{U}$, we may subdivide $[0,T]$ into subintervals on which f is approximately linear, see Figure 3. This gives the following heuristic count

$$\ln \#\text{curves in a neighborhood of } f \approx \frac{1}{\delta} \int_0^T S\left(\frac{f'+1}{2}\right) dt\,. \tag{4}$$

This "formula" shows that calling S *entropy* is indeed appropriate in the present context, since it gives the asymptotic count of microscopic states, i.e. zig-zag curves, corresponding to the same macroscopic state f.

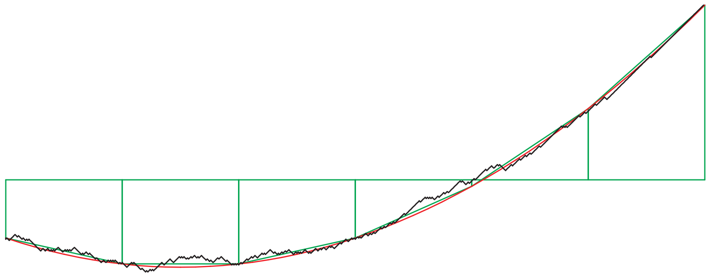

Figure 3: How many zig-zag curves stay close to the graph of f ?

To transform (4) into a meaningful mathematical statement, let $\mathrm{Lip}(1) \subset C[0,T]$ denote Lipschitz functions with constant 1 and let

$$\mathcal{K} \subset C[0,T]$$

consist of functions taking values 0 and X at the respective endpoints. This is a compact set containing all our random curves.

By Rademacher's theorem any $f \in \mathrm{Lip}(1)$ is differentiable almost everywhere and hence the functional

$$\mathcal{S}(f) = \int_0^T S\left(\frac{f'+1}{2}\right) dt - T S\left(\tfrac{1+V}{2}\right) \tag{5}$$

is well-defined on Lip(1). The second term in (5) is a constant that makes $\mathcal{S}$ vanish on a the linear function $x = Vt$. Concavity of S implies concavity and upper-semicontinuity of $\mathcal{S}$

$$\mathcal{S}(\lim f_n) \geq \lim \mathcal{S}(f_n)\,.$$

Now we can say what the statement (4) really means.

Let $A \subset \mathcal{K}$ be arbitrary. Then

$$\lim_{\delta \to 0} \delta \ln \mathrm{Prob}\{Z(t) \in A\} \subset \left[\sup_{A^\circ} \mathcal{S}, \sup_{\overline{A}} \mathcal{S} \right] \tag{6}$$

where lim denotes the set of all possible limit points and A°, $\overline{A}$ stand for interior and closure of A in $\mathcal{K}$, respectively.

When probabilists see a statement like (6), they say that $Z(t)$ satisfies a *large deviation principle*, or LDP for short. The functional $\mathcal{S}$ is called the *rate* (or *action*) functional.

2.6

All limit shapes that we will see in these lecture are found by, first, proving an LDP and then finding the maximum of the action functional. The existence and uniqueness of this maximum often follow from general compactness and convexity considerations, but its actual identification may be tricky.

Note the role played by the *topology* in the formulation of LDP. The finer it is the topology, the stronger is the LDP claim.

2.7

Of important practical interest is the following connection between large deviations and *Central Limit Theorems.* If we expand (3) around $x = Vt$, we see that randomness remains on the scale $\sqrt{t}$ and is a Gaussian random variable. This is forced by the simple fact that the asymptotics of the logarithm of probability is given by a function with a differentiable nondegenerate maximum.

A similar inference in infinite-dimensional setting, e.g. for the functional (5), has certainly a great heuristic appeal and is in common use in natural sciences to describe Gaussian corrections to limit shapes. Sometimes it is easy to turn it into a mathematical theorem and, in the case of a random walk, this correctly reproduces Brownian motion. For random surfaces, in particular for *stepped surfaces* that we will meet in Lecture 2, one then expects the 2-dimensional Gaussian Free Field to appear, and so on. However, already for stepped surfaces, this remains at present a challenging conjecture. It has only been settled for special classes of boundary conditions, see in particular

[20], using certain very specific features of the model and techniques much more delicate than taking second partials of the action functional.

3 The Ising crystal

3.1

The Ising model may be defined and studied in any dimension but dimensions 2 and 3 are more familiar. Originally studied as a model for ferromagnetism, the Ising model may also be interpreted as a model of an equilibrium crystal. We start we the description of the model in two dimensions.

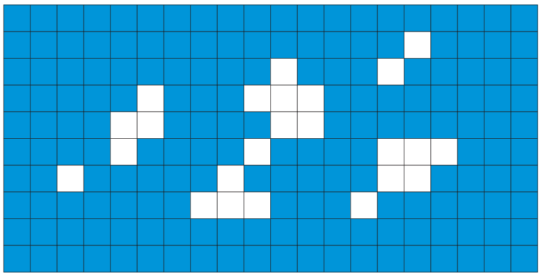

Figure 4: An Ising crystal configuration

Let every square in a rectangle as in Figure 4 be painted white or blue. We think of the white area as crystal surrounded by the blue stuff, whatever it may be. The rectangular container will eventually be taken very big; its exact shape doesn't matter. We fix the total number of white squares and to prevent them from sticking to the sides of the container we insist that all squares along the boundary be blue. This sums up the description of the possible states of the Ising crystal. The nontrivial element in the model is the assignment of probabilities to configurations.

3.2

By definition, a certain system is in thermal equilibrium with the rest of the world at some temperature T if the probability of any particular configuration C decays exponentially with the energy of C, namely

$$\mathrm{Prob}(C) = \frac{1}{Z(T)} \exp\left(-\frac{\mathrm{Energy}(C)}{kT}\right) .$$

Here k is a universal dimensional constant that bears the name of Boltzmann, the inventor of the above law, and

$$Z(T) = \sum_C \exp\left(-\frac{\text{Energy}(C)}{kT}\right) \tag{7}$$

is the normalization factor, known as *partition function*, that makes the probabilities sum to 1. One often prefers to use $\beta = (kT)^{-1}$ as a parameter.

In the Ising model case, the energy is the sum of interaction of all adjacent squares. The contribution of each pair of neighbors depends only on their colors. Since the total number of white squares is fixed, it is proportional to total length of the contours separating white from blue. These contours are plotted in red in Figure 5. The traditional normalization is

$$\text{Energy} = 2 \times \text{Length of contours}\,.$$

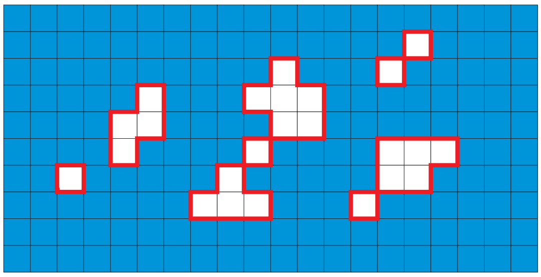

Figure 5: Configuration of Ising crystal may be represented by contours

3.3

In Ising model, as everywhere else in statistical mechanics, energy competes with entropy. White squares save energy by clumping, but low energy doesn't automatically imply high probability. Probability of any event is a sum and the number of terms is as important as the size of the terms. The neatly ordered energy saving configurations have low entropy, that is, there are, relatively speaking, very few of them. The outcome of their competition with disorder depends on the value of β. The larger β, the stronger the preference for order.

For Ising model in any dimension ≥ 2 there exists a certain critical temperature $T_c > 0$ above which disorder wins. In two dimensions, it corresponds to

$$\beta_c = \tfrac{1}{2} \ln \left(\sqrt{2} + 1\right) \approx 0.44 \,.$$

Below this critical temperature and for concentrations of white squares above a certain threshold[1], a crystal indeed forms as the size of the container goes to infinity. The macroscopic shape of this crystal is explicitly known in two dimensions. Up to scale, it is given by

$$\cosh \beta x + \cosh \beta y \leq \cosh^2 2\beta \big/ \sinh 2\beta \,, \tag{8}$$

see [26]. It starts out as a square at $T = 0$ and gets rounder and rounder as the temperature increases. Just below T_c is it a circle, see Figure 6.

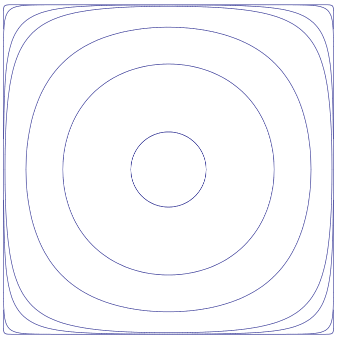

Figure 6: The shape of the Ising crystal for $\beta = .5, .7, 1, 2, 3, 10, 50$

Difficult and beautiful mathematics is required to prove these statements. I hope that the interested reader will open the books [6, 11] describing this story. Here we limit ourselves to the elementary case of $T = 0$.

3.4

At zero temperature, entropy all but disappears. Only configurations that strictly minimize the energy survive. A perfect square, shown in Figure 7, is an example. It certainly looks like a crystal, which is encouraging.

[1] It is shown in [12] that, in the d-dimensional Ising model, one needs on the order of $V^{\frac{d}{d+1}}$ white squares in a volume V to have condensation.

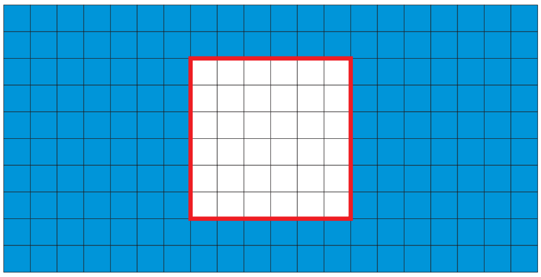

Figure 7: An energy minimizer

But what if the number of white squares is not a perfect square ? For example, if we only had 35 white squares, the possible minimizers will be 6×6 crystals with one of the corner atoms missing.

More generally, let us zoom in on a neighborhood of a corner. In fact, let us assume that the boundary conditions are those of a crystal corner. Namely all squares far enough from the origin are white in the first quadrant and blue in the rest of the plane. We then note that all staircase configurations as in Figure 8 are equally probable.

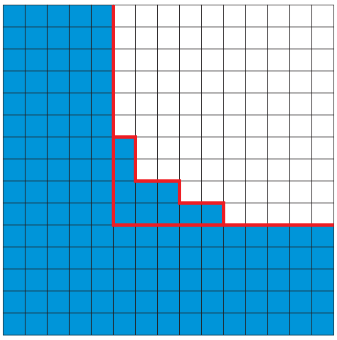

Figure 8: Zero-temperature 2D Ising with crystal corner boundary conditions

Looking from an angle, such staircase shapes are the same as random walks, conditioned to coincide with the graph of $|t|$ far away from $t = 0$. In particular, there will be many minimizers if the number of missing atoms is large. We see that some entropy remains even at zero temperature.

3.5

We may ask ourselves: what happens if the number of missing atoms is really large, while still very small compared to the size of the crystal ? Looking at Figure 8 from a 45° angle again, this amount to finding the limit shape of random walk conditioned to enclose a given area. The tools of in Section 2.5 immediately yield the solution.

The limit shape will maximize the action functional $\mathcal{S}$ for fixed enclosed area. That is, we are solving

$$\int S\left(\frac{f'(t)+1}{2}\right)dt + c\int f(t)\,dt \to \max \tag{9}$$

where c is the Lagrange multiplier for the area constraint.

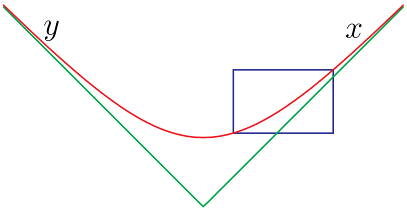

Figure 9: The constrained entropy maximizer

The solution of this variational problem is unique up to translation

$$t \mapsto t + a$$

and scale

$$t \mapsto \lambda t, \quad f \mapsto \lambda f, \quad c \mapsto \lambda^{-1} c\,.$$

In the original crystal corner coordinates x and y, it is given by

$$e^{-x} + e^{-y} = 1\,, \tag{10}$$

see Figure 9. This is the shape of a very large and very cold Ising crystal near its corner. In other words, this is how the region (8) looks when we zoom in on one of its corners by substituting

$$(x, y) \mapsto (2 - x/\beta, 2 - y/\beta)$$

while letting $\beta \to \infty$.

The little rectangular window in Figure 9 refers to the fact that the entropy maximizer with any boundary conditions and area constraint is obtained by finding a window like this that scales to the required parameters.

3.6

The staircase shapes from Figure 8 are known in combinatorics as diagrams of *partitions*. Read along the rows, the diagram in Figure 8 represents the partition

$$\lambda = (5, 3, 1, 1)$$

of the number $|\lambda| = 10$. In particular,

$$\exp\left(-\frac{x}{\pi\sqrt{6}}\right) + \exp\left(-\frac{y}{\pi\sqrt{6}}\right) = 1\,, \tag{11}$$

which is the curve (10) rescaled to enclose unit area, is known as Vershik's limit shape of a random partition of a large number [49].

Partitions are among the most basic objects of combinatorics and additive number theory. In particular, the number $p(n)$ of partitions of n is a quantity of classical interest. Using the generating function

$$\sum_{n\geq 0} p(n)\, q^n = \prod_{n>0} (1-q^n)^{-1}$$

Hardy and Ramanujan proved in 1918 that

$$p(n) \sim \frac{1}{4n\sqrt{3}}\, e^{\pi\sqrt{2n/3}}\,, \tag{12}$$

a result subsequently refined in 1937 by Rademacher, same Hans Rademacher who's 1919 theorem on almost everywhere differentiability of Lipschitz functions we quoted earlier.

Let us now, following [46], examine the connection between this result and LDP for random walks. We have proven that as $n \to \infty$ the diagrams of almost all partitions of n, rescaled by $\sqrt{n}$ in both directions, fit in an arbitrarily small neighborhood the maximizer $f^\star$ given by (11). Since their step size δ is given by

$$\delta = n^{-1/2}$$

we conclude from (4) and (6) that

$$\frac{\ln p(n)}{\sqrt{n}} \to \int S\left(\frac{1 + f^\star(t)'}{2}\right)\, dt = \pi\sqrt{\frac{2}{3}}\,, \tag{13}$$

in agreement with (12).

3.7

A simple, but important lesson that we learn from this is a qualitative explanation of the $\sqrt{n}$ asymptotics of $\ln p(n)$. Indeed, since the diagram of a typical partition of n is a random walk of length $O(\sqrt{n})$ it has $O(\sqrt{n})$ degrees of freedom, i.e. choices whether to go up or down. These choices are to a large degree independent and so they make $\ln p(n)$ scale as $\sqrt{n}$.

More generally, one expects the logarithm $\ln Z$ of the partition function like (7) to scale like the number of effective degrees of freedom as the size of system grows. For example, in the next Section we will deal with random surfaces instead of random curves. For them the prefactor in the analog of (4) changes to δ^{-2}, where δ is the mesh size.

4 3D Ising interfaces at zero temperature

4.1

In three dimensions, the contours separating white from blue become surfaces. In particular, Figure 10 shows a 3-dimensional analog of the corner from Figure 8. At zero temperature, only those interfaces survive that minimize their area for given boundary conditions. These may be described by the conditions that they project 1-to-1 in the $(1, 1, 1)$ direction. In other words, there are no overhangs. Such surfaces are called *stepped surfaces*. They may be viewed as an analog of the simple random walk with 1-dimensional space and 2-dimensional time.

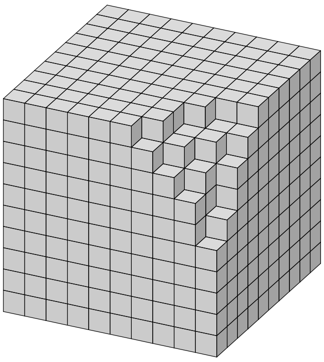

Figure 10: Same as Figure 8 but now in 3 dimensions.

It is clear from Figure 10 that there are many energy minimizers for a given number N of missing atoms. In fact, they are in a natural bijection with 3-dimensional or *plane partitions* of the number N. We may ask the same question we answered in Section 3.5, but now in the 3-dimensional setting. Namely, what is the limit shape of the zero-temperature corner as the number of missing atoms goes to infinity ?

4.2

What we need is a LDP for stepped surfaces. Let (x_1, x_2, x_3) be coordinates on $\mathbb{R}^3$ in which the crystal corner is the positive octant. We may parametrize a stepped surface by

$$x_3 = h(x,y)\,, \quad (x,y) = (x_3 - x_1, x_3 - x_2)\,.$$

The function h is called the *height function.* It is Lipschitz with constant 1. In fact, for any stepped surface the gradient of h takes only 3 values, namely the vertices of the triangle

$$\triangle = \operatorname{Conv}\{(0,0),(0,1),(1,0)\}\,.$$

Any pointwise limit f of rescaled height functions is Lipschitz. Its gradient, which is defined almost everywhere by Rademacher's theorem, takes values in $\triangle$. The triangle $\triangle$ replaces the segment $[0,1]$ which was the domain of the Shannon entropy function.

Cohn, Kenyon, and Propp proved in [8] an LDP for stepped surfaces with the action functional of the form

$$\mathcal{S}_{3\mathrm{D}}(f) = -\iint \sigma_{3\mathrm{D}}\left(\nabla f\right)\, dxdy\,, \tag{14}$$

for a certain function $\sigma_{3\mathrm{D}}$ on the triangle $\triangle$. Before we get to the description of $\sigma_{3\mathrm{D}}$, we remark that action functionals of the form (14) are typical for random surfaces with *local interaction*, that is, in the situation when both configurations and their probabilities are defined by a set of local rules.

The basis heuristic behind (14) is the one from Figure 3. To give a rough estimate of the number of stepped surfaces in the neighborhood of f, we may subdivide the domain of f into regions on which f is approximately linear. From locality, we expect that the number of nearby stepped surfaces will factor over these regions to leading order. For random walks, such factorization was exact. In general, proving it requires work. When that work it completed, it remains to tackle the case of a linear function f.

This is where the function σ_{3D} comes in. It precisely measures how much stepped surfaces like or dislike having a particular average slope. For this reason, it is called the *surface tension.* In principle, this function may be defined and studied without trying to identify it in terms of previously known special functions. For example, nobody knows an analytic formula for the surface tension of 3D Ising interfaces at positive temperature. But, remarkably, at zero temperature a beautiful description exists, thanks to an exact relation to the Kasteleyn theory of *planar dimers* [19].

4.3

In one sentence, σ_{3D} is the Legendre dual of the Ronkin function of the straight line. I will now explain what these words mean.

Let $P(z,w)$ be a polynomial in two variables. Its zero set $P(z,w)=0$ is, by definition, a plane algebraic curve. For example,

$$P(z,w) = z + w - 1 \tag{15}$$

defines a straight line. The Ronkin function of P is defined by

$$R(x,y) = \frac{1}{(2\pi i)^2} \iint_{\substack{|z|=e^x \\ |w|=e^y}} \log \bigl|P(z,w)\bigr| \, \frac{dz}{z} \, \frac{dw}{w} \, . \tag{16}$$

In other words, it is the average value of $\log|P|$ on the torus

$$\{|z| = e^x, |w| = e^y\} \subset \mathbb{C}^2 \, .$$

It is not difficult to prove, see [42], that for any polynomial P

(1) the Ronkin function R is convex,

(2) its gradient ∇R takes values in the Newton polygon of P,

(3) R is piecewise linear wherever the integral (16) is nonsingular.

Recall that, by definition, the Newton polygon of P is the convex hull of all exponents appearing in the equation of P, that is, all points $(a,b) \in \mathbb{Z}^2$

such that the monomial $z^a w^b$ is present in P. For the straight line (15) the Newton polygon is evidently the triangle $\triangle$.

Property (3) above says that R is piecewise linear away from the image of the curve $P(z,w)$ under the map

$$(z,w) \to (\ln|z|, \ln|w|)\,.$$

This image is called the *amoeba* of P, see Figure 11. The curve $P(z,w)=0$ is a Riemann surface of some genus and the holes in it may result in holes (nuclei) of the amoeba. The points where $P(z,w)=0$ intersects coordinate axes or the line at infinity produce the tentacles of the amoeba. The number of holes and tentacles is thus bounded by the genus and degree of the curve $P=0$, respectively. The shape of the amoeba imprints on the shape of the Ronkin function, see Figure 11.

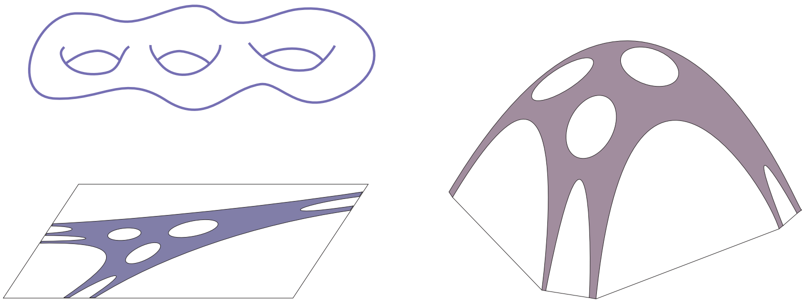

Figure 11: The amoeba and minus Ronkin function may look like this

4.4

What is the amoeba of the straight line ? The possible values of $|z|$ and $|w|$ for z,w satisfying $z+w=1$ are described by the triangle inequality. It implies that the amoeba is defined by

$$e^x + e^y \leq 1\,, \quad e^x \geq e^y + 1\,, \quad e^y \geq e^x + 1\,.$$

One can see it in Figure 12. It has no holes because a straight line has genus 0 and one tentacle in each direction because a straight line has degree 1.

We can now visualize the Ronkin function of the straight line, see Figure 13, left half. We will get to the right half of Figure 13 in a second, but first we need to take the Legendre transform of the Ronkin function which will give us the surface tension $\sigma_{3\mathrm{D}}$. It is plotted in Figure 14.

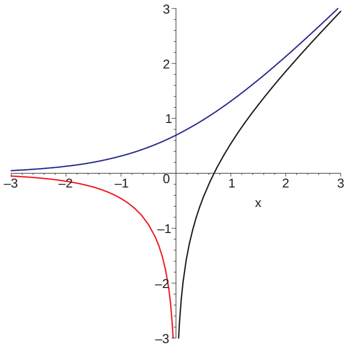

Figure 12: The amoeba of a straight line

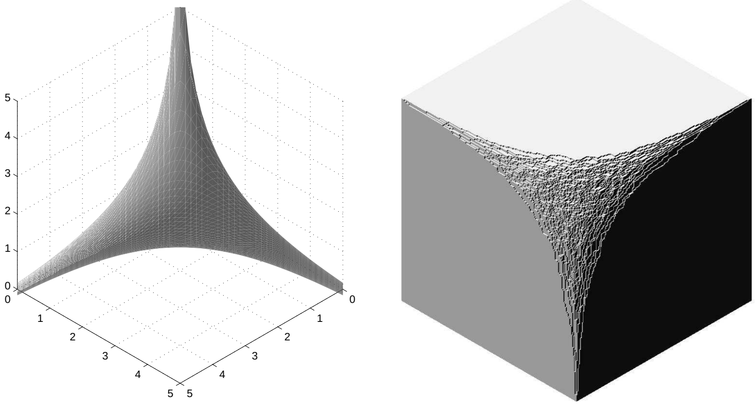

Figure 13: The Ronkin function of $z + w = 1$ as itself and as the crystal corner limit shape (simulation)

4.5

Now we have a variational characterization of the crystal corner limit shape. It is the maximizer of (14) among Lipschitz functions f tending to

$$f_0(x, y) = \max(0, x, y)$$

at infinity and enclosing a given volume between the graphs of f and f_0. In principle, this sounds like a highly nonlinear singular variational problem. Its solution, however, is readily available — it is the Ronkin function again. In fact, for a functional like (14), the Legendre transform of σ_{3D} is always a volume constrained maximizer (for its own boundary conditions). This is known as Wulff theorem and can be proven on very general grounds.

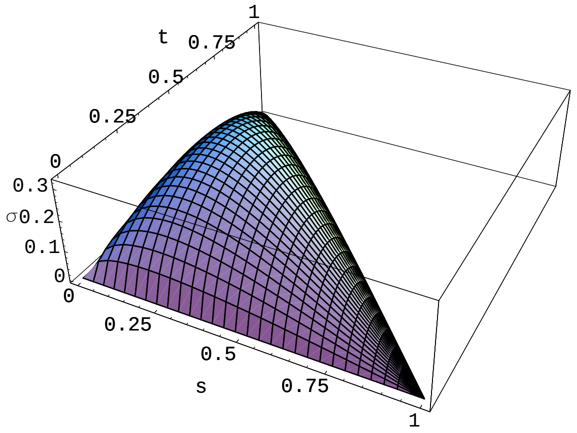

Figure 14: The graph of $-\sigma_{3D}$.

Same theorem connects the shape (10) to Shannon entropy (2). Incidentally, Legendre transform relates the surface tension of the Wulff shape, which we computed in (13) in a special case, to the volume enclosed by it.

Thus, we have identified, following Cerf and Kenyon [7], the limit shape of the 3D Ising corner at zero temperature. We may compare this theoretical result with the result of a computer simulation presented in Figure 13.

4.6

A closer examination of the simulation in Figure 13 reveals the following somewhat surprising feature. The parts of limit shape corresponding to flat pieces of the Ronkin function remain flat down to the microscopic scale ! Facets of natural crystals share this property: their height varies by only a few atomic layers. It is known that the flat pieces (facets) persist in the 3D Ising crystal for small positive temperatures and it was suggested in [2] that their height varies by at most 1 atomic layer when the temperature is sufficiently small. For an update on this, see [17, 18].

This formation of facets is new phenomenon that we did not observe in 2 dimensions. The boundary of (8) is a strictly convex analytic curve for all temperatures between 0 and T_c.

4.7

Note that the facet boundaries in Figure 13, which are also the amoeba boundaries from Figure 12, coincide with the 2-dimensional limit shape from Figure 9. A related observation is that near the origin, which is a singularity of the surface tension $\sigma_{3\mathrm{D}}$, it has the form

$$\sigma_{3\mathrm{D}}(s,t) = -(t+s)\, S\left(\frac{t}{t+s}\right) + o(|t+s|)\,,$$

where S is the Shannon entropy.

See [25, 26] for a discussion of what happens to this singularity of the surface tension at positive temperature.

4.8

In two dimensions, Figure 9 contained, inside a suitable window, the solution to (9) with all possible boundary conditions. By contrast, in three dimensions one may subject stepped surfaces to a much greater variety of boundary conditions. One observes that boundary conditions have a strong influence on the behavior of stepped surfaces.

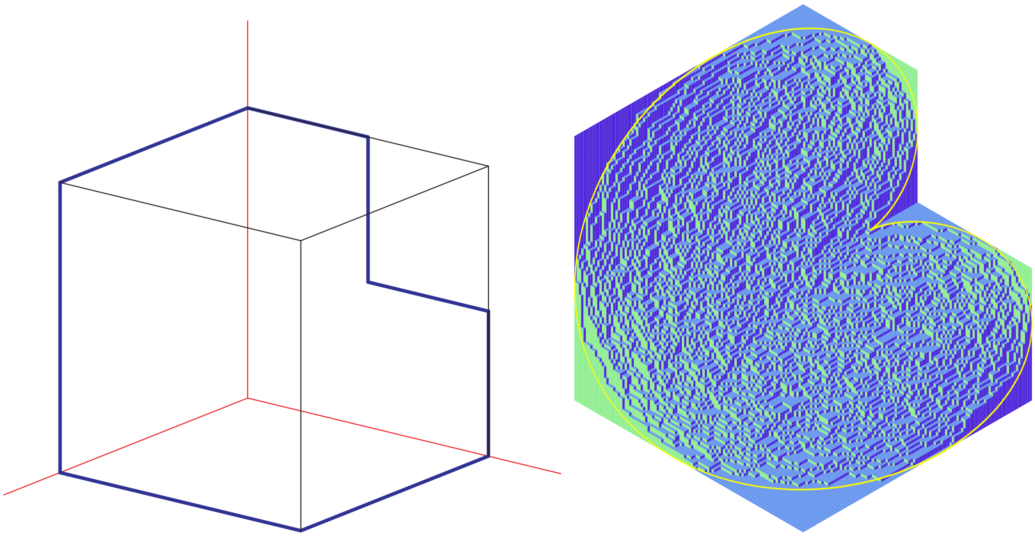

Figure 15: A polygonal boundary contour and a random stepped surface spanning it

Take for example the boundary contour shown in blue in Figure 15, left half. Imagine its edges belong to a lattice $\delta\mathbb{Z}^3$ with a very small mesh δ and let's look at a random stepped surface spanning it. The result of a computer simulation may be seen on the right in Figure 15.

Again, we observe facet formation along the boundaries. These facets, as before, are completely ordered. There is a sharp boundary between them and the disordered region, known as the *frozen boundary.* We will find the $\delta \to 0$ limit shape below and, in particular, we will see that the frozen boundary is an inscribed *cardioid.* This cardioid is plotted in yellow in Figure 15.

A similar phenomenon was observed by Cohn, Larsen, and Propp in [9]. In their case, the frozen boundary is a circle, which they named the *arctic circle.* In fact, as we will see, the frozen boundary is an algebraic curve for any polygonal boundary contour.

4.9

The LDP for stepped surfaces implies the limit shape for stepped surfaces spanning given boundary and enclosing a given volume is the unique minimizer of

$$\iint_\Omega \left[\sigma_{\mathrm{3D}}(\nabla f) + cf\right]\, dxdy \to \min \tag{17}$$

among all Lipschitz functions f taking given values on $\partial\Omega$, where Ω is the region enclosed by the projection of the boundary contour along the $(1,1,1)$ direction and c is the Lagrange multiplier.

This is a nonlinear and singular variational problem. In fact, the singularities of the surface tension σ_{3D} are precisely responsible for the formation of facets.

The following transformation of the Euler-Lagrange equation for (17) found in [21] is of a great help in the study of the facet formation. Namely, in the disordered region, the minimizer $f^\star$ satisfies

$$\nabla f^\star = \frac{1}{\pi}(\arg w, -\arg z)\,, \tag{18}$$

where the functions z and w solve the differential equation

$$\frac{z_x}{z} + \frac{w_y}{w} = c \tag{19}$$

and the algebraic equation (15). At the frozen boundary, z and w become real and the $\nabla f^\star$ starts to point in one of the coordinate directions.

4.10

The first-order quasilinear equation (19) is, essentially, the complex Burgers equation $z_x = zz_y$ and, in particular, it can be solved by complex characteristics as follows. There exists an analytic function $Q(z, w)$ such that

$$Q(e^{-cx}z, e^{-cy}w) = 0\,. \tag{20}$$

In other words, $z(x, y)$ can be found by solving (15) and (20). In spirit, this is very close to Weierstraß parametrization of minimal surfaces in terms of analytic data.

Frozen boundary only develops if Q is real. In this case, the solutions (z, w) and $(\bar{z}, \bar{w})$ coincide at the frozen boundary, so it is a *shock* for (19).

4.11

Let the boundary contour be connected and *polygonal*, that is, formed by segments going in the coordinate directions. If we allow segments of zero length, we may assume that the three possible directions repeat cyclically around the contour d times. For example, in Figure 15, we have one segment of zero length and $d = 3$.

With these hypotheses, it is proved in [21] that the analytic function Q above is a real polynomial of degree at most d. In addition, the equation $Q = 0$ defines a plane curve of genus 0, that is, a curve admitting a rational parametrization. The boundary segments determine where this curve intersects the coordinate axes and the line at infinity; they also determine the order in which these intersections occur. From these data one reconstructs the polynomial Q uniquely.

4.12

The frozen boundary is given by

$$Q^\vee(e^{cx}, e^{cy}) = 0$$

where $Q^\vee$ defines the *dual curve* of Q, i.e. the set of all (a, b) such that the line

$$ax + by = 1$$

is tangent to $Q(x, y) = 0$.

Duality is a beautiful classical operation on plane curves. The cardioid and its dual, which is a nodal cubic curve, may be seen in Figure 16. In particular, Q and $Q^\vee$ have the same genus, therefore the frozen boundary is a rational curve in exponentiated coordinates.

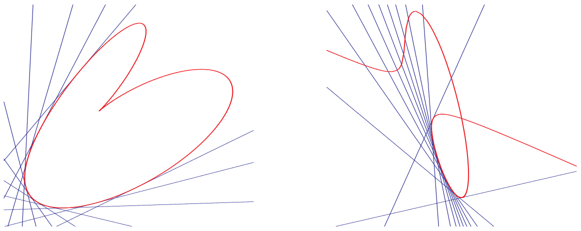

Figure 16: A cardioid and its dual

4.13

Formula (18) may be given the following geometric interpretation. For concreteness, let us consider the cardioid example when $d = \deg Q = 3$. The same number $d = 3$ is also the number of tangents to the cardioid $Q^\vee$ through a general point in the plane, see Figure 17. If the point is outside the cardioid, then all 3 tangents are real, see Figure 17. If, however, the point is inside the cardioid, as the point (e^{cx}, e^{cy}) would be for (x, y) in the disordered region, then only one real tangent to the cardioid will meet it. What happened to the other two ? They became a pair of complex conjugate tangents. I wish I knew how to draw complex tangents in the real plane. Without them, Figure 17 is missing its most important feature.

Whatever that complex tangent is, it is given by an equation of the form $a_1X + a_2Y + a_3 = 0$ and since the point $(X, Y) = (e^{cx}, e^{cy})$ lies on it, we know that

$$a_1e^{cx} + a_2e^{cy} + a_3 = 0$$

Three complex numbers summing to 0 define a triangle in the complex plane. The numbers (a_1, a_2, a_3) are defined only up to a common complex multiple and complex conjugation, hence our triangle is defined only up to similarity. But then its angles $\alpha_1, \alpha_2, \alpha_3$ are well-defined and (18) says that

$$(\alpha_1, \alpha_2, \alpha_3)$$

is the normal to the limit shape at the point above $(x, y) \in \Omega$.

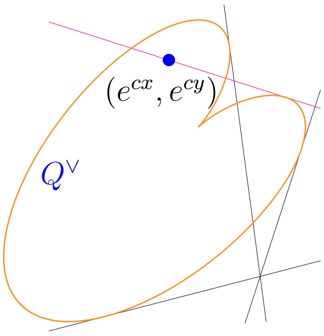

Figure 17: Where did the missing tangents go ?

As (x, y) approaches the frozen boundary, the complex tangent becomes real, the triangle degenerates, and the normal starts to point in one of the coordinate directions.

4.14

There was a reason for our discussion of amoebas of general plane algebraic curves $P(z, w) = 0$. Legendre transforms of their Ronkin function express the surface tension of *periodically weighted* stepped surfaces.

Crystals are periodic, but not necessarily with period one. For example, the sodium and the chloride ions in the table salt are arranged in the vertices of the cubic lattice in a checkerboard fashion, i.e. according to the parity of $x_1 + x_2 + x_3$, where $(x_1, x_2, x_3) \in \mathbb{Z}^3$ are the three coordinates.

We can introduce this feature for stepped surfaces by weighting each square tile by a periodic function of its position. Since the projection in the $(1, 1, 1)$ direction played a special role, we will require the weights to be invariant under the $(1, 1, 1)$ translation, which is not the case for table salt.

We say that our weights are periodic with period M if they are invariant under translation by any element of $M\mathbb{Z}^3$. In this case, the curve $P(z, w) = 0$, known as the *spectral curve*, has degree M. Its Ronkin function, i.e. the Ising corner limit shape may have as many as

$$\text{genus of a smooth degree } M \text{ curve} = (M - 1)(M - 2)/2$$

compact facets of the kind sketched in Figure 11.

The beautiful paper [37] is probably the first place where a degree three polynomial P and the corresponding bounded facet appeared in the physics literature. See [22] for a mathematical treatment of the general case. Among other things, it is proven in [22] that all of the $(M-1)(M-2)/2$ facets will, in fact, be present away from a certain explicit codimension 2 subvariety in the space of weights. On that resonant subvariety, one or more facets shrink to zero size.

This is a consequence of the following *maximality* of the spectral curves: for nonnegative real weights, the map from the spectral curve to its amoeba is at most 2-to-1. In other words, all spectral curves are Harnack, see [24] for many properties and characterizations of such curves.

5 Instanton counting

5.1

The goal of this section is to explain how limit shapes ideas can be of help in a very different area of mathematical physics, namely, quantum gauge theory. We begin by briefly recalling its most basic notions.

Familiar physical quantities, such as e.g. temperature, are, up to a certain continuous idealization, functions on the space-time $\mathbf{M}$. The air temperature reported in a given city on a given day maps space-time to a fixed real line, which is the same for all seasons and all locations. The direction and the velocity of the wind is an example of a vector-valued field. The weather report for your city will probably record it as taking value in a fixed 2-dimensional vector space with coordinate axes in the cardinal directions of the compass. This is because you probably don't live on the North or South Pole, where such description breaks down. More mathematically, surface wind at any point of the earth is a tangent vector to the Earth's surface, so globally it is a *section* of the Earth's *tangent bundle*. This tangent bundle is nontrivial, i.e. not isomorphic to $S^2 \times \mathbb{R}^2$, for topological reasons. Whence the necessity to deal even in classical physics with fields taking values in a point-dependent vector space, i.e. sections of various *vector bundles* over the space-time $\mathbf{M}$.

In gauge theories, the principle that a matter field $\psi(\mathbf{x})$ is not a function but rather a section of an appropriate vector bundle V over $\mathbf{M} \ni \mathbf{x}$ is promoted to the requirement of *gauge invariance*. It postulates that the physics should not be affected by transformations of the form

$$\psi(\mathbf{x}) \mapsto u(\mathbf{x})\, \psi(\mathbf{x})\,,$$

where $u(\mathbf{x})$ is an arbitrary point-dependent unitary transformation of the fibers of V (which typically come with an Hermitian inner product). In the most basic example of the trivial bundle $V = \mathbf{M} \times \mathbb{C}^r$, such transformation form the *gauge group* $\mathcal{G}$ of maps

$$u : \mathbf{M} \to U(r)$$

with point-wise multiplication.

5.2

Connections are responsible for holding the fibers of V together and transmitting interactions between matter fields.

A connection tells us how to differentiate a section of a vector bundle. In local coordinates

$$\mathbf{x} = (x_0, x_1, x_2, x_3)$$

these are $r \times r$ matrix-valued functions $A_i(\mathbf{x})$ that define covariant derivatives

$$\nabla_i = \frac{\partial}{\partial x_i} + A_i(\mathbf{x})\,, \quad A_i^* = -A_i\,. \tag{21}$$

From now on, we consider only the most basic case of the trivial bundle $V = \mathbb{R}^4 \times \mathbb{C}^r$ over the flat Euclidean space-time $\mathbf{M} = \mathbb{R}^4$, where such coordinate description is global. The gauge group $\mathcal{G}$ acts on connections by

$$\nabla \mapsto u(\mathbf{x})\, \nabla\, u(\mathbf{x})^{-1}\,.$$

5.3

By solving the ODE

$$\nabla \psi = 0$$

along a curve in $\mathbf{M}$, we can define parallel transport from the fiber of V over $\mathbf{x}$ to the fiber over $\mathbf{x}'$ along any path joining $\mathbf{x}$ to $\mathbf{x}'$. The anti-Hermitian condition in (21) implies unitarity of the parallel transport.

In general, the result of parallel transport depends on the the choice of the path from $\mathbf{x}$ to $\mathbf{x}'$, as Figure 18 schematically illustrates. In other words, the parallel transport along a closed curve may not be the identity transformation.

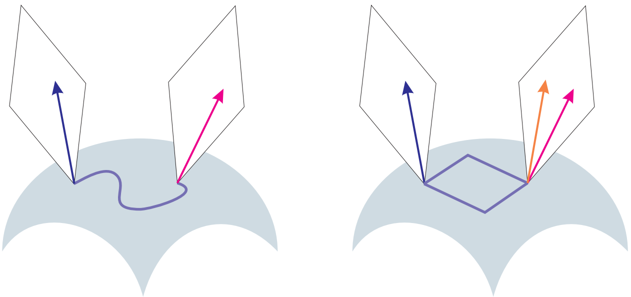

Figure 18: Parallel transport and curvature

The deviation of the parallel transport along an infinitesimal closed curve from the identity is measured by the *curvature*

$$F = \sum \left[\nabla_i, \nabla_j\right] \, dx_i \wedge dx_j \,,$$

which is a 2-form on $\mathbf{M}$ with values in $r \times r$ matrices.

5.4

The L^2-norm squared $\|F\|^2$ of the curvature is a natural gauge invariant energy functional for a connection A, first proposed by Yang and Mills. Poetically speaking, connections A glue together the fibers of V and, as a result, cause a certain stress $\|F\|^2$ in the fabric of space-time.

For example, in the $U(1)$-gauge theory, commonly known as the electromagnetism, the six entries of the 2-form F combine the electric and magnetic fields. In this case, $\|F\|^2$ becomes the familiar expression for the electromagnetic energy.

In the $U(r)$ case with $r > 1$, the energy $\|F\|^2$ has degree 4 in the entries of $A_i(\mathbf{x})$, reflecting the nonlinear, self-interacting nature of nonabelian gauge fields. This makes gauge theory a challenging problem already in the absence of any matter fields.

5.5

In parallel with (7), one "defines" the Euclidean partition function of the pure gauge theory to be

$$Z(\beta) = \int_{\text{connections}/\mathcal{G}} \exp\left(-\beta \, \|F\|^2\right) \, dA \tag{22}$$

The word Euclidean refers to the fact that our gauge theory is defined not on the Minkowski space-time with its indefinite metric, but rather on $\mathbb{R}^4$ with the standard positive definite metric $\|\mathbf{x}\|^2$. Pure means that, for the moment, we consider a gauge theory without any matter fields at all.

I've put quotation marks around "defines" to emphasize that the equality (22) is so far only symbolic. Indeed, the RHS is an integral over a problematic infinite-dimensional set that has not been given a natural measure dA.

A direct probabilistic approach, first proposed by K. Wilson, to actually defining (22) is to make it a theory of many interacting random matrices through a discretization of space-time. This is a fascinating topic about which I have nothing to say.

Instead, we will make our life simple by letting $\beta \to \infty$ while simultaneously imposing a certain topological constraint on the connections to keep the limit nontrivial.

5.6

The set-up is exactly parallel to our treatment of the Ising crystal. Back then, we fixed the overall number of the white squares while letting the temperature drop to zero. In the ferromagnetic interpretation of the Ising model, we fixed the overall magnetization of the sample before freezing it. We now plan to similarly constrain the connection A before letting $\beta \to \infty$.

The space of finite energy connections, that is, connections with $\|F\|^2 < \infty$, is in fact disconnected [48]. Its connected components are classified by a topological invariant

$$c = \frac{1}{8\pi^2} \int_{\mathbb{R}^4} \operatorname{tr} F^2 \,,$$

known as the 2nd Chern class or *instanton charge*. It provides a lower bound for the energy

$$8\pi^2 c \leq \|F\|^2 \,.$$

with an equality if and only if A is an *instanton*. Very schematically, the geometry of the Yang-Mills energy functional may be imagined like in Figure 19.

5.7

Instantons are immensely beautiful and much studied geometric objects. Approximately, an instanton of charge c may be imagined as a nonlinear super-

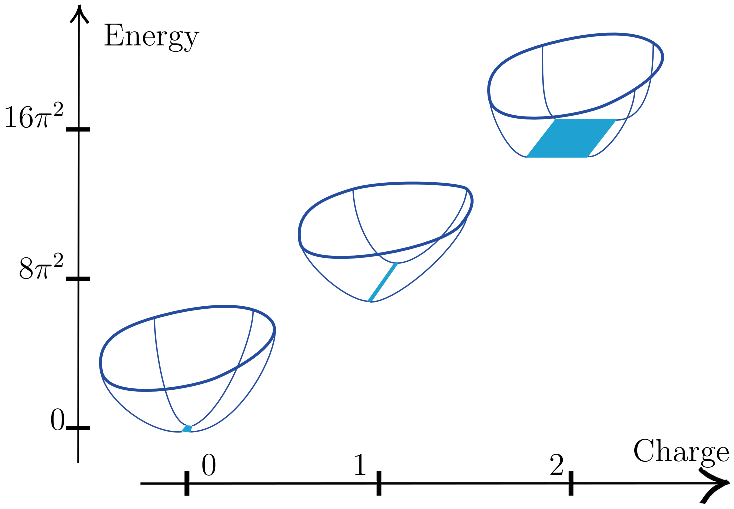

Figure 19: For given charge $c = 0, 1, 2, \ldots,$ absolute minima of $\|F\|^2$ are given by instantons, which modulo framed gauge equivalence form a manifold of dimension $4c \times$ rank.

position of c instantons of charge 1. In this sense, instantons are like a gas of little bumps, or twists, in the fabric of space-time. These bumps are localized in space-time, i.e. they may only be spotted for an instant, whence the name.

For a ridiculously oversimplified analogy, take the rubber bands in Figure 20. They minimize their energy for given topology. They can occur anywhere in this 1-dimensional space-time and feel each other's presence only when they are very close. The analog of the instanton charge in this example is the total winding number around the pole.

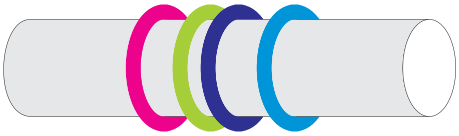

Figure 20: Instantons are a bit like a gas of rubber bands

Finally, just as in the Ising case crystal case the final step was to let the number of missing atoms go to infinity, we will eventually be interested in the effects produced by a large number of instantons.

5.8

For fixed charge, instantons dominate the $\beta \to 0$ limit in the partition function. Modulo the action of the gauge group $\mathcal{G}$, instantons of a given charge are parametrized by points of a certain finite-dimensional algebraic variety.

For our purposes, it is convenient and important to consider *framed instantons*. By a theorem of Uhlenbeck, any instanton on $\mathbb{R}^4$ extends, after a gauge transformation, to an instanton on

$$S^4 = \mathbb{R}^4 \cup \{\infty\} .$$

Thus we can talk about the value of an instanton at infinity.

Let $\mathcal{G}_0$ be the group of maps $g : S^4 \to U(r)$ such that $g(\infty) = 1$. Modulo $\mathcal{G}_0$, instantons on S^4 of given charge c are parametrized by a smooth manifold $\mathcal{M}(r,c)$ of real dimension $4rc$, known as the *moduli space* of framed instantons. The word framed refers to not taking the quotient by the action of constant gauge transformation.

Since all instantons of given charge have the same energy, it is natural to think that all of them will contribute a multiple of $\operatorname{vol} \mathcal{M}(r,c)$ to the partition function. However, the space $\mathcal{M}(r,c)$ is noncompact and its volume requires a regularization. This should be clear from the description of instantons as a gas of bumps. Those bumps can happen anywhere in space-time.

While the Ising corner geometry forced the vacant lattice sites to occur near the vertex of the crystal, there is no analogous intrinsic mechanism to prevent instantons from wondering off to infinity.

5.9

The traditional way to deal with a gas in statistical mechanics is to put it in an impervious reservoir, followed by letting the size of the reservoir go to infinity. This is precisely what we did to white squares in Figure 4. Imposing a similar cut-off for instantons is most unnatural and leads to terrible complications.

Nekrasov's idea was to use *equivariant integration* in lieu of a space-time cut-off. For the simplest of all examples, suppose we want to regularize the volume of $\mathbb{R}^2$. A gentle way to do it is to introduce a Gaussian well

$$\int_{\mathbb{R}^2} e^{-t\pi(x^2+y^2)} dx\, dy = \frac{1}{t} , \quad \Re t \geq 0 \tag{23}$$

and thus an effective cut-off at the $|t|^{-1/2}$ scale.

We now interpret this very familiar integral in the following way. We note that with respect to the standard symplectic form $\omega = dx \wedge dy$ the function

$$H(x,y) = \tfrac{1}{2}(x^2+y^2) = \tfrac{1}{2}|z|^2\,, \quad z = x+iy\,,$$

is the Hamiltonian generating the action

$$t \mapsto e^{it}$$

of $\mathbb{R}/2\pi\mathbb{Z}$ by rotations of $\mathbb{C} = \mathbb{R}^2$.

The origin $z = 0$ is the unique fixed point of this action. These observations make (23) the simplest instance of the Atiyah-Bott-Duistermaat-Heckman-Berline-Vergne *equivariant localization* formula [1]. We will state it in the following complex form.

Let $T = \mathbb{C}^*$ act on a complex manifold X with isolated fixed points X^T. Suppose that the action of $U(1) \subset T$ is generated by a Hamiltonian H with respect to a symplectic form ω. Then

$$\int_X e^{\omega - 2\pi t H} = \sum_{x\in X^T} \frac{e^{-2\pi t H(x)}}{\det t|_{T_x X}}\,, \tag{24}$$

where t should be viewed as an element of $\mathrm{Lie}(T) \cong \mathbb{C}$, an so it acts in the complex tangent space $T_x X$ to a fixed point $x \in X$.

While (24) is normally stated for compact manifolds X, example (23) shows that with care it can work for noncompact ones, too. Scaling both ω and H to zero, we get from (24) a formal expression

$$\int_X 1 \overset{\mathrm{def}}{=} \sum_{x\in X^T} \frac{1}{\det t|_{T_x X}}\,, \tag{25}$$

which does not depends on the symplectic form and vanishes if X is compact. This will be our definition of the equivariantly regularized volume.

5.10

The group

$$K = SU(2) \times SU(r)$$

acts on $\mathcal{M}(r,c)$ by rotations of $\mathbb{R}^4 = \mathbb{C}^2$ and constant gauge transformation, respectively. It also acts on a certain partial compactification

$$\overline{\mathcal{M}}(r,c) \supset \mathcal{M}(r,c)$$

known as the moduli space of *torsion-free sheaves*, see [16, 27]. The space $\overline{\mathcal{M}}(r,c)$ compactifies those directions that correspond to point-like instantons, while leaving the directions corresponding to run-away instantons uncompactified. In contrast to $\mathcal{M}(r,c)$, this larger moduli space has fixed points, so it is meaningful to apply formula (25) to it.

Equivariant localization with respect to a general $t \in \mathrm{Lie}(K)$

$$t = (\mathrm{diag}(-i\varepsilon, i\varepsilon), \mathrm{diag}(ia_1, \ldots, ia_r)) \tag{26}$$

combines the two following effects. First, it introduces a spatial cut-off parameter ε as in (23). Second, it introduces dependence on the instanton's behavior at infinity through the parameters a_i.

Our $\beta \to \infty$ approximation remains exact in certain gauge theories with enough *supersymmetry.* In that context, the parameters a_i correspond to the vacuum expectation of the Higgs field and thus are responsible for masses of gauge bosons. In short, the a_i's are live physical parameters.

5.11

We are now ready to introduces the *Nekrasov partition function* for the pure $U(r)$ theory. Instead of fixing the instanton charge c, it is more convenient to work with a generating function that combines the contributions of $c = 0, 1, 2, \ldots$. In the language of the classical statistical mechanics, it is more convenient to work with the *grand canonical ensemble* for the instanton gas, i.e. the ensemble in which the number of particles is not fixed by law but instead regulated by the energy cost of adding a extra particle.

We define

$$Z(\varepsilon; a_1, \ldots, a_r; \Lambda) = Z_{\mathrm{pert}} \sum_{c \geq 0} \Lambda^{2cn} \int_{\overline{\mathcal{M}}(r,c)} 1\,, \tag{27}$$

where Z_{pert} is a certain explicit product known as the perturbative contribution, Λ is the parameter of the generating function and the integral is defined by (25) applied to (26).

Since the energy of an instanton is linear in c, one can view (27) as the instanton contribution to (22) via the identification by

$$\Lambda = \exp(-4\pi^2\beta/r)\,.$$

5.12

By our regularization rule

$$\operatorname{vol}\mathbb{R}^4 = \int_{\mathbb{R}^4} 1 = \frac{1}{\varepsilon^2}\,.$$

Indeed, this is just the square of (23). Hence, from the argument of Section 3.7 we may expect that as $\varepsilon \to 0$

$$\ln Z(\varepsilon; a; \Lambda) \sim -\frac{1}{\varepsilon^2}\,\mathcal{F}(a;\Lambda)\,,$$

for a certain function $\mathcal{F}$, the traditional name for which which the *free energy*, or bulk free energy, that is, free energy per unit volume.

Nekrasov conjectured in [31] that the function $\mathcal{F}$ is the same as the *Seiberg-Witten prepotential*, first proposed in [43, 44] based on entirely different considerations. It is defined in terms of a certain family of algebraic curves.

5.13

Consider the following $(r-1)$-dimensional family of algebraic curves C in the (z,w)-plane

$$\Lambda^r \left(w + w^{-1}\right) = P(z)\,, \quad P(z) = z^r + u_2\, z^{r-2} + \cdots + u_r\,. \tag{28}$$

The parameter Λ here is fixed; it is the same as above. The coefficients

$$(u_2, \dots, u_r) \in \mathbb{C}^{r-1}\,,$$

of the polynomial P are the parameters of the family. See e.g. [45] and [47] for other situations in which this family naturally occurs.

The map $(z,w) \mapsto z$ is a 2-fold ramified over the roots of the equation

$$P(z) = \pm 2\Lambda^r\,.$$

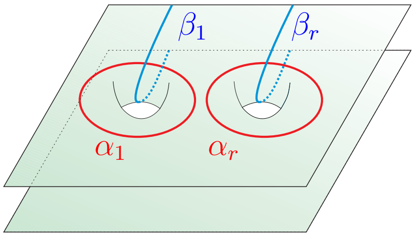

Figure 21: Cycles on Seiberg-Witten curve

Away from a hypersurface in the u-space, all $2r$ roots of this equation are distinct and hence (28) defines a smooth curve C of genus $r-1$. One then can choose cycles α_i and β_i on C as illustrated in Figure 21.

This cycles are not independent. Indeed, clearly $\sum \alpha_i = 0$. The cycles

$$\alpha_i - \alpha_{i+1}\,, \beta_i - \beta_{i+1}\,, \quad i = 1, \dots, r-1\,,$$

form a basis of homology. The periods of the differential

$$dS = \frac{1}{2\pi i}\, z\, \frac{dw}{w}$$

along either half of these cycles may be chosen as local coordinates on the base of family instead of the coefficients u.

The Seiberg-Witten prepotential $\mathcal{F}(a;\Lambda)$ is defined implicitly by the equations

$$a_i = \int_{\alpha_i} dS\,, \quad \frac{\partial}{\partial a_i}\mathcal{F} = -2\pi i \int_{\beta_i} dS\,.$$

From this description one can work out an asymptotic expansion of this function as

$$a_1 \ll a_2 \ll \cdots \ll a_r\,,$$

as well as its singularities and monodromy in the complex domain. In fact, the most valuable physical information is contained in the singularities of the free energy.

The surprising fact that free energy $\mathcal{F}$ is given in terms of periods of a hidden algebraic curve C is an example of *mirror symmetry*.

5.14

Nekrasov's conjecture was proven in [33] for a list of gauge theories with gauge group $U(r)$, namely, pure gauge theory, theories with matter fields in fundamental and adjoint representations of the gauge group, as well as 5-dimensional theory compactified on a circle. Simultaneously, independently, and using completely different ideas, the formal power series version of Nekrasov's conjecture was proven for the pure $U(r)$-theory by Nakajima and Yoshioka [28]. The methods of [33] were applied to classical gauge groups in [36] and to the 6-dimensional gauge theory compactified on a torus in [15]. Another algebraic approach, which works for pure gauge theory with any gauge group, was developed by Braverman [4] and Braverman and Etingof [5].

5.15

The main steps of the proof given in [33] are as follows. By construction (25), the partition function (27) is a sum over fixed-points of the torus action on $\overline{\mathcal{M}}(r, c)$. One begins by interpreting it as the partition function of an ensemble of random partitions. The $\varepsilon \to 0$ limit turns out to be the thermodynamic limit in this ensemble.

Next one proves, using the results of [23, 50, 51], an LDP for these random partitions, thereby showing, just as we did Section 3.6 for the Hardy-Ramanujan formula, that

$$\mathcal{F} = \min \mathcal{S}_{\text{inst}} \,,$$

for a certain convex functional $\mathcal{S}_{\text{inst}}$ on Lipschitz functions.

We then solve the variational problem explicitly. The solution is the algebraic curve C in disguise. Namely, the limit shape is essentially the graph of the function

$$\Re \int_{x_0}^{x} dS \,,$$

where dS is the Seiberg-Witten differential. Thus all ingredients of the answer appear very naturally in the proof.

The limit shape turns out to be an algebraic curve for precisely the same reason as the limit shapes from Section 4. In fact, the random partition ensemble in question may be produced by taking a random surface of the kind considered in Section 4 and slicing it along a particular direction. This offers an intriguing connection between the physics of crystals and the physics of instantons. For example, singularities of the free energy $\mathcal{F}$ occur when a facet shrinks to zero size.

I refer to [33, 38, 40] for details.

6 Outlook

6.1

It should be clear by now that what we've seen so far is just the beginning of a vast and rich land, in which several branches of mathematics and physics intersect and interact. And now, from the farthest point of our excursion, we can see, perhaps only in outline, some distant peaks near the horizon. Here are a few of them, picked according to my personal taste and expertise.

6.2

The principle of large deviation is a very robust probabilistic philosophy, applicable to large random systems of really diverse nature and origins. Throughout sciences, it is responsible for the transition from the microscopic description of a given systems to its large scale description in terms of calculus of variations and associated PDEs. Think of the transition from molecular physics to continuum mechanics. Anything so robust has to be sufficiently abstract and most of the time only rather abstract conclusions can be made about either the form of the action functional or about the solutions of the corresponding variational problem.

What we have seen in Section 4 is clearly the opposite of robust and is an evidence of something very special happening in the particular case of stepped surfaces. Obviously, there is something exactly solvable or *integrable* about the corresponding variational problem. Any time a problem in classical physics is integrable, it is a strong hint that *quantum integrability* could be nearby, which in the statistical context means that there could be a way to describe not only the law of large numbers but also the *fluctuations* around it in some closed analytic form.

Remarkably, many models of random surfaces or random partitions of the kind we discussed are quantum integrable, in the sense that one can work out a complete asymptotic expansion for the expectations of natural observables as the size of the system goes to infinity. For example, in the setting of Section 4 this may be achieved through a certain *quantization*, or a noncommutative deformation, of the curve Q, see [41].

The leading term in such expansions is determined by the limit shape, followed by a Gaussian correction, followed by terms that may, in fact, decay with the size of the system and so may not admit a clear probabilistic interpretation. The whole asymptotic series, however, is of great interest from the viewpoint of other applications, such as supersymmetric gauge theories and enumerative geometry.

6.3

The actual mathematical and physical mechanisms of quantum integrability are deep and diverse, and their proper discussion should be a topic of a separate lecture series. It also goes in the direction which is rather orthogonal to the probabilistic viewpoint which we took here. Quantum integrability is destroyed by the slightest variation in the model, including, for example, variations in boundary conditions. Nor is it crucial for it that the probabilities are nonnegative — another feature that should make it really suspect in the eyes of a true probabilist.

Probabilists should be nonetheless aware of the existence of such methods and of their power. Specifically in this context I would like to mention certain discrete versions of the Schwinger-Dyson equation developed by Nekrasov [32], see also earlier papers by Nekrasov and his collaborators [34, 35]. The paper [3] is a perfect demonstration of the force of such methods in probabilistic applications.

6.4

In Section 5 we discussed how some computations in 4-dimensional supersymmetric gauge theories may be usefully reformulated as the study of certain ensembles of random partitions. The plane partitions of Section 4, and certain more general random surfaces may similarly be associated to computations in certain 6- and 7-dimensional supersymmetric theories, known in mathematics as Donaldson-Thomas theories, or DT theories for short. These are enumerative theories of sheaves on smooth algebraic 3-folds which, from many perspectives, may be viewed as complex analogs of Chern-Simons theories of real 3-folds.

In particular, the formation of the limit shape for these random surfaces, and the algebraic description of the limit shape, may be seen as a probabilistic manifestation of *mirror symmetry*, see for example [39, 40]. Many other phenomena discussed before, such as the appearance of Seiberg-Witten curves from limit shapes of random surfaces, also find a natural explanation in the DT context.

It would be fair to say, however, that the worlds of random surfaces and of enumerative geometry intersect along a set of very small measure in each. Natural probabilistic deformation of the problem typically lose geometric significance and hidden structures, and vice versa. To me, this underscores the mathematical importance of the objects considered here as objects that can be looked at from many inequivalent points of view.

References

[1] A. Atyiah and R. Bott, *The moment map and equivariant cohomology*, Topology **23** (1984), no. 1, 1–28.

[2] T. Bodineau, R. Schonmann, S. Shlosman, *3D crystal: how flat its flat facets are?*, Comm. Math. Phys. **255** (2005), no. 3, 747–766.

[3] A. Borodin, V. Gorin, and A. Guionnet, *Gaussian asymptotics of discrete β-ensembles*,, `arXiv:1505.03760`.

[4] A. Braverman, *Instanton counting via affine Lie algebras. I. Equivariant J-functions of (affine) flag manifolds and Whittaker vectors*, Algebraic structures and moduli spaces, 113–132, CRM Proc. Lecture Notes, 38, Amer. Math. Soc., Providence, RI, 2004, `math.AG/0401409`.

[5] A. Braverman and P. Etingof, *Instanton counting via affine Lie algebras II: from Whittaker vectors to the Seiberg-Witten prepotential*, `math.AG/0409441`.

[6] R. Cerf, *The Wulff crystal in Ising and percolation models*, Lectures from the 34th Summer School on Probability Theory held in Saint-Flour, July 6–24, 2004, Lecture Notes in Mathematics, **1878**, Springer-Verlag, Berlin, 2006.

[7] R. Cerf and R. Kenyon, *The low-temperature expansion of the Wulff crystal in the 3D Ising model*, Comm. Math. Phys. **222** no. 1, 147–179 (2001).

[8] H. Cohn, R. Kenyon, J. Propp, *A variational principle for domino tilings*, Journal of AMS **14**(2001), no. 2, 297-346.

[9] H. Cohn, M. Larsen, J. Propp, *The shape of a typical boxed plane partition*, New York J. Math. **4** (1998), 137–165.

[10] E. D'Hoker and D. Phong, *Lectures on supersymmetric Yang-Mills theory and integrable systems*, Theoretical physics at the end of the twentieth century, 1–125, CRM Ser. Math. Phys., Springer, New York, 2002, `hep-th/9912271`.

[11] R. Dobrushin, R. Kotecký, S. Shlosman, *Wulff construction. A global shape from local interaction*, Translations of Mathematical Monographs, **104**, American Mathematical Society, Providence, 1992.

[12] R. Dobrushin and S. Shlosman, *Droplet condensation in the Ising model: moderate deviations point of view*, Proceedings of the NATO Ad-vanced Study Institute, Probability theory of spatial disorder and phase transitions, G. Grimmett ed., Kluwer Academic Publishers, vol. 20, 17–34, 1994.

[13] S. Donaldson and P. Kronheimer, *The geometry of four-manifolds*, Oxford Mathematical Monographs, The Clarendon Press, 1990.

[14] N. Dorey, T. Hollowood, V. Khoze, M. Mattis, *The calculus of many instantons*, Phys. Rep. **371** (2002), no. 4-5, 231–459, `hep-th/0206063`.

[15] T. Hollowood, A. Iqbal, C. Vafa, *Matrix models, geometric engineering, and elliptic genera*, J. High Energy Phys. 2008, no. 3, `hep-th/0310272`.

[16] D. Huybrechts and M. Lehn, *The geometry of moduli spaces of sheaves*, Aspects of Mathematics, Vieweg, Braunschweig, 1997.

[17] D. Ioffe and S. Shlosman, *Ising model fog drip: the first two droplets*, In and Out of Equilibrium 2, Progress in Probability 60, 365–382, ed. M.E. Vares, V. Sidoravicius, Birkhäuser, 2008.

[18] D. Ioffe and S. Shlosman, *Ising model fog drip: the puddle*, in preparation.

[19] P. Kasteleyn, *Graph theory and crystal physics*, Graph Theory and Theoretical Physics, 43–110 Academic Press, 1967

[20] R. Kenyon, *Height fluctuations in the honeycomb dimer model*, Comm. Math. Phys. **281** (2008), no. 3, 675–709.

[21] R. Kenyon and A. Okounkov, *Limit shapes and complex Burgers equation*, Acta Math. **199** (2007), no. 2, 263–302.

[22] R. Kenyon, A. Okounkov, and S. Sheffield, *Dimers and amoebae*, Ann. of Math., **163** (2006), no. 3, 1019–1056.

[23] B. Logan and L. Shepp, *A variational problem for random Young tableaux*, Adv. Math.
textbf26, 1977, 206–222.

[24] G. Mikhalkin, *Amoebas of algebraic varieties and tropical geometry*, Different faces of geometry, 257–300, Int. Math. Ser., Kluwer/Plenum, New York, 2004, `math.AG/0403015`.

[25] S. Miracle-Sole, *Surface tension, step free energy and facets in the equilibrium crystal*, J. Stat. Phys. **79**, 183–214 (1995).

[26] S. Miracle-Sole, *Facet shapes in a Wulff crystal*, Mathematical results in statistical mechanics (Marseilles, 1998), World Sci. Publishing, 1999, pp. 83–101.

[27] H. Nakajima, *Lectures on Hilbert schemes of points on surfaces*, University Lecture Series, **18** AMS, Providence, 1999.

[28] H. Nakajima and K. Yoshioka, *Instanton counting on blowup. I. 4-dimensional pure gauge theory*, Invent. Math. **162** 313–355 (2005), `math.AG/0306198`.

[29] H. Nakajima and K. Yoshioka, *Lectures on instanton counting*, Algebraic structures and moduli spaces, 31–101, CRM Proc. Lecture Notes, 38, AMS, Providence, RI, 2004, `math.AG/0311058`.

[30] H. Nakajima and K. Yoshioka, Instanton counting on blowup. II. *K*-theoretic partition function, Transform. Groups **10** (2005), no. 3-4, 489V-519. `math.AG/0505553`.

[31] N. Nekrasov, *Seiberg-Witten prepotential from instanton counting*, Adv. Theor. Math. Phys. **7** (2003), no. 5, 831–864, `hep-th/0206161`.

[32] N. Nekrasov, in preparation.

[33] N. Nekrasov and A. Okounkov, *Seiberg-Witten Theory and Random Partitions*, The Unity of Mathematics (ed. by P. Etingof, V. Retakh, I. M. Singer) Progress in Mathematics, Vol. 244, Birkhäuser. 2006, `hep-th/0306238`.

[34] N. Nekrasov and V. Pestun, *Seiberg-Witten geometry of four dimensional N=2 quiver gauge theories*, `arXiv:1211.2240`

[35] N. Nekrasov, V. Pestun, and S. Shatashvili, *Quantum geometry and quiver gauge theories*, `arXiv:1312.6689`.

[36] Nekrasov, N., Shadchin, S., *ABCD of instantons*, Commun. Math. Phys. **252** (2004) 359–391, `hep-th/0404225`.

[37] B. Nienhuis, H. J. Hilhorst, H. W. J. Blöte, *Triangular SOS models and cubic-crystal shapes*, J. Phys. A **17** (1984), no. 18, 3559–3581.

[38] A. Okounkov, *The uses of random partitions*, XIVth International Congress on Mathematical Physics, 379–403, World Sci. 2005, `math-ph/0309015`.

[39] A. Okounkov, *Random surfaces enumerating algebraic curves*, Proceedings of Fourth European Congress of Mathematics, EMS, 751–768, `math-ph/0412008`.

[40] A. Okounkov, *Random partitions and instanton counting*, International Congress of Mathematicians, Vol. III, 687–711, EMS Zürich, 2006. `math-ph/0601062`.

[41] A. Okounkov, *Noncommutative geometry of random surfaces*, `arXiv:0907.2322`.

[42] M. Passare and H. Rullgøard, *Amoebas, Monge-Ampère measures, and triangulations of the Newton polytope*, Duke Math. J. **121** (2004), no. 3, 481–507.

[43] N. Seiberg and E. Witten, *Electric-magnetic duality, monopole condensation, and confinement in $\mathcal{N} = 2$ supersymmetric Yang-Mills theory*, Nucl. Phys. B426 (1994) 19–52; Erratum-ibid. B430 (1994) 485–486.

[44] N. Seiberg and E. Witten, *Monopoles, duality and chiral symmetry breaking in $\mathcal{N} = 2$ supersymmetric QCD*, Nucl. Phys. B431 (1994) 484–550.

[45] M. Sodin and P. Yuditskii, *Functions that deviate least from zero on closed subsets of the real axis*, St. Petersburg Math. J. **4** (1993), no. 2, 201–249.

[46] S. Shlosman, *The Wulff construction in statistical mechanics and in combinatorics*, Russ. Math. Surv. **56** no. 4, 709–738 (2001)

[47] M. Toda, *Theory of nonlinear lattices*, Springer, Berlin, 1981.

[48] K. Uhlenbeck, *The Chern classes of Sobolev connections*, Comm. Math. Phys. **101** (1985), no. 4, 449–457.

[49] A. Vershik, *Statistical mechanics of combinatorial partitions and their limit configurations*, Func. Anal. Appl. **30**, no. 2, 1996, 90–105.

[50] A. Vershik, S. Kerov, *Asymptotics of the Plancherel measure of the symmetric group and the limit form of Young tableaux*, Soviet Math. Dokl. **18**, 1977, 527–531.

[51] A. Vershik, S. Kerov, *Asymptotics of the maximal and typical dimension of irreducible representations of symmetric group*, Func. Anal. Appl. **19**, 1985, no.1.

[52] E. Witten, *Dynamics of quantum field theory*, Quantum fields and strings: a course for mathematicians, (ed. by P. Deligne, P. Etingof, D. Freed, L. Jeffrey, D. Kazhdan, J. Morgan, D. Morrison and E. Witten), AMS and IAS, vol. 2, 1119–1424, 1999.