CONCENTRATION OF MEASURE[*]

Nathanaël Berestycki and Richard Nickl
with an appendix by Ben Schlein

University of Cambridge

Version of: December 1, 2009

Contents

1 Introduction 3

1.1 Foreword 3

1.2 Definitions 4

1.3 Spherical Isoperimetry 5

1.4 Concentration of Gaussian Measures 8

1.5 Concentration of Product Measures 10

2 Isoperimetric and Poincaré Inequalities 14

2.1 Poincaré implies concentration 15

2.2 Poincaré inequality and eigenvalues 17

2.3 Cheeger's inequality 19

2.4 Discrete case: Markov chains on graphs 22

3 Logarithmic Sobolev Inequalities 27

3.1 Introduction and definition 27

3.2 Log-Sobolev implies Gaussian concentration 29

[*] Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge. Email: n.berestycki@statslab.cam.ac.uk, r.nickl@statslab.cam.ac.uk. These are informal lecture notes, and we claim no originality for any of its contents. Please let us know of any errors in these notes.

4 Concentration of Gaussian Measures and Processes 32

4.1 The Ornstein-Uhlenbeck Semigroup . . . 33

4.2 The logarithmic Sobolev inequality for Gaussian Measures in $\mathbb{R}^n$ . . 36

4.3 Proof of Borell's Inequality . . . 39

4.4 Gaussian Measures in Banach Spaces . . . 40

5 Log-Sobolev Inequalities in Product Spaces 43

5.1 Entropy in product spaces . . . 43

5.2 Further results and applications . . . 45

6 Talagrand's Inequality for Empirical Processes 49

6.1 Empirical Processes . . . 49

6.2 Talagrand's inequality and variations thereof . . . 50

6.3 A Proof of Talagrand's Inequality . . . 52

6.3.1 A bound for the Laplace transform of empirical processes . . 52

6.3.2 A Bernstein-type version of Talagrand's inequality . . . 54

6.3.3 Completion of the Proof of Theorem 14 . . . 57

6.4 Moment Bounds via Rademacher Symmetrization . . . 59

6.4.1 Rademacher Processes . . . 59

6.4.2 Moment Bounds for Empirical Processes . . . 61

6.4.3 A 'statistical version' of Talagrand's Inequality . . . 62

6.5 Sums of i.i.d. Banach-Space valued Random Variables . . . 65

6.6 Estimation of a Probability Measure . . . 66

A Concentration inequalities for Wigner random matrices 69

A.1 Wigner's semicircle law . . . 70

A.2 A general concentration inequality for eigenvalues . . . 71

A.3 Concentration of eigenvalues at the right scale . . . 74

1 Introduction

1.1 Foreword

The *concentration of measure phenomenon* is a notion that was put forward in the 1970's by V. Milman while investigating the asymptotic (i.e., high dimensional) geometry of Banach spaces. It is a notion that has since then proved tremendously useful in a huge variety of contexts, partly due to the work of mathematicians such as M. Gromov, M. Talagrand and M. Ledoux. Let us mention, among other fields: combinatorics, functional and discrete analysis, statistics, probability theory, geometry, statistical physics, and probably much more. It is also a mathematical theory of its own, with its major theorems (e.g., Talagrand's inequality), ramifications into other branches of mathematics, and outstanding open problems.

In this course we will try to convey some basic notions and results related to this phenomenon, and showcase some of its applications in different contexts. Roughly speaking, the course will be divided into three parts of approximately equal length. Michel Ledoux's monograph [31] will be a major source of inspiration throughout the notes.

1. The first part consists of the 'geometric theory' (in particular, isoperimetry and Cheeger's constant) and its functional-analytic counterpart, e.g., Poincaré and Log-Sobolev inequalities.

2. In the second part we show how the analytic theory using 'log-Sobolev' inequalities can be used to establish two main results of the theory: First, we prove "Borell's inequality" [5] by establishing the Log-Sobolev inequality (and thus measure concentration) for Gaussian measures, in a dimension-free way. This has striking ramifications to (and is partly motivated by questions in) probability theory, infinite dimensional analysis and statistics, among other things. Second, we state, discuss and summarize the main ingredients of the proof of what has become known as "Talagrand's inequality" [38] (although there are really many inequalities proved by Talagrand!). Talagrand [37] summarized the intuition behind as 'a new look at independence': *a random variable that smoothly depends on a large number of independent random variables (but not too much on any of them), is 'essentially' constant, in a 'dimension-free' way.* It is clear that any formalization of such a statement is to have profound impact on probability theory and, in particular, on statistics.

3. In the third and final part, we will examine some non-trivial applications to empirical processes in statistical theory, first passage percolation in probability theory, and random matrix theory.

Often, proofs can be formulated either analytically or probabilistically. We will strive to present both points of view, as much as possible. It is usually the case that one proof informs on the other, and vice-versa, so the audience and readers are warmly encouraged to make the effort to understand both points of view!

1.2 Definitions

We start with a few informal examples that describe the notion of measure concentration from a few different points of view.

From a probabilistic perspective, let us agree that a random variable Z defined on some probability space satisfies a concentration inequality if for some constant m – which will typically be EZ or the median of Z – we have for every $u \geq 0$

$$\mathbb{P}\left\{|Z - m| \geq u\right\} \leq c \exp\left\{-\frac{u^2}{2v}\right\}$$

or equivalently

$$\mathbb{P}\left\{|Z - m| \leq u\right\} \geq 1 - c \exp\left\{-\frac{u^2}{2v}\right\} \tag{1}$$

where the constant v is usually related to the variance of Z, and where $c > 0$ should be a small numerical constant. Sometimes the exponent will not be of the Gaussian form u^2, but at least a function of polynomial growth of u should feature in the exponent to convey the notion of concentration. We also emphasize that a concentration inequality should hold *for every* $u \geq 0$: it is different from a 'large deviation inequality' that holds only 'asymptotically' (for large enough u).

From a more geometric perspective, we can say that a measure μ on some metric space (X, d) satisfies a measure concentration principle if, for any set A such that $\mu(A) \geq 1/2$, we have

$$\mu(A_r) \geq 1 - c \exp\left\{-\frac{r^2}{2v}\right\} \tag{2}$$

where A_r denotes the r-enlargement of A: that is, points $x \in X$ within distance r of A. This point of view should be of some use to the combinatorists and analysts: indeed, it is highly reminiscent of the celebrated 'sharp transition thresholds' of Friedgut-Kalai [16].

Since this helps to get an intuition on the phenomenon of measure concentration, it is perhaps worth briefly discussing this result. Consider the random graph process $G(n, p)$, where every edge of the complete graph K_n is present with probability p - the 'density' of edges. Let A is an arbitrary monotone graph property (in the sense that if H satisfies A and H is a subgraph of G then G automatically satisfies A - think for instance of the property 'being connected' or 'containing a cycle'). Then A has a sharp threshold of occurrence. That is, there is p_c such that $G(n, p_c - \varepsilon)$ does not satisfy A with high probability, while $G(n, p_c + \varepsilon)$ satisfies A with high probability, for every $\varepsilon > 0$. Here and in what follows, 'with high probability' means with probability tending to 1 as $n \to \infty$. This statement says, roughly speaking, that if $G(n, p)$ satisfies A with probability bounded away from 0 as $n \to \infty$ for some $p > 0$, then increasing p just slightly will guarantee that A is satisfied with overwhelming probability. The concentration of measure should be viewed similarly. Suppose a set A carries a small but 'bounded away from 0' probability: then an ε-enlargement of A carries almost full probability.

Note that from this description, the connection to isoperimetric problems is already quite intuitive: there is measure concentration if and only if the measure of any set increases very rapidly with its enlargement, i.e., if their 'boundaries' (in a sense that can be made rigorous) carry significant weight.

Naturally, the two points of view (1) and (2) are strictly equivalent by defining $\mu(A) = \mathbb{P}(Z \in A)$.

1.3 Spherical Isoperimetry

It is time to speak about an example. We start with one of the simplest geometric examples, which is the concentration of measure on a high-dimensional sphere. Let $\mathbb{S}^n$ be the unit sphere in $\mathbb{R}^{n+1}$ equipped with the geodesic metric d. Let Z be distributed according to the uniform probability measure μ^n on $\mathbb{S}^n$, and let A be any measurable subset of $\mathbb{S}^n$ that satisfies $\mu^n(A) \geq 1/2$, so that A is a set with significant area. Recall further our notation for the enlargement (or the neighbourhood)

$$A_\varepsilon = \{x \in \mathbb{S}^n : d(x, y) < \varepsilon \text{ for some } y \in A\}.$$

Then we claim that

Theorem 1. *Under the above assumptions,*

$$\mathbb{P}(Z \in A_\varepsilon) = \mu^n(A_\varepsilon) \geq 1 - e^{-(n-1)\varepsilon^2/2}. \tag{3}$$

One particularly striking consequence of this result is that, from a measure point of view, most of the mass of the sphere is concentrated around points within distance $O(1/\sqrt{n})$ of the equator. This (and some further developments) has prompted M. Gromov to say that the observable diameter of the sphere is of order $1/\sqrt{n}$, in contrast with its physical diameter of order 1. See Section 1.4 in Ledoux [31] for more about this notion.

Proof. This proof relies on the spherical isoperimetric inequality and a computation. The spherical isoperimetric inequality, discovered by ... Paul Lévy in 1919, is the n-dimensional generalization of the following statement: among all curves on the unit sphere $\mathbb{S}^2 \subset \mathbb{R}^3$, the ones which enclose the largest area for a given length are circles (which enclose spherical caps). More formally, on a metric measured space (X, d, μ), for a given measurable set $A \subset X$, define the Minkowski boundary measure

$$\mu^+(A) = \liminf_{\varepsilon \to 0} \frac{1}{\varepsilon} \mu(A_\varepsilon \setminus A).$$

We say that X has isoperimetric profile $I : [0, \infty) \to [0, \infty)$ (a function on the real numbers) if for every $m > 0$ and every A such that $\mu(A) = m$ we have

$$\mu^+(A) \geq I(m),$$

and I is the largest function for which this holds. Then the spherical isoperimetric inequality sates that on $X = \mathbb{S}^n$, $I(m) = v'(v^{-1}(m))$, where $v(r)$ is the volume of a (spherical ball or cap) $B(x, r)$ for any $x \in \mathbb{S}^n$ and r such that $\mu(B(x, r)) = m$. Hence $v^{-1}(m)$ is the radius r such that $v(r) = m$. In practice, this means that if $\mu(A) = m$ and if r is such that $\mu(B(x, r) = m$ then

$$\mu(A_\varepsilon) \geq \mu(B(x, r + \varepsilon)).$$

This is, of course, the analogue of the classical isoperimetry in $\mathbb{R}^n$, and we won't include a proof of it here. Once it is accepted, it suffices to prove (3) for A a spherical ball of mass $\geq 1/2$. Thus the rest of the proof basically follows from the following computation on the volume of spherical caps.

Let $v(r) = \mu(A)$ where $A = B(x, r)$ for some (any) $x \in \mathbb{S}^n$. Then for $0 < r < \pi$,

$$v(r) = \frac{1}{s_n} \int_0^r \sin^{n-1} \theta d\theta,$$

where $s_n = \int_0^\pi \sin^{n-1} \theta d\theta$. Since we have assumed that $\mu(A) \geq 1/2$, then $r = \pi/2 + s$ for some $s \geq 0$. Now, let $\varepsilon > 0$ and let $s' = s + \varepsilon$

$$\begin{aligned} 1 - v(r + \varepsilon) &= s_n^{-1} \int_{r+\varepsilon}^{\pi} \sin^{n-1} \theta d\theta \\ &= s_n^{-1} \int_{s+\varepsilon}^{\pi/2} \cos^{n-1} \theta d\theta \end{aligned}$$

Making the change of variable $\theta = \tau/\sqrt{n-1}$, and using the inequality $\cos u \leq$

$e^{-u^2/2}$ for $0 \leq u \leq \pi/2$, we find

$$\begin{aligned}\int_{s+\varepsilon}^{\pi/2} \cos^{n-1}\theta d\theta &= \frac{1}{\sqrt{n-1}} \int_{(s+\varepsilon)\sqrt{n-1}}^{(\pi/2)\sqrt{n-1}} \cos^{n-1}\left(\frac{\tau}{\sqrt{n-1}}\right) d\tau \\ &\leq \frac{1}{\sqrt{n-1}} \int_{(s+\varepsilon)\sqrt{n-1}}^{\infty} e^{-\tau^2/2} d\tau \\ &\leq \frac{\sqrt{\pi}}{\sqrt{2(n-1)}} e^{-(n-1)(s+\varepsilon)^2/2} \leq \frac{\sqrt{\pi}}{\sqrt{2(n-1)}} e^{-(n-1)\varepsilon^2/2}\end{aligned}$$

We now bound s_n from below. Integrating by parts twice, we get $s_n = ((n-2)/(n-1))s_{n-2}$, from which

$$\sqrt{n-1} s_n \geq \sqrt{n-3} s_{n-2}$$

Using this inequality inductively, we obtain $s_n \geq 2/\sqrt{n-1}$. Putting the pieces together gives us:

$$1 - v(r+\varepsilon) \leq e^{-(n-1)\varepsilon^2/2}$$

and thus

$$\mu(A_\varepsilon) \geq \mu(B(x, r+\varepsilon)) \geq 1 - e^{-(n-1)\varepsilon^2/2}$$

which is what we wanted. □

A pretty interesting consequence of this is the following: any Lipschitz function on the sphere (i.e., and function whose local oscillations are small) must be concentrated! Recall that a function $F : \mathbb{S}^n \to \mathbb{R}$ is said to be a C-Lipschitz function, if

$$\sup_{x \neq y, x, y \in \mathbb{S}^n} \frac{|F(x) - F(y)|}{d(x, y)} \leq C.$$

Let m_F be the median of F, that is, m_F is a number which satisfies

$$\mu^n \{x : F(x) \geq m_F\} \geq 1/2 \quad \textit{and} \quad \mu^n \{x : F(x) \leq m_F\} \geq 1/2.$$

(Note that m_F doesn't have to be unique but always exists).

Corollary 1.

$$\mu^n \{x \in \mathbb{S}^n : |F(x) - m_F| \geq r\} \leq 2 \exp\left(-\frac{(n-1)r^2}{2C^2}\right) \tag{4}$$

where m_F is the median of F.

Proof. Let $A = \{F \leq m_F\}$ i.e., $A = \{x \in \mathbb{S}^n : F(x) \leq m_F\}$. If $\varepsilon > 0$ and $d(x, A) \leq \varepsilon$ then there exists $y \in A$ such that $d(x.y) \leq \varepsilon$. Since F is C-Lipschitz,

$$F(x) \leq F(y) + C\varepsilon \leq m_F + C\varepsilon.$$

Taking the contrapposite, $\{F > m_F + C\varepsilon\} \subset A_\varepsilon^c$, so that, taking probabilities:

$$\mu(F > m_F + C\varepsilon) \leq 1 - \mu(A_\varepsilon) \leq \exp\left(-\frac{(n-1)r^2}{2}\right).$$

Similarly,

$$\mu^n(F < m_F - C\varepsilon) \leq 1 - \mu(A_\varepsilon) \leq \exp\left(-\frac{(n-1)r^2}{2}\right).$$

Changing r into r/C and combining these two inequalities, we obtain

$$\mu^n \{x \in \mathbb{S}^n : |F(x) - m_F| \geq r\} \leq 2\exp\left(-\frac{(n-1)r^2}{2C^2}\right)$$

as requested. □

So again – choosing $r = t/\sqrt{n}$ – we see that the measure of the set of points in $\mathbb{S}^n$ for which $F(x)$ deviates more than $t/\sqrt{n}$ from a constant (here m_F) is less than or equal to $2e^{-\frac{t^2}{2}(1-\frac{1}{n})}$. A visual interpretation of this pheonomeon – due to Gromov – is the following: The unit sphere $\mathbb{S}^n$ cannot be 'observed' by us for $n \geq 3$. However, we could think of a visual machine that sends points on high-dimensional spheres into one-dimensional information (so an 'observable'). We would think that any 'reasonable machine' is such that local oscillations in the domain result in oscillations of the same size in the image – meaning that the machine is a Lipschitz function. The last inequality then says that 'the observable diameter of $\mathbb{S}^n$ is effectively of size $1/\sqrt{n}$, whereas its diameter as a metric space is constant'.

1.4 Concentration of Gaussian Measures

A centered Gaussian or normal real random variable with variance σ^2 (we shall often write $N(0, \sigma^2)$) is defined by its probability density

$$\phi_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2}{2\sigma^2}\right\}.$$

These random variables are central to probability theory and statistics for various reasons. We start with a simple calculus exercise.

Proposition 1. *If X is a normally distributed random variable with mean EX and variance σ^2, then X concentrates around a constant, namely its mean, in the sense that, for every $u \geq 0$*

$$\Pr\{|X - EX| \geq u\} \leq \exp\left\{-\frac{u^2}{2\sigma^2}\right\}.$$

Proof. We can assume $EX = 0$ and, replacing X by X/σ, also that $\sigma = 1$. If $\Phi(u) = P(X \leq u)$ denotes the cumulative distribution function of X then we want to prove $2\Phi(-u) \leq e^{-u^2/2}$ for every $u \geq 0$. This is obviously true for $u = 0$, and follows from differentiating the inequality

$$2\Phi(-u) = \sqrt{\frac{2}{\pi}} \int_u^\infty e^{-x^2/2} dx \leq e^{-u^2/2}$$

with respect to u in the range $0 < u \leq \sqrt{2/\pi}$. For $u > \sqrt{2/\pi}$ we use the well known bound

$$\Phi(-u) \leq \frac{1}{\sqrt{2\pi}} \frac{e^{-u^2/2}}{u} \tag{5}$$

which follows from integration by parts

$$\int_u^\infty x e^{-x^2/2} \frac{1}{x} dx = \frac{e^{-u^2/2}}{u} - \int_u^\infty \frac{1}{x^2} e^{-x^2/2} dx \leq \frac{e^{-u^2/2}}{u}.$$

□

While the more classical bound (5) is more exact in the 'large deviation' context (that is, for u large), it is not optimal for *every* $u > 0$. The last proposition shows that in the Gaussian case a genuine concentration inequality of type (1) holds with exponent $-u^2$, $v = \sigma^2$ and leading constant $c = 1$ – which is often referred to as *Gaussian concentration.*

Clearly Proposition 1 follows merely from the definition of Gaussian random variables and elementary calculus, so is not at all surprising. However, we shall see that Proposition 1 is a special case of a very general, *dimension free* phenomenon that Gaussian measures and processes concentrate around their mean.

For instance we shall prove the following result: Let $(B, \|\cdot\|_B)$ be a (separable) Banach space and let $X : (\Omega, \mathcal{A}, \Pr) \to B$ be a random variable taking values in B (i.e., a measurable mapping where B is equipped with its Borel-σ-field). We say that the random variable X is centered *Gaussian* if $L(X)$ is normally distributed with mean zero for every continuous linear form $L : B \to \mathbb{R}$. It is easily seen that this requirement characterizes normal laws in finite-dimensional real vector spaces (e.g., [11] Theorem 9.5.13), but otherwise serves as the natural generalization. A consequence of Borell's inequality will be that

$$\Pr\left\{|\|X\|_B - E\|X\|_B| \geq u\right\} \leq \exp\left\{-\frac{u^2}{2\sigma^2}\right\} \tag{6}$$

where

$$\sigma^2 = \sup_{L:\|L\|'_B\leq 1} EL^2(X)$$

are the 'weak' variances (and where $\|\cdot\|'_B$ is the norm of the dual space of B).

We should note that (6) does almost but not exactly yield Proposition 1 as a corollary, but the unifying approach via the theory of Gaussian processes will: We shall prove for (almost) arbitrary centered Gaussian processes $\{X_t\}_{t\in T}$ where T is *any* index set that

$$\Pr\left\{\left|\sup_{t\in T} X_t - E\sup_{t\in T} X_t\right| \geq u\right\} \leq \exp\left\{-\frac{u^2}{2\sigma^2}\right\}$$

where $\sigma^2 = \sup_{t\in T} EX_t^2$. This result clearly implies Proposition 1 but will also be seen to imply (6), and has substantial implications in the theory of Gaussian processes (such as Brownian motion, for instance).

1.5 Concentration of Product Measures

Let now $X_1, ..., X_n$ be independent and identically distributed random variables with law P supported in $[-1, 1]$ (as long as the range is bounded the restriction to $[-1, 1]$ is immaterial, as one can always rescale), and denote by $\Pr := P^n$ the product probability measure representing the joint law of $X_1, ..., X_n$.

The sample mean $\frac{1}{n}\sum_{i=1}^n X_i$ is a fundamental quantity in almost all of statistics, but also in the study of random walks and in many other problems in probability theory. It turns out that the sample mean represents another instance of the 'dimension-free measure concentration'.

As a first step in this direction, let us briefly prove a classical result due to Hoeffding [26], in dimension one.

Proposition 2. *Let $X_1, ..., X_n$ be i.i.d. random variables bounded in absolute value by one and with $EX = 0$. Then we have, for every $n \in \mathbb{N}$ and every $u > 0$*

$$\Pr\left\{\frac{1}{n}\sum_{i=1}^n X_i \geq u\right\} \leq \exp\left\{-\frac{u^2}{2}n\right\}. \tag{7}$$

and hence also

$$\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \geq u\right\} \leq 2\exp\left\{-\frac{u^2}{2}n\right\}. \tag{8}$$

Proof. We first derive a bound on the moment generating function of vX for $v > 0$, namely

$$Ee^{vX} \le e^{v^2/2}. \tag{9}$$

To see this, observe that convexity of the exponential function implies

$$e^{vx} \le \frac{1-x}{2}e^{-v} + \frac{1+x}{2}e^{v}, \ -1 \le x \le 1.$$

Taking expectations and using $EX = 0$ we see

$$Ee^{vX} \le \frac{1}{2}e^{-v} + \frac{1}{2}e^{v} = e^{-v+\log\left(\frac{1}{2}+\frac{1}{2}e^{2v}\right)} =: e^{g(v)}$$

where $g(v) = -v + \log\left(\frac{1}{2} + \frac{1}{2}e^{2v}\right)$. Clearly $g(0) = g'(0) = 0$, and also $g''(v) \le 1$ for all $v \ge 0$, so that a Taylor series expansion gives $g(v) = (v^2/2)g''(\tilde{v}) \le v^2/2$, completing the proof of (9).

Now to prove Hoeffding's inequality, recall Markov's inequality $\Pr\{|X| > C\} \le E|X|/C$. Then, for every $t > 0$, $v > 0$, using that the X_i's are i.i.d and (9),

$$\begin{aligned}\Pr\left\{\sum_{i=1}^n X_i \ge t\right\} &= \Pr\left\{e^{v\sum_{i=1}^n X_i} \ge e^{tv}\right\} \le e^{-tv}Ee^{v\sum_{i=1}^n X_i} \\ &= e^{-tv}\Pi_{i=1}^n Ee^{vX_i} \le e^{-tv}\Pi_{i=1}^n e^{v^2/2} \\ &= e^{-tv+\frac{v^2 n}{2}} = e^{-\frac{t^2}{2n}}\end{aligned}$$

by choosing $v = t/n$. Finally (7) simply follows from multiplying the inequality featuring in the probability in question by n and applying the last inequality with $t = nu$. □

The result says that a sample mean of n bounded i.i.d. random variables satisfies a concentration inequality of type (1) around its mean $m = EX$ with exponent u^2, $v = 2/n$ and leading constant $c = 2$. In particular, the 'degree of concentration', measured by the exponent in the inequality, improves as dimension (i.e., 'sample size') n increases.

While Proposition 2 is a genuine concentration inequality, it is still somewhat unsatisfactory: One would think that the size of the random fluctuations of $\frac{1}{n}\sum_{i=1}^n X_i$ around the constant EX should depend on the *variance* σ^2 of X, and that therefore σ^2 should be featuring in the exponent on the r.h.s. of (8), as it does in Proposition 1 for example. This more refined result is known as Bernstein's inequality: If $X_1, ..., X_n$ are as in Theorem 2, and if $\sigma^2 = Var(X)$, then

$$\Pr\left\{\frac{1}{n}\sum_{i=1}^n X_i \ge u\right\} \le \exp\left\{-\frac{u^2}{2\sigma^2 + \frac{2u}{3}}n\right\}. \tag{10}$$

and thus also

$$\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq u\right\} \leq 2\exp\left\{-\frac{u^2}{2\sigma^2 + \frac{2}{3}u}n\right\}. \tag{11}$$

The proof – which is only marginally more involved than the one of Proposition 2 – can be found, for instance, in [34]. To interpret this inequality, it is instructive to rescale the centered sample mean, so that one has the equivalent inequality

$$\Pr\left\{\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} X_i \geq v\right\} \leq \exp\left\{-\frac{v^2}{2\sigma^2 + \frac{2v}{3\sqrt{n}}}\right\}. \tag{12}$$

This suggests that – as $n \to \infty$ – this inequality approaches Gaussian concentration. This of course has a deeper meaning that is linked to the central limit theorem, which states that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \to^d N(0, \sigma^2)$$

as $n \to \infty$, which gives a heuristic explanation of (12) at least for large n. In fact one may at first sight hope for 'pure' Gaussian concentration $e^{-v^2/2\sigma^2}$ in (12), and wonder why the $\frac{2v}{3\sqrt{n}}$ term occurs in the denominator of the exponent. That this term cannot be improved uniformly in u, however, can be seen as follows: Take X_i i.i.d. Bernoulli variables with success probability $p := p_n$ very small depending on n, say $p_n = 1/n$. A standard result in probability is that the distribution of $\sum_{i=1}^{n} X_i$ can in this case be very well approximated by the distribution of a Poisson variable Y with rate 1, in fact one can show that

$$\left|\Pr\left\{\sum_i X_i \geq u\right\} - P\left\{Y \geq u\right\}\right| \leq 1/n.$$

Since the tail of a Poisson random variable is of order e^{-u} and not e^{-u^2}, we cannot expect 'pure' Gaussian concentration in (12) if σ^2 is very small and/or depends on n (as in this Bernoulli example). Note also that this does not contradict the 'limiting' Gaussian concentration suggested by the central limit theorem, as the distribution of the X_i's in this 'counter-example' changes with n.

This shows that product measures of bounded random variables in one dimension satisfy concentration inequalities. It was mentioned after Proposition 1 that one-dimensional Gaussian concentration extends to arbitrary dimensions without a price by virtue of Borell's inequality, and one may ask the same question for product measures. *One of the most striking achievements in the theory of measure concentration in the last decades is Talagrand's inequality [38], which shows* *that the same is true for product measures in infinite dimensions.* For instance Talagrand's inequality for *empirical processes* (these processes, which will be defined precisely below, parallel the role of Gaussian processes in the treatment of Gaussian measures in a way) will imply that for $X_1, ..., X_n$ i.i.d. centered random variables in a Banach space $(B, \|\cdot\|_B)$ that are bounded ($\|X\|_B \leq 1$) one has a genuine concentration inequality (comparable to (11) above):

$$\Pr\left\{\left|\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\right\|_B - E\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\right\|_B\right| \geq u\right\} \leq \exp\left\{-\frac{u^2}{2V+2u}n\right\} \tag{13}$$

where V is a quantity that needs some more careful discussion (but will depend on the variance of X in a suitable way).

Someone who knows the field of probability in Banach spaces will immediately recognize the following: The central limit theorem for a sum $\sum_{i=1}^{n} X_i$ of general B-valued centered random variables holds only under certain conditions on the geometry of B (see [1] and [32]), whereas the concentration inequality (13) will be seen to hold (in more or less) arbitrary Banach spaces. This offers one of the deepest insights in the theory treated in this course: *concentration of product measures is – contrary to what one may expect at first – NOT related to the central limit theorem, but a self-standing probabilistic phenomenon.* A complete understanding of this involves some subtleties and requires more context, so we postpone exact statements and definitions to Section 6.

2 Isoperimetric and Poincaré Inequalities

In this section we discuss a class of inequalities calles Poincaré inequalities, and their relation to, on the one hand, spectral quantities (ie., having to do with the eigenvalues of a certain operator), and, on the other hand, isoperimetric problems. We will notably discuss the celebrated Cheeger inequality in this context. We start by defining what is a Poincaré inequality. Recall that the variance of a random variable f with respect to a probability measure μ is $\mathrm{Var}_\mu(f) = \mathbb{E}_\mu((f-m)^2) = \mathbb{E}_\mu(f^2) - \mathbb{E}_\mu(f)^2$.

Definition 1. *A measure μ on a space X is said to satisfy a Poincaré inequality with constant c if*

$$\mathrm{Var}_\mu(f) \leq c \int_X |\nabla f|^2 d\mu \tag{14}$$

for every $f : X \to \mathbb{R}$ such that the various terms in (14) make sense.

Usually we will have in mind that X is a smooth Riemaniann manifold with metric g which is either compact or open and bounded, and that μ is a probability measure on X. However, not much is lost in thinking that X is a finite graph and that ∇f denotes the discrete derivative $f(y) - f(x)$ for x and y neighbours in f (the integral then becomes a sum over the edges). We will come back to the discrete case in more specific details below.

Poincaré proved that if X is the closure of an open domain in $\mathbb{R}^n$ with smooth boundary, equipped with the standard Euclidean metric and Lebesgue measure, then X satisfies (14) for some $c > 0$. Here is an elementary proof in dimension $n = 1$, in which case we may assume without loss of generality that $X = (0,1)$. Then if f is $\mathcal{C}^1$ and m denotes the average of f, we have:

$$\begin{aligned}
\int_0^1 |f(s) - m|^2 ds &= \int_0^1 \left| f(s) - \int_0^1 f(t)dt \right|^2 ds \\
&\leq \int_0^1 \int_0^1 |f(s) - f(t)|^2 ds dt \\
&= \int_0^1 \int_0^1 \left| \int_s^t f'(u) du \right|^2 ds dt \\
&\leq \int_0^1 \int_0^1 |t - s| \int_s^t f'(u)^2 du ds dt
\end{aligned}$$

by applying Cauchy-Schwarz's inequality (twice). Thus, by Fubini's theorem:

$$\begin{aligned}\int_0^1 |f(s)-m|^2 ds &\leq 2\int_0^1 f'(u)^2 du \int_0^1\int_0^1 |t-s|\mathbf{1}_{\{s\leq u\leq t\}} dt ds \\ &\leq 2\int_0^1 f'(u)^2 du [u(1-u)-\frac{u^2}{2}(1-u)] \\ &\leq \frac{1}{2}\int_0^1 f'(u)^2 du,\end{aligned}$$

and (14) holds with $c = 1/2$. (Better optimal constants are possible).

Intuition. To say that there is a Poincaré inequality amounts to the following statement. Say a number $\sigma > 0$ is given. We try to build a function f on X such that $\mathbb{E}(f) = 0$ and the standard deviation of$f$ is σ (fixed). Then there is a Poincaré inequality if there is always a "smoothest" function which achieves this standard deviation. In particular any function with zero mean and standard deviation σ has to be rougher than this function, in the sense that its gradient has larger L^2 norm. This is a highly intuitive fact for domains in $\mathbb{R}^d$, so unsurprisingly the Poincaré inequality holds with a great degree of generality - as we will soon see. But first, we explain why we are interested in this inequality.

Remark 1. *In (14), the L^2 norm of the gradient makes sense even if the function is not $\mathcal{C}^2$, e.g. if the weak derivatives exist and are L^2 functions as in the Sobolev space $W^{1,2} = H^1$. In proofs below, we will freely apply the Poincaré inequality for functions in $W^{1,2}$.*

2.1 Poincaré implies concentration

Theorem 2. *Assume that (X, g) satisfies a Poincaré inequality (14) for some $c > 0$ and that μ is absolutely continuous with respect to the volume element. Then if $\mu(A) \geq 1/2$, we have for all $\varepsilon > 0$*

$$\mu(A_\varepsilon) \geq 1 - \exp\left(-\frac{\varepsilon}{3\sqrt{c}}\right). \tag{15}$$

In other words, a Poincaré inequality always implies an exponential concentration for (X, μ).

Proof. Let A, B be subsets of X such that $d(A, B) = \varepsilon$ (later, we will take B to be the complement of A_ε). Define a function $f : X \to \mathbb{R}$ by the following. Let $a = \mu(A)$ and let $b = \mu(B)$. Let f be a function such that $f(x) = 1/a$ on A and

$f(x) = -1/b$ on B. In between A and B, interpolate as smoothly as possible, say linearly. For instance, take

$$f(x) = \frac{1}{a} - \frac{1}{\varepsilon}\left(\frac{1}{a} + \frac{1}{b}\right)\min(\varepsilon, d(x, A)).$$

f is not smooth but is in the space $W^{1,2}$, so we can apply the Poincaré inequality (14) to it (this is where we use the fact that μ is absolutely continuous with respect to the volume element). This will give us a lower bound on the total roughness of f and thus may tell us how b and a differ - in particular, we hope to show concentration, so that b is indeed much smaller than a.

Note that since f is constant on $A \cup B$, we have $\nabla f = 0$ on this set. Moreover,

$$|\nabla f| \leq \frac{1}{\varepsilon}\left(\frac{1}{a} + \frac{1}{b}\right),$$

μ-almost surely. Thus, integrating:

$$\int |\nabla f|^2 d\mu \leq \frac{1}{\varepsilon^2}\left(\frac{1}{a} + \frac{1}{b}\right)^2 (1 - a - b).$$

On the other hand, if m denotes the mean of f,

$$\begin{aligned}\mathrm{Var}_\mu(f) &= \int (f-m)^2 d\mu \\ &\geq \int_A (f-m)^2 d\mu + \int_B (f-m)^2 d\mu \\ &\geq a(1/a - m)^2 + b(-1/b - m)^2\end{aligned}$$

Calculus shows that the right-hand side is minimized for $m = 0$. Thus, after expanding, we find:

$$\mathrm{Var}_\mu(f) \geq \frac{1}{a} + \frac{1}{b}.$$

Plugging this into (14), this implies:

$$\frac{1}{a} + \frac{1}{b} \leq c\frac{1}{\varepsilon^2}\left(\frac{1}{a} + \frac{1}{b}\right)^2 (1 - a - b)$$

or equivalently

$$\frac{\varepsilon^2}{c} \leq \left(\frac{1}{a} + \frac{1}{b}\right)(1 - a - b) = \frac{(1-a-b)(a+b)}{ab} \leq \frac{1-a-b}{ab} \leq \frac{1-a}{ab} - \frac{1}{a}$$

where we have used that $a + b \leq 1$. Rearranging, we get (since $a \geq 1/2$):

$$b \leq \frac{1-a}{a} \frac{1-a}{1/a + \varepsilon^2/c} \leq \frac{1}{1 + a\varepsilon^2/c} \leq \frac{1-a}{1 + \varepsilon^2/(2c)}.$$

If $B = A_\varepsilon^c$ and $\varepsilon^2/(2c) = 1$ or $\varepsilon = \sqrt{2c}$, then we find

$$\mu(A_\varepsilon^c) \leq \mu(A^c)\frac{1}{2}$$

and, iterating, $1 - \mu(A_{k\varepsilon}) \leq 2^{-k-1}$. Thus if $r > 0$ and if $k\varepsilon \leq r < (k+1)\varepsilon$, noting that $r \mapsto \mu(A_r^c)$ is monotone non-increasing,

$$1 - \mu(A_r) \leq 2^{-k-1} \leq \exp\left(-\frac{\log 2}{\sqrt{2c}} r\right).$$

Since $\log 2/\sqrt{2} = 0.49... > 1/3$, we get

$$1 - \mu(A_r) \leq \exp\left(-\frac{r}{3\sqrt{c}}\right)$$

as claimed. □

Note. Ledoux's proof of this result, p. 48, contains several typos that affect the proof.

2.2 Poincaré inequality and eigenvalues

On every Riemaniann manifold (X, g) there is one particularly natural measure which is the volume element dv. If furthermore, X is compact then dv is a finite measure and we can define $\mu = dv/V$, where V is the total volume, so that μ is a probability measure. In that case, we claim that (X, g, μ) always satisfies a Poincaré inequality, and furthermore that the constant $c > 0$ is the first eigenvalue of an operator $-\Delta$, where Δ is defined on functions on X and is called the Laplace-Beltrami operator, or shortly the *Laplacian*. Formally, if f is a smooth function on X, then

$$\Delta f = \operatorname{div}(\nabla f)$$

as for the usual Laplacian in $\mathbb{R}^n$. Alternatively, one may think of a general finite graph X and a Markov chain defined on X through its transition probability matrix P. In that case $\Delta = P - I$, where I is the identity matrix.

In all those cases, spectral theory guarantees that the eigenvalues of $-\Delta$ are all nonnegative, and form a discrete set. Thus they can be ordered in a nondecreasing way $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots$. Usually we are interested only on strictly positive eigenvalues. We have written here $\lambda_i \leq \lambda_{i+1}$ because the eigenvalues may have nontrivial multiplicity. However, it is a fact that the first eigenvalue λ_1 has multiplicity 1: see Proposition VII.4.1 in [8] or Corollary 2 in Chapter I of [7].

Remark 2. *if the Riemaniann manifold X is not compact, it becomes necessary to impose boundary conditions in order to get a discrete spectrum. These are typically of the* Dirichlet *type, meaning that $f|_{\partial X} = 0$, or of* Neumann *type: that is, $\partial f/\partial n = 0$ on ∂X. In both cases there is a discrete spectrum, however in the Neumann case the first eigenvalue is $\lambda_1 = 0$, corresponding to the constant eigenfunction equal to 1 on X. On a compact manifold, or in the Dirichlet case for open bounded manifolds, then it is a fact that $\lambda_1 > 0$.*

A good and short introduction to the Laplacian on Riemaniann manifolds can be found in the book by Chavel [7], Chapter I. Chapters VII.1 and VII.2 in another book by Chavel, [8], contains many more details, but is still very accessible.

Theorem 3. *Assume that (X, g) is a compact Riemaniann manifold. Then a Poincaré inequality (14) holds with $c = 1/\lambda_1$. Moreover, $1/\lambda_1$ is the smallest constant for which (14) holds. In particular, there is the concentration inequality: for all measurable $A \subset X$ such that $\mu(A) \geq 1/2$, and for all $\varepsilon > 0$*

$$\mu(A_\varepsilon) \geq 1 - \exp\left(-\frac{\varepsilon\sqrt{\lambda_1}}{3}\right). \tag{16}$$

Proof. We may regard $\mathcal{C}^2$ functions on X as elements of $L^2(X)$, which is a Hilbert space when equipped with $(f, g) = \int fg d\mu$. Green's formula states that for any $\mathcal{C}^2$ functions f and g,

$$\int (\Delta f) g d\mu = - \int (\nabla f, \nabla g) d\mu.$$

(Since the manifold is compact there is no boundary term). In particular, $-\Delta$ is self-adjoint and nonnegative for this scalar product, hence the eigenspaces are orthogonal to one another, and there exists a complete orthonormal basis of $L^2(X)$ which consists of eigenfunction $f_1, \ldots$ that are associated with eigenvalues $\lambda_1, \ldots$. Thus for every $f \in L^2(X)$ one may write

$$f = \sum_{j=1}^{\infty} (f, f_j) f_j$$

and by Parseval's identity

$$\|f\|^2 = \sum_{j=1}^{\infty} (f, f_j)^2.$$

Now, let $\mathcal{E}(f, g) = \int (\nabla f, \nabla g) d\mu$, the so-called Dirichlet energy. Then $\mathcal{E}$ is bilinear symetric, so if $r \geq 1$ one has, letting $\alpha_j = (f, f_j)$

$$\begin{aligned}
0 \leq \mathcal{E}&\left(f - \sum_{j=1}^{r} \alpha_j f_j, f - \sum_{j=1}^{r} \alpha_j f_j\right) \\
&= \mathcal{E}(f,f) - 2\sum_{j=1}^{r} \alpha_j \mathcal{E}(f, f_j) + \sum_{j,k=1}^{r} \alpha_j \alpha_k \mathcal{E}(f_j, f_k) \\
&= \mathcal{E}(f,f) + 2\sum_{j=1}^{r} \alpha_j (f, \Delta f_j) - \sum_{j,k=1}^{r} \alpha_j \alpha_k (f_j, \Delta f_k) \\
&= \mathcal{E}(f,f) - 2\sum_{j=1}^{r} \alpha_j^2 \lambda_j + \sum_{j=1}^{r} \alpha_j^2 \lambda_j
\end{aligned}$$

Thus $\mathcal{E}(f,f) \geq \sum_{j=1}^{r} \alpha_j^2 \lambda_j$ for all $r \geq 1$, from which we deduce:

$$\sum_{j=1}^{\infty} \alpha_j^2 \lambda_j < \infty$$

and

$$\mathcal{E}(f,f) \geq \sum_{j=1}^{\infty} \alpha_j^2 \lambda_j \geq \lambda_1 \sum_{j=1}^{\infty} \alpha_j^2 = \lambda_1 \|f\|^2$$

by Parseval's identity. Note that if f has zero mean, then $\|f\|^2 = \operatorname{Var}_\mu f$. By adding to a zero-mean f a constant number, note that neither the variance nor the Dirichlet energy are changed, from which we deduce that for any function f on X:

$$\operatorname{Var}_\mu f \leq \frac{1}{\lambda_1} \mathcal{E}(f,f)$$

as claimed. To see that this is sharp, take $f = f_1$. □

2.3 Cheeger's inequality

We now explain the connection of the Poincaré inequality to isoperimetric problems. This explains Theorem 2 from a more geometric perspective. Recall the definition of the isoperimetric constant:

$$h = \inf_{\Omega \subset X} \frac{\mu^+(\Omega)}{\mu(\Omega)},$$

where Ω is open with compact closure within X.

It is not hard to see that if X is compact then h, as defined here, must be equal to 0 (take, for instance, X itself or X minus a small ball of radius $\varepsilon > 0$, and let $\varepsilon \to 0$.) Thus in the compact case, one needs to change slightly the definition, for instance by requiring that $\mu(\Omega) \leq 1/2$. But for the sake of simplicity we only speak about the open bounded case.

It can be shown that if X is a n-dimensional Riemaniann manifold ($n \geq 2$), then

$$h = \inf_{\Omega} \frac{A(\partial\Omega)}{\mathrm{Vol}(\Omega)} \tag{17}$$

where A is the $n-1$-dimensional Riemaniann area measure for submanifolds of X. We have seen how we can expect to have exponential concentration of the form $\mu(A_r) \geq 1 - e^{-hr}$ for $\mu(A) \geq 1/2$. The following fundamental inequality shows that (up to the the constants in the exponential) what we get from a Poincaré inequality is stronger.

Theorem 4. (Cheeger's inequality) *The following inequality holds in complete generality:*

$$h^2/4 \leq \lambda_1,$$

where λ_1 is the first eigenvalue of $-\Delta$ with Dirichlet boundary conditions.

Proof. The proof consists of two results which are interesting in their own right. First, introduce a more general 'isperimetric constant' h_ν for $\nu > 1$:

$$h_\nu = \inf_{\Omega} \frac{A(\partial\Omega)}{\mathrm{Vol}(\Omega)^{1-1/\nu}} \tag{18}$$

If $0 < h_\nu < \infty$ then ν may be called an 'isoperimetric dimension" of X, and it is fairly obvious that $h_\nu > 0$ can only occur if $\nu \geq n$, the dimension of the manifold. The standard isoperimetric number of (17)is obtained by taking $\nu = \infty$ in (18). The first one says that the isoperimetric constant can be viewed as an L^1 equivalent of the Poincaré constant.

Lemma 1. (Federer-Fleming theorem)

$$h_\nu = \inf_{f \neq 0} \frac{\int |\nabla f| d\mu}{\|f\|_{\nu/(\nu-1)}}$$

where the infimum is taken over $f \in \mathcal{C}_c^\infty$.

Proof. The proof is closely related to the arguments in the proof of Theorem 2, so we skip it. (This is an adaptation of Theorem II.2.1 in [8]). □

The second result is the Nirenberg-Sobolev inequality, which is a particular case of the well-known Sobolev embedding (on which we will come back later).

Lemma 2. *There is the following inequality: for any function $\phi \in \mathcal{C}_c^\infty$, then*

$$\|\nabla\phi\|_2 \geq \frac{\nu-2}{2(\nu-1)} h_\nu \|\phi\|_{2\nu/(\nu-2)}.$$

Proof. We follow Lemma VI.1.1 in [8]. We may assume that $\nu > 2$ and that $h_\nu > 0$ without loss of generality. Let $f = |\phi|^p$ where

$$p = \frac{2(\nu-1)}{\nu-2}.$$

Then $p \geq 2$, and $f \in \mathcal{C}_c^\infty$ a.e.-$d\mu$. By the Federer-Fleming theorem,

$$\|\nabla f\|_1 \geq h_\nu \|f\|_{\nu/(\nu-1)}. \tag{19}$$

But note that

$$|\nabla f| = p|\phi|^{p-1}|\nabla\phi|, \quad \text{a.e.-}d\mu.$$

Therefore, integrating and applying the Cauchy-Schwarz inequality, we obtain:

$$\int |\nabla f| d\mu \leq p\|\nabla\phi\|_2 \|\phi\|_{2(p-1)}^{p-1}.$$

Putting this together with (19), we get

$$p\|\nabla\phi\|_2 \|\phi\|_{2(p-1)}^{p-1} \geq h_\nu \|f\|_{2\nu/(\nu-2)}$$

Since $f = |\phi^p|$, we get

$$\|\nabla\phi\|_2 \geq \frac{1}{p} h_\nu \frac{\|f\|_{\nu/(\nu-1)}}{\|\phi\|_{2(p-1)}^{p-1}} = \frac{1}{p} h_\nu \|\phi\|_{2\nu/(\nu-2)}$$

as claimed. □

Now, letting $\nu \to \infty$ in the Nirenberg-Sobolev inequality we obtain:

$$\|\nabla\phi\|_2 \geq \frac{1}{2} h \|\phi\|_2, \tag{20}$$

for any $\phi \in \mathcal{C}^\infty$. Let $f = f_1$, the first eigenfunction of the Dirichlet Laplacian. We claim that this inequality (20) also holds for $\phi = f$. This is because, since f has Dirichlet boundary conditions, we can find a sequence of functions $\phi_n \in \mathcal{C}^\infty$ such that $\phi_n \to f$ in Sobolev norm, i.e., $\nabla(\phi_n - f)$ and $\phi_n - f$ both tend to 0 in the L^2 norm. Thus applying (20) to ϕ_n and letting $n \to \infty$, this also applies to f. However, by Green's formula

$$\|\nabla f\|_2 = \sqrt{\lambda_1} \|f\|_2,$$

which, after squaring in (20), completes the proof. □

2.4 Discrete case: Markov chains on graphs

Much of the theory developed above has an equivalent for the discrete case, where the manifold X is replaced with a graph $G = (V, E)$ and the gradient of a function $f : V \to \mathbb{R}$ defined on the vertices of the graph, is given by the discrete derivative

$$\nabla f(e) = f(y) - f(x), \quad e = (x, y) \in E.$$

Let us assume for simplicity that G is undirected (we will only use the gradient squared, so the orientation doesn't matter anyway), and consider the simple random walk on G: thus if $p(x, y)$ denotes the transition probabilities: $p(x, y) = 1/\deg(x)$ if $y \sim x$, and 0 otherwise. It has a reversible equilibrium probability distribution $\mu(x)$. That is,

$$\mu(x)p(x, y) = \mu(y)p(y, x) \tag{21}$$

and $\mu(x)$ is proportional to $\deg(x)$. Define the Dirichlet energy of this function f is given by

$$\mathcal{E}(f, f) = \frac{1}{2}\sum_{x,y}(f(y) - f(x))^2 p(x, y)\mu(x).$$

A Poincaré inequality states that

$$\operatorname{Var} f \le c\, \mathcal{E}(f, f) \tag{22}$$

for all functions $f : V \to \mathbb{R}$. Recall that the discrete Laplacian is the matrix

$$L = P - I,$$

where I is the $V \times V$ identity matrix and $P = (p(x, y))_{x,y\ inV}$ is the transition matrix. The matrix L is the (discrete) generator of the random walk, in the sense that for any bounded function $f : V \to \mathbb{R}$,

$$f(X_n) - \sum_{i=0}^{n-1} Lf(X_i), \quad n = 0, 1, \ldots$$

is a martingale in the natural filtration of the random walk $(X_0, \ldots)$. By the Perron-Frobenius theorem, P has all its eigenvalues smaller or equal to 1, so we can order the eigenvalues of L in nondecreasing order $\lambda_0 = 0 < \lambda_1 \le \lambda_2 \ldots$.

Definition 2. *The number $\lambda_1 > 0$ is called the* spectral gap *of the random walk.*

The spectral gap comes up in a number of probabilistic and geometric problems. In probability, the spectral gap is intimately connected to asymptotics of the *mixing time* of the random walk, i.e., how long does it take for a random walk to reach its stationary distribution.

We start with an analogue of the Poincaré inequality described in Theorem 3.

Theorem 5. *There is the equality:*

$$\lambda_1 = \inf_{f \neq 0} \frac{\mathcal{E}(f,f)}{\operatorname{Var} f}$$

The proof is a copy of the continuum version of this result (Theorem 3). In particular, this gives us a (discrete) Poincaré inequality (22) satisfied with $c = 1/\lambda_1$.

We now prove a concentration inequality (closely related Theorem 2) which holds in complete generality for functions that are sufficiently 'smooth", i.e., Lipschitz in a suitable sense. For $f : V \to \mathbb{R}$, define

$$|||f|||_\infty^2 = \frac{1}{2} \sup_{x\in X} \sum_{y \in X} |f(y) - f(x)|^2 p(x,y) \tag{23}$$

Theorem 6. *Assume that (p,μ) is reversible on the finite graph G, and let $\lambda_1 > 0$ be the spectral gap. Then if $|||F|||_\infty < \infty$ we have:*

$$\mu\left(F > \int F d\mu + \varepsilon\right) \leq 3e^{-\varepsilon |||F|||_\infty \sqrt{\lambda_1}/2}. \tag{24}$$

Proof. While (24) can be proved along the same lines as Theorem 2, we propose a different approach which is more adapted to this case. Let

$$\Phi(\lambda) = \int e^{\lambda F} d\mu$$

be the Laplace transform of the random variable F under the distribution μ. We may assume by homogeneity that $|||F|||_\infty = 1$. By the Poincaré inequality applied to $e^{\lambda F/2}$ we get

$$\lambda_1 \operatorname{Var}(e^{\lambda F/2}) \leq \mathcal{E}(e^{\lambda F/2}, e^{\lambda F/2}),$$

that is,

$$\lambda_1(\Phi(\lambda) - \Phi(\lambda/2)^2) \leq \mathcal{E}(e^{\lambda F/2}, e^{\lambda F/2}).$$

On the other hand, if we compute the Dirichlet energy of $e^{\lambda F/2}$ we get

$$\begin{aligned}
\mathcal{E}(e^{\lambda F/2}, e^{\lambda F/2}) &= \sum_{F(x)<F(y)} (e^{\lambda F(x)/2} - e^{\lambda F(y)/2})^2 p(x,y)\mu(x) \quad \text{by symmetry} \\
&\leq \frac{\lambda^2}{2} \sum_{F(x)<F(y)} e^{\lambda F(y)} (F(y) - F(x))^2 p(x,y)\mu(x) \quad \text{using } 1 - e^{-x} \leq x \\
&= \frac{\lambda^2}{2} \sum_y e^{\lambda F(y)} \sum_x (F(y) - F(x))^2 \mathbf{1}_{\{F(x)<F(y)\}} p(y,x)\mu(y) \quad \text{by (47)} \\
&\leq \lambda^2 |||F|||_\infty^2 \sum_y e^{\lambda F(y)} \mu(y) = \lambda^2 \Phi(\lambda).
\end{aligned}$$

Thus we get

$$\lambda_1(\Phi(\lambda) - \Phi(\lambda/2)^2) \leq \lambda^2\Phi(\lambda).$$

Rephrasing this, we get the inequality:

$$\Phi(\lambda) \leq \frac{1}{1-\lambda^2/\lambda_1}\Phi(\lambda/2)^2,$$

for every $\lambda < \sqrt{\lambda_1}$. This can now be iterated, yielding:

$$\Phi(\lambda) \leq \prod_{k=0}^{n-1}\left(\frac{1}{1-\lambda^2/(4^k\lambda_1)}\right)^{2^k}\Phi(\lambda/2^n)^{2^n}$$

Note that $\Phi(\lambda) = 1 + \lambda\mathbb{E}(F) + o(\lambda)$ as $\lambda \to 0$ (by the Lebesgue convergence theorem), and we may assume without loss of generality that $\mathbb{E}(F) = 0$. Thus letting $n \to \infty$:

$$\Phi(\lambda) \leq \prod_{k=0}^{\infty}\left(\frac{1}{1-\lambda^2/(4^k\lambda_1)}\right)^{2^k}$$

and note that this infinite product converges for every $\lambda > 0$ (but we have only show that the above inequality holds if $\lambda < \sqrt{\lambda_1}$). If we also set $\lambda = \sqrt{\lambda_1}/2$ then the right-hand side is a universal constant slightly smaller than 3 (slightly bigger than e) and we find

$$\Phi(\sqrt{\lambda_1}/2) \leq 3.$$

The proof of Theorem 2 is now an easy application of Markov's inequality:

$$\begin{aligned}\mu\left(F > \varepsilon\right) &\leq \mu\left(e^{\lambda F} > e^{\varepsilon}\right)\\ &\leq e^{-\lambda\varepsilon}\Phi(\lambda) = 3e^{-\varepsilon\sqrt{\lambda_1}/2}\end{aligned}$$

as desired. The proof extends immediately to arbitrary function F with $|||F|||_\infty < \infty$. □

We now mention a couple of useful and rather unexpected consequences of this Theorem, which relates the spectral gap of the random walk to the diameter of the graph. Let $G = (V, E)$ be an undirected finite graph and let μ be the equilibrium measure of simple random walk on G: thus, $\mu(x) = \alpha/\deg(x)$, where $\alpha > 0$ is a normalising constant. We say that μ is *nearly constant* if there exists $C > 0$ such that

$$\mu(x) \leq C\min_{y\in V}\mu(y).$$

In particular this is guaranteed to happen if the graph G is regular (in which case we can take $C = 1$).

Theorem 7. *Let λ_1 be the spectral gap of the random walk, and let δ be the diameter of the graph (maximal distance between two vertices). Then*

$$\lambda_1 \leq \left(\frac{8 \log(12C \text{ Card } V)}{\sqrt{2}\delta} \right)^2 .$$

Proof. We first show how to extend Theorem 6 to a concentration away from the median. Let

$$D(x,y) = \sup_{\|f\|_\infty \leq 1} [f(y) - f(x)]$$

Then for any set A such that $\mu(A) \geq 1/2$ we have, denoting A_ε the D-enlargement of size ε of A,

$$\mu(A_\varepsilon) \geq 1 - 3\exp(-\varepsilon\sqrt{\lambda_1}/4). \tag{25}$$

Indeed, $F(x) = \min(r, D(x,A))$ satisfies $|||f|||_\infty \leq 1$ so Theorem 6 applies to it. On the other hand,

$$\int F d\mu \leq (1 - \mu(A))r$$

thus

$$\begin{aligned} \mu(A_r^c) &\leq \mu(F > r) \\ &\leq \mu\left(F > \int F d\mu + r\mu(A)\right) \\ &\leq \mu\left(F > \int F d\mu + r/2\right) \leq 3\exp(-r\sqrt{\lambda_1}/4) \end{aligned}$$

by Theorem 6. This proves (25). In particular, if F is 1-Lipschitz for the distance D, the same reasoning as for the sphere shows that

$$\mu(|F - m_F| > r) \leq 6\exp(-r\sqrt{\lambda_1}/4). \tag{26}$$

Now, fix $a, b \in V$, and let $2r = D(a,b)$. Consider the function $F(y) = D(y,b)$, and let m be a median for F. Note that since F is 1-Lipschitz for the distance D, we have:

$$\mu(a)\mu(b) \leq \mu \otimes \mu\{(x,y) : |F(x) - F(y)| \geq 2r\}.$$

Decomposing on the event where $|F(x)-m| \leq r$, $|F(y)-m| \leq r$, or $|F(x)-m| > r$ and $|F(y) - m| > r$, we get to the upper-bound:

$$\mu(a)\mu(b) \leq 2\mu(|F - m| > r) \leq 12\exp(-r\sqrt{\lambda_1}/4),$$

using (26). On the other hand, note that $D(a,b) \geq d(a,b)\sqrt{2}$. Indeed, for every

$$\begin{aligned}|||f|||_\infty^2 &= \frac{1}{2} \sup_{x\in V} \sum_{y \in V} (f(y) - f(x))^2 p(x,y) \\ &\leq \frac{1}{2} \sup\{|f(y) - f(x)|; x \sim y\}^2 =: \|\nabla f\|_\infty^2\end{aligned}$$

and thus

$$D(x,y) = \sup_{\|f\|_\infty \leq 1} [f(y) - f(x)] \geq \sup_{\|\nabla f\|_\infty^2 \leq 2} [f(y) - f(x)] = \sqrt{2} d(x,y).$$

Thus

$$\mu(a)\mu(b) \leq 2\mu(|F - m| > r) \leq 12 \exp(-\sqrt{2} d(a,b)\sqrt{\lambda_1}/8).$$

Since μ is almost constant, note however that $\mu(x) \geq 1/(C \quad \text{Card } V)$. Thus taking a, b a pair of vertices which realise the diameter, gives us the result. □

The spectral gap of a random walk is an important quantity, and contains much information about the properties of the random walk. For instance, its inverse $R = 1/\lambda_1$ is called the relaxation time, and is closely related to the *mixing time* of the random walk. Intuitively speaking, this is the time it takes for the process to reach its equilibrium distribution. Equivalently, if we think of heat diffusion on the graph, starting from a situation where a particular vertex has temperature 1 and every other vertex has temperature 0, this is the time it takes for the temperature distribution to take its stationary values.

Thus the spectral gap gives information about how long to run a particular Markov chain to get a stationary sample. As such, the relaxation is a quantity of fundamental importance in statistics or computer science: the MCMC is one of the 10 most widely used algorithms in the world. The upper-bound from Theorem 7 gives a lower bound on the relaxation time R. In practice this will rarely be sharp, but it maybe useful to give theoretical polynomial or logarithmic lower bounds on the running time of an MCMC algorithm.

3 Logarithmic Sobolev Inequalities

3.1 Introduction and definition

Log-Sobolev inequalities, introduced by Leonard Gross in 1975 [22], are one of the essential tools for proving concentration phenomena, not only because they require in some sense less understanding about the underlying geometry of the measured space, but also because they yield sharper results for concentration, i.e., Gaussian rather than exponential. They are particularly well-suited for infinite-dimensional analysis.

We start by recalling the classical Sobolev embedding for functions in $\mathbb{R}^n$. Let Ω be a bounded open domain in $\mathbb{R}^n$ with sufficiently regular boundary (Lipschitz boundary, or satisfying the 'cone' condition for Brownian motion). Let $W^{k,p}(\Omega)$ be the Sobolev space of order (k, p) where $k \geq 1$ and $p \in [1, \infty)$: this is the set of functions for which all k (weak) derivatives are in $L^p(\Omega)$. In particular, $W^{1,2}$ is denoted by H^1 is a Hilbert space. Then the Sobolev embedding states the following:

Theorem 8. *Let $f \in W^{k,p}(\Omega)$. Then $f \in W^{\ell,q}(\Omega)$ for any ℓ, q such that $\ell \leq k$ and*

$$\frac{1}{p} < \frac{1}{q} + \frac{k-\ell}{n}. \tag{27}$$

Moreover, this embedding is continuous*: that is,*

$$\|f\|_{\ell,q} \leq c\|f\|_{k,p} \tag{28}$$

where the Sobolev norm $\|f\|_{k,p}$ is given by the sum of the L^p norms of all derivatives of order k.

In the case where $k = 1$ and $p = 2$, which is the case we are in fact really interested in, we have already proved this result (it is the Nirenberg-Sobolev inequality). Let us make a few important comments on this result. First, this result tells us that if $f \in W^{k,p}$ for some $k \geq 1$ and $p \in [1, \infty)$, then its derivatives of order $\ell < k$ are in fact not only in L^p but also in L^q where q is defined by (27), and note that $q > p$. In the case $k = 1$, $p = 2$, this says

$$\|f\|_{2+\varepsilon_n} \leq c_n\|\nabla f\|_2$$

where $\varepsilon_n > 0$ and $\varepsilon_n \sim 4/n$ as $n \to \infty$. Thus lower-order derivatives are more integrable. Interestingly, *how much* more integrable depends on the dimension n, as shown by (27). Note that as $n \to \infty$ (in order to work in dimension-free setup) we have $1/q \to 1/p$, and thus the Sobolev gain appears to become negligible in the limit. In fact things are more subtle, as we will see that if $k = 1$ and $p = 2$, we can still get a *logarithmic* additional control when $n \to \infty$: under the Gaussian measure, we will see with the Gaussian Log-Sobolev inequality (Theorem 30)

$$\int f^2 \log(f^2) d\mu \leq c \int (\nabla f)^2 d\mu \tag{29}$$

for some $c > 0$, provided the right normalisation is chosen ($\mathbb{E}(f^2) = 1$). Interestingly, this logarithmic gain compared to the Poincaré inequality has huge consequences for the concentrative properties. In particular we will see that this in general leads to Gaussian concentration (Theorem 9).

We will essentially take (29) as our definition of a Log-Sobolev inequality. Given the $f \log f$ term, there is a natural probabilistic interpretation in terms of the entropy of the random variable f. This leads us to the following definition.

Definition 3. *Let (X, g) be a Riemaniann manifold, equipped with a probability measure μ. We say that μ satisfies a Log-Sobolev inequality with constant $C > 0$ if for every smooth function f such that the terms below are well-defined:*

$$\mathrm{Ent}_\mu(f^2) \leq 2C \int |\nabla f|^2 d\mu. \tag{30}$$

Here $\mathrm{Ent}(f)$ *denotes the entropy of the function f:*

$$\mathrm{Ent}_\mu(f) = \mathbb{E}(f \log f) - \mathbb{E}(f) \log \mathbb{E}(f).$$

More generally, this definition makes sense on any metric probability measure space (X, d, μ), where $|\nabla f|$ denotes the generalized gradient:

$$|\nabla f(x)| = \limsup_{y \to x} \frac{|f(y) - f(x)|}{d(x, y)},$$

Note that the entropy of a random variable (function) is always nonnegative by Jensen's inequality (since the function $x \mapsto x \log x$ is convex) and is defined as soon as $\mathbb{E}(X \log^+(1 + X)) < \infty$. It is also homogeneous of degree 1 (as a norm should be).

A first, unsurprising result, is that a Log-Sobolev inequality is always stronger than a Poincaré inequality.

Proposition 3. *Assume that μ satisfies a Log-Sobolev inequality (30) for all smooth functions f with constant $C > 0$. Then for all bounded smooth functions f*

$$\mathrm{Var}_\mu(f) \leq C \int |\nabla f|^2 d\mu,$$

i.e., the Poincaré inequality is satisfied.

Proof. The proof is simple and relies on a Taylor expansion of the Log-Sobolev inequality (30) applied to $1+\varepsilon f$, where f is any bounded function with zero mean. Indeed, note first that $\nabla(1+\varepsilon f) = \varepsilon \nabla f$ so the right-hand side of (30) is $2C\varepsilon^2 \int |\nabla f|^2 d\mu$. The left-hand side, on the other hand, has the following asymptotics as $\varepsilon \to 0$. First, using the probabilistic notations $\mathbb{P}$ and X instead of μ and f, by the Lebesgue convergence theorem

$$\begin{aligned}\mathbb{E}[(1+\varepsilon X)^2 \log((1+\varepsilon X)^2)] &= 2\mathbb{E}[(1+2\varepsilon X)(\varepsilon X - \varepsilon^2 X^2/2)] + o(\varepsilon^2)\\ &= 2\varepsilon \mathbb{E}(X) + 4\varepsilon^2 \mathbb{E}(X^2) - \varepsilon^2 \mathbb{E}(X^2) + o(\varepsilon^2)\\ &= 3\varepsilon^2 \mathbb{E}(X^2) + o(\varepsilon^2) \quad (\text{since } \mathbb{E}(X) = 0).\end{aligned}$$

Also,

$$\begin{aligned}\mathbb{E}[(1+\varepsilon X)^2] \log \mathbb{E}[(1+\varepsilon X)^2] &= (1+\varepsilon^2 \mathbb{E}(X^2)) \log(1+\varepsilon^2 \mathbb{E}(X^2))\\ &= (1+\varepsilon^2 \mathbb{E}(X^2))(\varepsilon^2 \mathbb{E}(X^2)) + o(\varepsilon^2)\\ &= \varepsilon^2 \mathbb{E}(X^2) + o(\varepsilon^2)\end{aligned}$$

so that

$$\mathrm{Ent}_{\mathbb{P}}[(1+\varepsilon X)^2] = 2\varepsilon^2 \mathbb{E}(X^2) + o(\varepsilon^2).$$

From this we conclude (using (30)): for all $\varepsilon > 0$

$$2\varepsilon^2 \mathbb{E}(X^2) + o(\varepsilon^2) \le 2C\varepsilon^2 \int |\nabla f|^2 d\mu$$

so that necessarily

$$\mathrm{Var}_\mu f = \mathbb{E}(X^2) \le C \int |\nabla f|^2 d\mu.$$

Since this inequality is unchanged by the addition of a constant to f, it holds for any smooth function f with compact support, thereby proving Poincaré's inequality. □

3.2 Log-Sobolev implies Gaussian concentration

We now state the result which explains our interest in log-Sobolev inequalities. Let (X, d, μ) be a general measured metric space. If $\phi : (X, d) \to (Y, |\cdot|)$ then we may define the (generalised) gradient length of ϕ at the point x to be:

$$|\nabla \phi(x)| = \limsup_{y \to x} \frac{|\phi(y) - \phi(x)|}{d(x,y)},$$

which matches the usual definition when $X = \mathbb{R}^n$ and $Y = \mathbb{R}$. The important property that is used in this proof is that the generalized gradient satisfies the following chain rule: if $f : X \to \mathbb{R}$ and $\phi : \mathbb{R} \to \mathbb{R}$ is smooth, then

$$|\nabla \phi(f)| \le |\phi'(f)||\nabla(f)|$$

pointwise. Recall that if a function is K-Lipschitz

Theorem 9. *Let (X, d, μ) be a metric measure space in which the Log-Sobolev inequality (30) holds, where $|\nabla f|$ is interpreted as above. Then every $K-$Lipschitz function is integrable and if $F : X \to \mathbb{R}$ is such a function, we have:*

$$\mu\left\{x : F(x) \geq \int F d\mu + r\right\} \leq \exp\left(-\frac{r^2}{2CK^2}\right). \tag{31}$$

Proof. We start by proving the result for F bounded. The following argument is due to Herbst and is described in Theorem 5.3 from Ledoux [31]. Let $\Phi(\lambda)$ denote the Laplace transform of F:

$$\Phi(\lambda) = \int e^{\lambda F} d\mu.$$

Our goal will be to show that if $\int F d\mu = 0$, and F is bounded,

$$\Phi(\lambda) \leq e^{CK^2\lambda^2/2}, \tag{32}$$

and the result then follows from Markov's inequality and optimising in λ: indeed for every $\lambda > 0$: assuming (32) then we have

$$\mathbb{P}(F > r) \leq e^{-\lambda r}\Phi(\lambda) \leq \exp(-\lambda r + CK^2\lambda^2/2).$$

The term in the exponent is optimal when $-r + CK^2\lambda = 0$, i.e. if $\lambda = r/(CK^2)$, in which case we obtain:

$$\mathbb{P}(F > r) \leq \exp(-r^2/(2CK^2)),$$

as claimed (the general, not necessarily bounded case is then an easy consequence of the monotone convergence theorem). Thus it suffices to prove (32). To do this, assume by homogeneity that $K = 1$ and consider the function

$$G(\lambda) = \mathbb{E}(e^{\lambda F - C\lambda^2/2}).$$

If f is defined as $f^2 = e^{\lambda F - C\lambda^2/2}$ so that $G(\lambda) = \mathbb{E}(f^2)$, then by the Log-Sobolev inequality

$$\mathrm{Ent}_\mu(f^2) \leq 2C \int |\nabla f|^2 d\mu. \tag{33}$$

Now, note that by definition of the entropy,

$$\mathrm{Ent}_\mu(f^2) = \mathbb{E}(e^{\lambda F - C\lambda^2/2}(\lambda F - C\lambda^2/2)) - G(\lambda)\log G(\lambda)$$

Let us now get an upper-bound on the the right-hand side of (33). Note first that by the genearlized chain rule,

$$\begin{aligned}\int |\nabla f|^2 d\mu &\leq \frac{\lambda^2}{4}\int |\nabla F|^2 e^{\lambda F - C\lambda^2/2} d\mu \\ &\leq \frac{\lambda^2}{4}\int e^{\lambda F - C\lambda^2/2} d\mu = \lambda^2 G(\lambda)/4.\end{aligned}$$

since $|\nabla F| \leq K$ almost everywhere, by Rademacher's theorem, and $K = 1$ by assumption.Thus, using (33) and our lower bound on the entropy, we get

$$\mathbb{E}(e^{\lambda F - C\lambda^2/2}(\lambda F - C\lambda^2/2)) - G(\lambda)\log G(\lambda) \leq C\lambda^2 G(\lambda)/2$$

and thus

$$\mathbb{E}(e^{\lambda F - C\lambda^2}(\lambda F - C\lambda^2)) - G(\lambda)\log G(\lambda) \leq 0$$

but note that the first term in the left hand side is now simply equal to $\lambda G'(\lambda)$, hence:

$$\lambda G'(\lambda) - G(\lambda)\log G(\lambda) \leq 0.$$

Equivalently, if

$$H(\lambda) = \frac{\log G(\lambda)}{\lambda}; \quad \lambda > 0;$$

and $H(0) = G'(0)/G(0) = \mathbb{E}(F) = 0$. Then we get

$$H'(\lambda) = \frac{1}{\lambda^2 G(\lambda)}[\lambda G'(\lambda) - G(\lambda)\log G(\lambda)] \leq 0$$

and thus $H(\lambda) \leq H(0) = 0$ for all $\lambda \geq 0$. In particular, $G(\lambda) \leq 1$ for all $\lambda \geq 0$. That is,

$$\Phi(\lambda) \leq e^{CK^2\lambda^2/2}, \tag{34}$$

which proves (32). □

Remark 3. *Below we will prove that the infinite-dimensional Gaussian measure satisfies a Log-Sobolev inequality. In fact a similar argument may be used to prove that any compact manifold X has a Log-Sobolev inequality. Furthermore, it can be shown that if*

$$\rho = \inf_{f \neq 0} \frac{\int |\nabla f|^2 d\mu}{\mathrm{Ent}_\mu(f^2)}$$

(so that a Log-Sobolev inequality holds if and only if $\rho > 0$), and if n is the dimension of the manifold X then

$$\lambda_1 \geq \rho \geq \frac{n\gamma}{n-1}$$

where γ is any lower-bound on the Ricci curvature of X.

4 Concentration of Gaussian Measures and Processes

In this section we shall investigate several ways in which Gaussian measures enjoy dimension-free concentration phenomena. There are several mathematical approaches to establish this (geometric, probabilistic, analytic), and we take an analytic approach via Ornstein-Uhlenbeck semigroups with some probabilistic elements coming from Gaussian processes.

First some definitions: Let T be any (nonempty) set and let $(\Omega, \mathcal{A}, \Pr)$ be a probability space. Then a Gaussian process G is a mapping $G : T \times (\Omega, \mathcal{A}, \Pr) \to \mathbb{R}$ such that for any *finite* set of points $(t_1, ..., t_k)$, the vector $(G(t_1), ..., G(t_k))$ has a multivariate normal distribution. We shall say that G is *centered* if $EG(t) = 0$ for every $t \in T$. All standard Gaussian processes (such as the usual variants of Brownian motion and sheets, Brownian bridges, etc.) as well as Gaussian measures in normed linear spaces B (by simple duality arguments, see Subsection 4.4) can be accommodated in this framework.

As we shall see a unified way to establish dimension-free measure concentration for Gaussian measures is through studying the suprema of Gaussian processes, that is by studying the random variable $\sup_{t \in T} G(t)$ (or alternatively, $\sup_{t \in T} |G(t)|$). To ensure that $\sup_{t \in T} G(t)$ is well defined, let us assume that T is countable. [Note that otherwise this supremum is not necessarily a proper random variable. Usually G will have continuous sample paths and then the supremum can a fortiori be realized as one over a countable set, so that this is not really a restriction.]

Since $E \sup_t G(t)$ is generally not zero we can expect concentration of $\sup_t G(t)$ only around its mean $E \sup_t G(t)$ (or, alternatively, around its median). Indeed we shall prove the following fundamental result, which is due independently to Borell [5] and Sudakov & Cirelson. Note that it holds with 'minimal structure' required for the Gaussian process (i.e., no 'stationarity' requirement, no assumption on the increments, no sample-continuity etc.). The only condition is that the maximum of the process exists almost surely (which is necessary to even formulate the result).

Theorem 10 (Borell's inequality)**.** *Let $G(t), t \in T$, be a centered Gaussian process indexed by the countable set T, and such that $\sup_{t \in T} G(t) < \infty$ almost surely. Then $E \sup_{t \in T} G(t) < \infty$, and for every $r \geq 0$ we have*

$$\Pr \left\{ \left| \sup_{t \in T} G(t) - E \sup_{t \in T} G(t) \right| \geq r \right\} \leq 2e^{-r^2/2\sigma^2}$$

where $\sigma^2 = \sup_{t \in T} E(G^2(t)) < \infty$.

We note that the same inequality holds if $G(t)$ is replaced by $|G(t)|$ everywhere in the theorem.

Before we prove Theorem 10 below let us add some comments. The quantity σ^2 is sometimes called the 'weak' variance of G, as the supremum over T is outside of the expectation of the second moment of $G(t)$. Indeed one of the hidden strengths of this result is that the dependence on the variance of the process G is simply through σ^2: Think of the simplest case where $T = 1, ..., n$ and where the $X(t)'s$ are i.i.d. centered normal: then $E \max_{1 \le t \le n} G^2(t)$ grows logarithmically in n (being the maximum of n i.i.d random variables) whereas $\max_{1 \le t \le n} EX^2(t)$ is bounded by a fixed constant!

We should also emphasize that Theorem 10 does not give any information about the size of $E \sup_t G_t$, it just says that $\sup_t G_t$ concentrates around its expectation. If one is interested in estimating the size of $E \sup_t G_t$ one can use Dudley's entropy integral [12] and refinements using 'generic chaining' [39], but this is not the content of these notes.

4.1 The Ornstein-Uhlenbeck Semigroup

We shall prove Theorem 10 in Subsection 4.3 by applying a powerful logarithmic Sobolev inequality for Gaussian measures on $\mathbb{R}^n$. To establish this log-Sobolev inequality we need some simple analytic tools from semigroup theory that are summarized in what follows. We refer to Section 1.4 in [4] for an excellent reference.

Let $\gamma = \gamma_n$ be the canonical Gaussian measure on $\mathbb{R}^n$ and denote by $L^p(\gamma_n)$ the space of p-fold γ-integrable functions on $\mathbb{R}^n$. The Ornstein-Uhlenbeck semigroup $(P_t)_{t \ge 0}$ is the family of integral operators defined by

$$P_t(h)(x) = \int_{\mathbb{R}^n} h(e^{-t}x + (1 - e^{-2t})^{1/2} y) d\gamma(y), \quad t \in [0, \infty], x \in \mathbb{R}^n.$$

There are other (equivalent) ways to define these operators, the one used here is also know as the 'Mehler formula'. Clearly whenever h is a bounded (measurable) function these operators are well-defined. More generally, recall that γ equals the image of the measure $\gamma \otimes \gamma$ on $\mathbb{R}^n \times \mathbb{R}^n$ under the mapping $(x, y) \mapsto e^{-t}x + \sqrt{1 - e^{-2t}}y$ (write $e^{-t}x + \sqrt{1 - e^{-2t}}y = x \sin \theta(t) + y \cos \theta(t)$ for $\theta(t) = \arcsin(e^{-t})$ and use that for X, Y i.i.d. centered Gaussian vectors and every θ the random variable $X \sin \theta + Y \cos \theta$ has the same distribution as X). Therefore

$$\int |h(x)| d\gamma(x) = \int \int |h(e^{-t}x + \sqrt{1 - e^{-2t}}y)| d\gamma(x) d\gamma(y)$$

so that Fubini's theorem implies that $P_t(h)$ exists and is γ-integrable whenever h is. The same holds for $L^1(\gamma)$ replaced by $L^p(\gamma)$. Similarly we obtain for $f \in L^1(\gamma)$ that

$$\int P_t(h)(x) d\gamma(x) = \int \int h(e^{-t}x + \sqrt{1 - e^{-2t}}y) d\gamma(x) d\gamma(y) = \int h(x) d\gamma(x),$$

or in other words

$$P_\infty(P_t(h)) = P_\infty(h) \tag{35}$$

so that γ is a (in fact the unique) *invariant measure* for family of operators $(P_t)_{t\geq 0}$. From the above and dominated convergence and a simple approximation argument we also obtain, for $f \in L^p(\gamma)$

$$\lim_{t\to\infty} P_t(f) = P_\infty(f) \quad in\ L^p(\gamma). \tag{36}$$

On the other 'end' of the parameterization we have, since γ is a probability measure,

$$P_0(h)(x) = h(x). \tag{37}$$

Using furthermore the change of variables

$$(y, z) \mapsto e^{-s}\frac{\sqrt{1-e^{-2t}}}{\sqrt{1-e^{-2t-2s}}}y + \frac{\sqrt{1-e^{-2s}}}{\sqrt{1-e^{-2t-2s}}}z$$

which also realizes γ as the image of $\gamma \otimes \gamma$ one has

$$\begin{aligned} P_t(P_s h)(x) &= \int P_s h\left(e^{-t}x + \sqrt{1-e^{-2t}}y\right) d\gamma(y) \\ &= \int\int h\left(e^{-s}e^{-t}x + e^{-s}\sqrt{1-e^{-2t}}y + \sqrt{1-e^{-2s}}z\right) d\gamma(z)d\gamma(y) \\ &= \int h\left(e^{-t-s}x + \sqrt{1-e^{-2t-2s}}w\right) d\gamma(w) \end{aligned}$$

so that

$$P_t(P_s h)(x) = P_{t+s}h(x), \tag{38}$$

which implies that $(P_t)_{t\geq 0}$ has the properties of a semigroup. The (infinitesimal) generator of the semigroup is the operator defined by

$$Lh := \lim_{s\to 0}\frac{P_s(h) - h}{s} \quad (in\ L^2(\gamma)) \tag{39}$$

for $h \in L^2(\gamma)$ for which this limit exists. This is equivalent to

$$LP_t(f) = \frac{d}{dt}P_t(f) \tag{40}$$

(apply (39) to $h = P_t(h)$ and use (38) for one direction and recall $P_0(h) = h$ for the other).

The following Proposition gives a useful characterization of L in terms of a second differential operator and an integration by parts formula for this operator (which is called the Ornstein-Uhlenbeck operator). We define a Gaussian Sobolev space: Let $W_2^1(\mathbb{R}^n, \gamma)$ be the completion of the space of infinitely-differentiable functions with compact support in $\mathbb{R}^n$ with respect to the norm

$$\|f\|_{1,2}^2 := \int_{\mathbb{R}^n} f^2(x) d\gamma(x) + \int_{\mathbb{R}^n} |\nabla f(x)|^2 d\gamma(x).$$

Say $f \in W_2^2(\mathbb{R}^n, \gamma)$ if $f, Df \in W_2^1(\mathbb{R}^n, \gamma)$.

Proposition 4. *Denote by Δ the Laplace operator on $\mathbb{R}^n$ and by ∇ the gradient operator. Let L be the infinitesimal generator (39) of the Ornstein-Uhlenbeck semigroup. Then we have $L = \Delta - x\nabla$ and the integration by parts formula*

$$-\int fL(g) d\gamma = \int \nabla f \cdot \nabla g d\gamma$$

holds true for every $f \in W_2^1(\mathbb{R}^n, \gamma)$, $g \in W_2^2(\mathbb{R}^n, \gamma)$.

Proof. The identity $L = \Delta - x\nabla$ has several proofs: There is for instance a probabilistic one using Ito's formula and martingales, a funtional-analytic one using Hermite polynomials as orthonormal bases for $L^2(\gamma)$, and one that simply uses Taylor expansions. We shall give here a rather pedestrian calculus proof for 'smooth enough' f (the general result following from approximation arguments): clearly 'machinery' can give faster proofs here but we prefer to stay self-contained. In view of (40) it is sufficient to establish that

$$\frac{d}{dt} P_t(f) = (\Delta - x\nabla)(P_t(f))$$

holds. Define

$$P_\theta(f)(x) = \int f(x \sin\theta + y \cos\theta) d\gamma(y)$$

and furthermore the mapping $\theta(t) = \arcsin(e^{-t})$ so that $P_{\theta(t)}(f) = P_t(f)$ and

$$\frac{d}{dt} P_t(f) = \frac{d}{dt} P_{\theta(t)}(f) = \frac{d}{d\theta} P_\theta(f) \frac{d\theta}{dt} = \frac{d}{d\theta} P_\theta(f) \left(-\frac{\sin\theta}{\cos\theta} \right) \Bigg|_{\theta=\theta(t)} \tag{41}$$

by the chain rule. Denote by $\langle \cdot, \cdot \rangle$ the regular inner product in $\mathbb{R}^n$. Then

$$\begin{aligned} \frac{d}{d\theta} P_\theta(f) &= \int \frac{d}{d\theta} f(x \sin\theta + y \cos\theta) d\gamma(y) \\ &= \int \langle \nabla f(x \sin\theta + y \cos\theta), x \cos\theta - y \sin\theta \rangle d\gamma(y) \end{aligned}$$

$$
\begin{aligned}
&= \cos\theta \int \langle \nabla f(x\sin\theta + y\cos\theta), x\rangle d\gamma(y) \\
&\quad - \sin\theta \int \langle \nabla f(x\sin\theta + y\cos\theta), y\rangle \, d\gamma(y) \\
&= \frac{\cos\theta}{\sin\theta} \langle \nabla \int f(x\sin\theta + y\sin\theta) d\gamma(y), x\rangle \\
&\quad - \frac{\cos\theta}{\sin\theta} \Delta \int f(x\sin\theta + y\cos\theta) d\gamma(y) \\
&= \frac{\cos\theta}{\sin\theta} \left(\langle \nabla P_\theta f, x\rangle - \Delta P_\theta(f) \right) \qquad (42)
\end{aligned}
$$

where we have used, in the last but one step, that

$$
\nabla_x f(x\sin\theta + y\cos\theta) = \frac{\nabla_x \left(f(x\sin\theta + y\cos\theta) \right)}{\sin\theta}
$$

as well as integration by parts for the i-th coordinates

$$
\begin{aligned}
&\int_{\mathbb{R}^{n-1}} \int_{-\infty}^{\infty} \frac{d}{d_i} f(x\sin\theta + y\cos\theta) y_i \frac{e^{-\sum_{i=1}^n y_i^2/2}}{(2\pi)^{n/2}} dy \\
&= \cos\theta \int_{\mathbb{R}^n} \frac{d^2}{(d_i)^2} f(x\sin\theta + y\cos\theta) d\gamma(y) \qquad (43) \\
&= \frac{\cos\theta}{(\sin\theta)^2} \frac{d^2}{(d_i)^2} \int_{\mathbb{R}^n} f(x\sin\theta + y\cos\theta) d\gamma(y).
\end{aligned}
$$

Plugging the last expression in (42) into (41) gives

$$
\frac{d}{dt} P_t(f) = -\left(\langle \nabla P_\theta f, x\rangle - \Delta P_\theta(f) \right) \Bigg|_{\theta=\theta(t)} = (\Delta - x\nabla)(P_t(f))
$$

which completes the proof of the first claim. Given the identity $L = \Delta - x\nabla$ the second claim of the proposition now follows immediately from integration by parts on $\int \nabla f \cdot \nabla g d\gamma$. □

4.2 The logarithmic Sobolev inequality for Gaussian Measures in $\mathbb{R}^n$

We will now prove that the canonical Gaussian measure $\gamma := \gamma_n$ on $\mathbb{R}^n$ satisfies the log-Sobolev inequality condition of Theorem 9. This in turn implies that arbitrary Lipschitz-maps of Gaussian measures concentrate in a dimension-free way, and will in particular be seen to imply Theorem 10.

The following theorem is originally due to Gross [22] (it is sometimes called ”Gross’ logarithmic Sobolev inequality”).

Theorem 11. *Let γ be the canonical Gaussian measure on $\mathbb{R}^n$ with mean vector zero and covariance matrix equal to the identity. Then for every $f \in W_2^1(\mathbb{R}^n, \gamma)$ we have*

$$Ent_\gamma(f^2) \leq 2 \int |\nabla f|^2 d\gamma.$$

Proof. We shall assume first that f is infinitely differentiable, bounded and satisfies $\inf_x f(x) \geq c > 0$ to simplify technicalities (in particular so that we can interchange differentiation and integration below as we wish). We comment on the general case at the end of the proof.

If we set $h = f^2$ then it is in fact sufficient to establish

$$Ent_\gamma(h) \leq \frac{1}{2} \int \frac{|\nabla h|^2}{h} d\gamma \tag{44}$$

since

$$\frac{|\nabla(f^2)|^2}{f^2} = \frac{|2f\nabla f|^2}{f^2} = 4|\nabla f|^2.$$

We use the Ornstein-Uhlenbeck semigroup $(P_t)_{t\geq 0}$ introduced in the previous subsection. Define

$$E(t) = \int P_t(h) \log P_t(h) d\gamma$$

for which we have $E(0) = \int_{\mathbb{R}^n} h \log h d\gamma$ in view of (37) and

$$E(\infty) = \lim_{t\to\infty} \int_{\mathbb{R}^n} P_t(h) \log P_t(h) d\gamma = \int_{\mathbb{R}^n} h d\gamma \log \int_{\mathbb{R}^n} h d\gamma$$

in view of (36). For f bounded and infinitely differentiable $E(t)$ is continuously differentiable for every $t \geq 0$ by standard arguments involving dominated convergence so that

$$-\int_0^\infty \frac{d}{dt} E(t) dt = E(0) - E(\infty)$$

which implies

$$Ent_\gamma(h) = E(0) - E(\infty) = -\int_0^\infty \frac{d}{dt} \left[\int_{\mathbb{R}^n} P_t(h) \log P_t(h) d\gamma \right] dt. \tag{45}$$

Using (40) we have

$$\begin{aligned} \frac{d}{dt}(P_t(h) \log P_t(h)) &= \frac{d}{dt}(P_t(h)) \log P_t(h) + P_t(h) \frac{\frac{d}{dt} P_t(h)}{P_t(h)} \\ &= LP_t(h) \log P_t(h) + \frac{d}{dt} P_t(h) \end{aligned}$$

and interchanging differentiation and integration we arrive at

$$\frac{d}{dt}\left[\int_{\mathbb{R}^n} P_t(h)\log P_t(h)d\gamma\right] = \int_{\mathbb{R}^n} LP_t(h)\log P_t(h)d\gamma + \int_{\mathbb{R}^n} \frac{d}{dt}P_t(h)d\gamma.$$

To the first summand we apply integration by parts from Proposition 4 to obtain

$$\int LP_t(h)\log P_t(h)d\gamma = -\int \nabla P_t(h)\cdot\nabla(\log P_t(h))d\gamma = -\int \frac{|\nabla P_t(h)|^2}{P_t(h)}d\gamma$$

For the second summand we interchange integration and differentiation again to obtain from invariance of γ (i.e. (35)) that

$$\int \frac{d}{dt}P_t(h)d\gamma = \frac{d}{dt}P_\infty(P_t(h)) = \frac{d}{dt}P_\infty(h) = 0.$$

We conclude that the integrand in (45) equals

$$\frac{d}{dt}\int_{\mathbb{R}^n}(P_t(h)\log P_t(h))d\gamma = -\int_{\mathbb{R}^n}\frac{|\nabla P_t(h)|^2}{P_t(h)}d\gamma. \tag{46}$$

To proceed with the proof, note that the chain rule (and another interchange of differentiation and integration) clearly implies that

$$\nabla P_t(h) = e^{-t}P_t(\nabla h)$$

so that

$$|\nabla P_t(h)| \leq e^{-t}P_t(|\nabla h|) \tag{47}$$

Furhermore – writing shorthand $v(y) = e^{-t}(\cdot)+(1-e^{-2t})^{1/2}y$ – the Cauchy-Schwarz inequality implies

$$P_t(|\nabla h|)^2 = \int |\nabla h(v(y))|\sqrt{\frac{h(v(y))}{h(v(y))}}d\gamma(y) \leq P_t(h)P_t\left(\frac{|\nabla h|^2}{h}\right). \tag{48}$$

Combining (45), (46), (47) and (48) we conclude

$$\begin{aligned} Ent_\gamma(h) &= \int_0^\infty\int_{\mathbb{R}^n}\frac{|\nabla P_t(h)|^2}{P_t(h)}d\gamma dt \\ &\leq \int_0^\infty e^{-2t}\int\left(P_t\left(\frac{|\nabla h|^2}{h}\right)\right)d\gamma dt = \frac{1}{2}\int\frac{|\nabla h|^2}{h}d\gamma \end{aligned}$$

where we have used invariance of γ (i.e., (35)) in the last step. This establishes (44).

The result for $f \in W_2^1(\mathbb{R}^n,\gamma)$ satisfying $f \geq 0$ then follows from Fatou's lemma and since one can devise a sequence ϕ_i of bounded infinitely differentiable functions satisfying $\phi_i \geq 1/i$ s.t. $\phi_i \to f$ in $\|\cdot\|_{1,2}$ and almost everywhere. For an arbitrary function $f \in W_2^1(\mathbb{R}^n,\gamma)$ the result follows from the fact that $|f| \in W_2^1(\mathbb{R}^n)$ and $|\nabla|f|| = |\nabla f|$. □

We note that the same proof implies a log-Sobolev inequality with constant $2/c$ for log-concave measures $d\mu = e^{-U} d\lambda$ where $Hess(U) \geq cI$. In the proofs one uses, instead of the Ornstein-Uhlenbeck semigroup, the semigroup with infinitesimal generator $\Delta - \nabla U \cdot \nabla$.

4.3 Proof of Borell's Inequality

We are now in a position to apply the log-Sobolev inequality Theorem 9 to prove Theorem 10. Let $G(t), t \in T$, be a centered Gaussian process indexed by the countable set T, and such that $\sup_{t \in T} G(t) < \infty$ almost surely. Fix a finite set of points $t_1, ..., t_n$ in T and denote by $\mathcal{G} = (G(t_1), ..., G(t_n))$ the associated Gaussian random vector with positive semi-definite covariance $\Gamma = V'V$. Now if $\mathcal{N}$ is a random vector in $\mathbb{R}^n$ distributed according to the canonical Gaussian measure γ, then by the usual properties of normal random variables the distribution of $V\mathcal{N}$ is the same as the one of $\mathcal{G}$. Define the mapping

$$F(x) = \max_{1 \leq i \leq n} (Vx)_i, \quad x \in \mathbb{R}^n, \tag{49}$$

to which we will apply Theorem 9 with $X = \mathbb{R}^n$ and $\mu = \gamma$. To verify that F is Lipschitz from $\mathbb{R}^n$ to $\mathbb{R}$ we use the Cauchy-Schwarz inequality that

$$|(Vx)_i - (Vy)_i| = \left| \sum_j V_{i,j}(x_j - y_j) \right| \leq \sqrt{\sum_j V_{i,j}^2} |x - y|.$$

Since furthermore

$$\sum_j V_{i,j}^2 = Var(G_{t_i}) \leq \sigma^2$$

we see that the mapping F is Lipschitz with Lipschitz constant σ and hence satisfies the conditions of Theorem 9 with $K = \sigma$ which – combined with Theorem 11 (and thus $C = 1$) – gives that

$$\Pr \left(\max_{1 \leq i \leq n} G_{t_i} \geq E \max_{1 \leq i \leq n} G_{t_i} + r \right) \leq \exp \left\{ -\frac{r^2}{2\sigma^2} \right\}. \tag{50}$$

The same inequality applied to $-F$ gives

$$\Pr \left(\max_{1 \leq i \leq n} G_{t_i} \leq E \max_{1 \leq i \leq n} G_{t_i} - r \right) \leq \exp \left\{ -\frac{r^2}{2\sigma^2} \right\}. \tag{51}$$

This is already 'almost' Borell's inequality: it holds only for a finite maximum of points, but the fact that the last bound is completely dimension free already shows the 'infinite-dimensional nature' of the inequality.

To continue with the proof we show that $E\sup_t G_t$ is finite: choose r_0 large enough so that $1-e^{-r_0^2/2\sigma^2} > 1/2$ and m large enough so that $\Pr\{\sup_t G_t > m\} \leq 1/2$ (which is possible since $\sup_t G_t$ is finite almost surely). Then we have from (51)

$$\begin{aligned}
\frac{1}{2} &< 1-e^{-r_0^2/2\sigma^2} \leq \Pr\left\{\max_{1\leq i\leq n} G_{t_i} > E\max_{1\leq i\leq n} G_{t_i} - r_0\right\} \\
&= \Pr\left\{\left\{\max_{1\leq i\leq n} G_{t_i} > E\max_{1\leq i\leq n} G_{t_i} - r_0\right\} \cap \left\{\sup_t G_t \leq m\right\}\right\} \\
&\quad + \Pr\left\{\left\{\max_{1\leq i\leq n} G_{t_i} > E\max_{1\leq i\leq n} G_{t_i} - r_0\right\} \cap \left\{\sup_t G_t > m\right\}\right\} \\
&\leq \Pr\left\{E\max_{1\leq i\leq n} G_{t_i} \leq m + r_0\right\} + 1/2
\end{aligned}$$

so that

$$E\max_{1\leq i\leq n} G_{t_i} \leq m + r_0$$

(a 'nonrandom' event A that has positive probability 'happens with probability one' since $0 < \Pr(A) = \int 1_A d\Pr = 1_A \int d\Pr = 1_A = 1$). Since m and r_0 do not depend on n we can take a finite subset $T_n = \{t_1, ..., t_n\} \nearrow T$ such that $\max_{t\in T_n} G_t \nearrow \sup_{t\in T} G_t$ and then by monotone convergence also $E\max_{1\leq i\leq n} G_{t_i} \nearrow E\sup_t G_t$ to conclude

$$E\sup_{t\in T} G(t) < \infty. \tag{52}$$

Now to pass from (50) and (51) to Theorem 10 take $T_n \nearrow T$ as above so that, for every $\varepsilon > 0$

$$\begin{aligned}
\Pr\left\{\sup_{t\in T} G_t - E\sup_{t\in T} G_t > r+\varepsilon\right\} &\leq \liminf_n \Pr\left\{\max_{t\in T_n} G_t - E\sup_{t\in T} G_t > r+\varepsilon\right\} \\
&= \liminf_n \Pr\left\{\max_{t\in T_n} G_t - E\max_{t\in T_n} G_t > r\right\} \\
&\leq e^{-r^2/2\sigma^2}
\end{aligned}$$

using (50). Repeating this argument for the lower deviations (using (51)) completes the proof of Theorem 10.

4.4 Gaussian Measures in Banach Spaces

We shall show here how Theorem 10 implies results for general Banach-space valued Gaussian random variables. Let $(B, \|\cdot\|_B)$ be a (for simplicity) *separable* Banach space with norm $\|\cdot\|_B$, and let $X : (\Omega, \mathcal{A}, \Pr) \to B$ be a random variable (a measurable mapping where B is equipped with its Borel-σ-field). The law of X is called a centered Gaussian measure γ on B if $\gamma \circ L^{-1}$ is normally distributed with mean zero for every continuous linear functional $L : B \to \mathbb{R}$.

If B' is the dual space of B equipped with its operator norm $\|\cdot\|_{B'}$, then we can use the Hahn-Banach theorem to represent the norm of an element $x \in B$ by $\|x\|_B = \sup_{L:\|L\|_{B'} \le 1} |L(x)|$. For a Gaussian random variable we thus have

$$\|X\|_B = \sup_{L:\|L\|_{B'} \le 1} |L(X)| \tag{53}$$

and we can study the supremum of the Gaussian process $G_L := L(X), L \in T$ where $T = \{L : \|L\|_{B'} \le 1\}$ is the unit ball of B'. For every $x \in B$, the linear mapping $L \mapsto L(x)$ is weak-*-continuous on the weak-*-compact metric space T and thus the supremum (53) exists (everywhere and hence also almost surely). By continuity this supremum can be realized as one over a countable subset of T, which we shall also denote by T. Now Theorem 10 (and the observation that $L \in T$ implies $-L \in T$ by linearity so that $\sup_{L\in T} G_L = \sup_{L \in T} |G_L|$) proves that any Gaussian random variable in a Banach space has a finite first moment

$$E\|X\|_B < \infty$$

and satisfies the concentration inequality

$$\Pr\{|\|X\|_B - E\|X\|_B| \ge r\} \le 2e^{-r^2/2\sigma^2} \tag{54}$$

where $\sigma^2 = \sup_{L \in \mathcal{L}} EL(X)^2$ are the 'weak variances'. In other words, the norm of *every* Gaussian random variable in a separable Banach space concentrates around its mean with a perfect 'one-dimensional' Gaussian tail.

We should note that σ^2 is *always finite*: Since $\|X\|_B < \infty$ almost surely there exists M finite such that $\Pr\{\|X\|_B > M\} \le 1/2$. For any continuous $L \in T$ we thus have $\Pr\{|f(X)| > M\} \le 1/2$ where M does not depend on L. Let $\sigma^2(L) = EL^2(X)$ so that $L(X)/\sigma(L)$ is standard normal. Then we can take R small enough such that

$$\begin{aligned}
1/2 &< \Pr\{L(X)/\sigma(L) \ge R\} \\
&= \Pr\{L(X)/\sigma(L) \ge R, |L(X)| \le M\} + \Pr\{L(X)/\sigma(L) \ge R, |L(X)| > M\} \\
&\le \Pr\{\sigma(L) \le M/R\} + 1/2
\end{aligned}$$

so again the 'constant' event $\sigma(L) \le M/R$ has positive probability and thus probability one. Since $M/R < \infty$ does not depend on L we conclude that $\sup_{L \in T} EL^2(X) \le M/R < \infty$.

We should emphasize again that the estimate (54) does not convey any information about the size of $E\|X\|_B$ (other than that it is finite). If one needs more information about the size of $E\|X\|_B$ (in dependence of σ for instance), the answer depends on the geometry of the Banach space B, and can often be estimated by direct methods from probability in Banach spaces, see [1], [32]. Alternatively one can use moment inequalities for suprema of Gaussian processes, see, e.g., [12] and [39].

5 Log-Sobolev Inequalities in Product Spaces

5.1 Entropy in product spaces

The reason that Log-Sobolev inequalities are so convenient is that, in addition to their power to prove concentration results, they are fairly easy to establish for product measures once they are established for each individual space. This is essentially a consequence of the fact that "entropy behaves well with respect to independence". Furthermore, we will see that in doing so, the Log-sobolev constants are dimension-free, so it will suffice to prove a log-Sobolev inequality in one dimension, from which the infinite-dimensional case follows automatically.

We consider the following setup. Let $(X_1, d_1, \mu_1), \ldots (X_n, d_n, \mu_n)$ be metric probability measure spaces and let $X = X_1 \times \cdots \times X_n$ and denote the product measure by $\mu = \mu_1 \otimes \ldots \otimes \mu_n$. X is also endowed with a natural "ℓ^2" metric: that is, for $x, y \in X$, $d(x,y)^2 = \sum_{i=1}^n d_i(x_i, y_i)^2$. We assume that a Log-Sobolev inequality holds for every space X_i: that is, if $f : X_i \to \mathbb{R}$, then

$$\mathrm{Ent}_{\mu_i}(f^2) \leq 2C_i \int_{X_i} |\nabla f|^2 d\mu_i$$

for some $C_i > 0$, and where $|\nabla f|$ is the generalized gradient,

$$|\nabla f(x)| = \limsup_{y \to x} \frac{|f(x) - f(y)|}{d_i(x,y)}$$

for $x \in X_i$.

Theorem 12. *For any locally Lipschitz function $f : X \to \mathbb{R}$, we have*

$$\mathrm{Ent}_\mu(f^2) \leq 2\left(\max_{1 \leq i \leq n} C_i\right) \int_X |\nabla f|^2 d\mu$$

where $|\nabla f|^2 = \sum_{i=1}^n |\nabla_i f|^2$, i.e., the coordinatewise gradient where all variables except the ith one are frozen.

Proof. The first result is a "variational" characterization of entropy.

Lemma 3. *Let f be a nonnegative random variable on some probability space. Then*

$$\mathrm{Ent}_{\mathbb{P}}(f) = \sup\left\{\int fg d\mathbb{P}; \int e^g d\mathbb{P} \leq 1\right\}. \tag{55}$$

Proof. By homogeneity we may assume that $\mathbb{E}(f) = 1$, in which case $\mathrm{Ent}(f) = \mathbb{E}(f \log f)$. By Young's inequality:

$$uv \leq u \log u - u + e^v,$$

we get for $\int e^g d\mathbb{P} \leq 1$,

$$\int fgd\mathbb{P} \leq \mathrm{Ent}_{\mathbb{P}}(f) - 1 + \int e^g d\mathbb{P} \leq \mathrm{Ent}_{\mathbb{P}}(f).$$

Thus the supremum in the right-hand side of (55) is less or equal than $\mathrm{Ent}_{\mathbb{P}}(f)$. By considering $g = \log f$, we get the other direction. □

The second result is that the entropy is additive in the following sense. Given $x \in X$, let f_i denote the function on X_i defined by freezing all variables except the ith one. Thus

$$f_i(\cdot) = f(x_1, \ldots, x_{i-1}, \cdot, x_{i+1}, \ldots, x_n).$$

Lemma 4.

$$\mathrm{Ent}_\mu(f) \leq \sum_{i=1}^n \int_X \mathrm{Ent}_{\mu_i}(f_i) d\mu \tag{56}$$

Proof. This is Proposition 5.6 in Ledoux [31], and we follow his proof. Let g be a random variable such that $\int e^g d\mu \leq 1$. For $x_i, \ldots, x_n$ fixed and $1 \leq i \leq n$, then define

$$g^i(x_1, \ldots, x_n) = \log \frac{\int e^{g(y_1,\ldots,y_{i-1},x_i,\ldots,x_n)} d\mu_1 \ldots d\mu_{i-1}}{\int e^{g(y_1,\ldots,y_i,x_{i+1},\ldots,x_n)} d\mu_1 \ldots d\mu_i}$$

where the the numerator of the fraction is interpreted when $i = 1$ as $g(x_1, \ldots, x_n)$. Then by cascading the sum we get

$$\sum_{i=1}^n g^i(x_1, \ldots, x_n) = g(x_1, \ldots, x_n) - \log \int e^{g(y_1,\ldots,y_n)} d\mu_1 \ldots d\mu_n \geq g(x_1, \ldots, x_n)$$

since we have assumed that $\int e^g d\mu \leq 1$. Moreover, for any fixed $x = (x_1, \ldots, x_n) \in X$,

$$\int_{X_i} e^{(g^i)_i} d\mu_i = \int d\mu_i(z_i) \left[\frac{\int e^{g(y_1,\ldots,y_{i-1},z_i,x_{i+1},,\ldots,x_n)} d\mu_1 \ldots d\mu_{i-1}}{\int e^{g(y_1,\ldots,y_{i-1},y_i,x_{i+1},,\ldots,x_n)} d\mu_1 \ldots d\mu_i} \right] = 1,$$

by Fubini's theorem. It follows that:

$$\begin{aligned} \int fgd\mu \leq \sum_{i=1}^n \int fg^i d\mu &= \sum_{i=1}^n \int \left(\int_{X_i} f_i (g^i)_i d\mu_i \right) d\mu \\ &\leq \sum_{i=1}^n \int \mathrm{Ent}_{\mu_i}(f_i) d\mu \end{aligned}$$

which proves Lemma 4. □

To finish the proof of Theorem 12, we apply Lemma 4 to f^2 and use the Log-Sobolev inequality on each X_i. The result follows at once. □

5.2 Further results and applications

We record here some further results on entropy in product spaces, all of which will be used in proofs below.

Lemma 5. *For any real-valued function f defined on S^n and for $P = \mu_1 \times \cdots \times \mu_n$, we have*

$$\mathrm{Ent}_P(e^f) \leq \sum_{i=1}^{n} \int R_i(e^{f_i})(x) dP(x),$$

where, for $x = (x_1, ..., x_n)$,

$$R_i(e^{f_i})(x) := \int\int_{f_i(x_i) \geq f_i(y_i)} [f_i(x_i) - f_i(y_i)]^2 e^{f_i(x_i)} d\mu_i(x_i) d\mu_i(y_i),$$

with $f_i = f(x_1, \ldots, x_{i-1}, \cdot, x_{i+1}, \ldots, x_n)$.

Proof. By Lemma 4 it suffices to prove this lemma for $n = 1$. By Jensen's inequality,

$$\begin{aligned}
\mathrm{Ent}_P(e^f) &= \int f e^f dP - \int e^f dP \log \int e^f dP \\
&\leq \int f e^f dP - \int e^f dP \int f dP \\
&= \frac{1}{2} \int\int [f(x) - f(y)][e^{f(x)} - e^{f(y)}] dP(x) dP(y) \\
&\leq \int\int_{f(x) \geq f(y)} [f(x) - f(y)][e^{f(x)} - e^{f(y)}] dP(x) dP(y)
\end{aligned}$$

using Fubini in the last step. But, for $v \geq u$,

$$e^v - e^u = \int_u^v e^x dx \leq (v-u) e^v$$

so that $(v-u)(e^v - e^u) \leq (v-u)^2 e^v$, which gives the result. □

There is another 'variational' definition of entropy and its consequence which is sometimes useful. Let ξ be a convex function on a finite or infinite interval (e.g., $\xi(u) = u \log u$ on $[0, \infty)$), differentiable on its interior, and let the range of f be contained in it. Then, assuming existence,

$$\int \xi(f) d\mu - \xi\left(\int f d\mu\right) = \inf_t \int \left[\xi(f) - \xi(t) + (t - f)\xi'(t)\right] d\mu.$$

To see this note that the integral at the right hand side for $t = \int f d\mu$ is the left hand side. Now, note that the convex function $y = \xi(x)$ at $\int f d\mu$ is larger than or equal to the value at $\int f d\mu$ of the tangent line to the graph of this function at $(t, \xi(t))$, which gives

$$\xi\left(\int f d\mu\right) \geq \xi(t) + \left(\int f d\mu - t\right)\xi'(t),$$

proving the claim. Applied to entropy, this gives the following 'variational definition of entropy':

Lemma 6. $\mathrm{Ent}_\mu f = \inf_{t\geq 0} \int \left[f \log f - (\log t + 1) f + t\right] d\mu$. *Here μ is a probability measure and $f \geq 0$.*

Corollary 2. *Under the hypotheses of Lemma 4, and with $\phi(u) := e^{-u} + u - 1$, we have (for any $\lambda \in \mathbb{R}$)*

$$\mathrm{Ent}_P e^{\lambda f} \leq \sum_{i=1}^{n} \int \phi\left(\lambda(f(x) - f(y_i(x))\right) e^{\lambda f(x)} dP(x),$$

where $y_i(x) = (x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n)$.

Proof. By Lemma 4, it suffices to consider $n = 1$, in which case $y(x) = 0$. By Lemma 6,

$$\mathrm{Ent}_\mu e^f = \inf_{t\geq 0} \int [f e^f - (\log t + 1)e^f + t] d\mu = \inf_{u\in\mathbb{R}} \int \phi(f - u) e^f d\mu,$$

where the last identity results from changing $\log t$ to u. Now, take $u = f(0)$. □

We finish this section with a concentration result for general functionals of product measures that can derived from the log-Sobolev theory. It gives a first rigorous formalization of the intuition mentioned in the introduction that *a random variable that smoothly depends on a large number of independent random variables is 'essentially' constant, in a 'dimension-free' way.*

Theorem 13. *Let $F : [0,1]^n \mapsto \mathbb{R}$ be a separately convex (i.e., convex in each coordinate) and 1-Lipschitz function and let P be a product probability measure on $[0,1]^n$. Then, for every $r > 0$,*

$$P\left\{F \geq \int F dP + r\right\} \leq e^{-r^2/16n}.$$

Proof. If f is convex in each coordinate and smooth, then

$$f_i(x_i) - f_i(y_i) \leq (x_i - y_i) f_i'(x_i),$$

where we are using the notation of Lemma 5. That lemma gives

$$\mathrm{Ent}_P(e^f) \leq \int \int \sum_{i=1}^{n} (x_i - y_i)^2 \left(\frac{\partial f}{\partial x_i}(x) \right)^2 e^{f(x)} dP(x) dP(y),$$

Therefore, we have the log Sobolev inequality for e^f:

$$\mathrm{Ent}_P(e^f) \leq \int |\nabla f(x)|^2 dP(x).$$

Now, if f is 1-Lipschitz, $|\nabla f(x)| \leq 1$ a.e. by Rademacher's theorem, and by convolving with a Gaussian kernel if necessary, we can assume $|\nabla f(x)| \leq 1$ everywhere. Take $f = \lambda F - \lambda^2$ (wiht F convolved with a Gaussian kernel if it is not smooth) to get

$$\mathrm{Ent}_P(e^{\lambda F - \lambda^2}) \leq \lambda^2 \int e^{\lambda F - \lambda^2} dP.$$

Define $\Lambda(\lambda) = \int e^{\lambda F - \lambda^2} dP$. By definition of entropy, the last inequality becomes

$$\int e^{\lambda F - \lambda^2} (\lambda F - \lambda^2) dP - \Lambda(\lambda) \log \Lambda(\lambda) \leq \lambda^2 \Lambda(\lambda).$$

But

$$\Lambda'(\lambda) = \frac{1}{\lambda} \int (\lambda F - 2\lambda^2) e^{\lambda F - \lambda^2} dP = \frac{1}{\lambda} \int (\lambda F - \lambda^2) e^{\lambda F - \lambda^2} dP - \frac{1}{\lambda} \lambda^2 \Lambda(\lambda).$$

This yields the following differential inequality:

$$\lambda \Lambda' - \Lambda \log \Lambda \leq 0.$$

Take $H(\lambda) := \frac{1}{\lambda} \log \Lambda(\lambda)$ and note, as in previous theorems that $H(0) = \lim_{\lambda \to 0} H(\lambda) = EF$. Since

$$H'(\lambda) = -\frac{1}{\lambda^2} \log \Lambda(\lambda) + \frac{1}{\lambda} \frac{\Lambda'(\lambda)}{\Lambda(\lambda)},$$

the above inequation becomes

$$H'(\lambda) \leq 0, \quad H(0) = EF.$$

This means that $H(\lambda)$ is a non-increasing function and, since $H(0) = EF$, we obtain

$$H(\lambda) \leq EF, \quad \lambda > 0,$$

$$\log \Lambda(\lambda) \leq \lambda EF, \quad \lambda > 0,$$
$$Ee^{\lambda F - \lambda^2} \leq e^{\lambda EF}, \quad \lambda \geq 0.$$

Now, applying this inequality in combination with Chebyshev,

$$P\{F \geq EF + r\} = P\{e^{\lambda F - \lambda^2} \geq e^{\lambda EF + \lambda r - \lambda^2}\} \leq \frac{e^{\lambda EF}}{e^{\lambda EF + \lambda r - \lambda^2}} = e^{-\lambda r + \lambda^2},$$

and the result follows by taking $\lambda = r/2$. The theorem is proved for F convolved with a Gaussian kernel in the non-smooth case; convolve with Gaussian kernels G_h converging in law to δ_0 and take limits to complete the proof. □

6 Talagrand's Inequality for Empirical Processes

6.1 Empirical Processes

Throughout this and the subsequent subsections, let $X_1, ..., X_n$ be i.i.d. random variables taking values in some measurable space $S, \mathcal{A}$ and with law P (defined on $\mathcal{A}$). Denote their joint product law by Pr. Suppose we are given a uniformly bounded class $\mathcal{F}$ of functions defined on S, w.l.o.g. $\sup_{s\in S}\sup_{f\in\mathcal{F}}|f(s)| \le 1$. The *empirical process* is defined as

$$\nu_n(f) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Ef(X))\right), \qquad f \in \mathcal{F}. \tag{57}$$

In words these are centered and scaled sample means indexed by the class of functions $\mathcal{F}$. *Empirical Process Theory* studies the probabilistic properties of these stochastic processes, with a particular view on the *uniform* fluctuations of ν_n, that is, the random variables $\sup_{f\in\mathcal{F}}|\nu_n(f)|$.

As the index sets $\mathcal{F}$ of these processes are abstract we cannot use the classical theory of stochastic processes (which uses spaces of continuous functions, the cadlag space, Skorohod-topologies etc..) We should note that considering 'abstract' $\mathcal{F}$ is not 'general abstract nonsense', but arises naturally in many situations, particularly in statistics. Examples are $\mathcal{F}$ equal to a family of indicators of subsets of $\mathbb{R}^d$ – for instance $\{1_{(-\infty,t]} : t \in \mathbb{R}^d\}$, or convex subsets of $\mathbb{R}^d$, or sets with smooth boundaries –, as well as various classes of functions: for instance $\mathcal{F}$ a ball in a Lipschitz-space, or $\mathcal{F}$ a finite-dimensional vector space of functions, or $\mathcal{F} = \{K((\cdot - y)/h) : y \in \mathbb{R}, h > 0\}$, or $\mathcal{F}$ a class of functions indexed by some parameter space Θ (think of Maximum Likelihood Estimation), and many others. In a way considering abstract empirical processes is useful in a similar way as it was to consider abstract Gaussian processes in the previous section, and since sample means occur almost everywhere in statistics, proves to be broadly applicable.

It is clear that for a fixed $f \in \mathcal{F}$, we have that

$$\nu_n(f) \to^d N(0, \sigma^2(f))$$

as $n \to \infty$ by the central limit theorem, and where $\sigma^2(f) = E(f(X) - Ef(X))^2$. To make this central limit theorem 'uniform' in $\mathcal{F}$ is a highly nontrivial problem (even the formulation of it!). As this is not the focus of these lecture notes we refer to the monograph by Dudley [12] for a rigorous treatment of this problem and of the theory of empirical processes. Let us mention only that rather specific conditions have to be imposed on the class $\mathcal{F}$ for this 'uniform central limit theorem' to hold, and by definition one says that $\mathcal{F}$ is a 'Donsker'-class if ν_n converges in law in the space of bounded functions on $\mathcal{F}$ (equipped with the uniform norm) to a generalized Brownian bridge process.

Instead of studying the weak convergence of the empirical process, we ask a different question here: Can we generalize Hoeffding's inequality Proposition 2 or even Bernstein's inequality to general suprema $\sup_{f\in\mathcal{F}}|\nu(f)|$ of empirical processes (and then also to sums of i.i.d. Banach space valued random variables)? And if yes, does this result relate to the fact that there is a central limit theorem (so do we have to require that $\mathcal{F}$ is Donsker), or is the concentration of product measures a general 'dimension-free' phenomenon unrelated to the central limit theorem? One of the most striking substantiations of the concentration of measure phenomenon due to Talagrand [38] is that concentration inequalities for empirical processes can be proved in generality (without invoking CLT-type conditions).

6.2 Talagrand's inequality and variations thereof

Talagrand's inequality for empirical processes in its general form (with universal constants) [38] is

Theorem 14. *[Talagrand's Inequality] Let $\mathcal{F}$ be P-centered, countable and uniformly bounded by one. Then there exists an absolute constant K such that*

$$\Pr\left(\left|\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n f(X_i)\right| - E\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n f(X_i)\right|\right| > r\right) \le \frac{1}{K}\exp\left\{-\frac{r}{K}\log\left(1+\frac{r}{E\Sigma^2}\right)\right\}$$

where $\Sigma^2 = \sup_{f\in\mathcal{F}}\sum_{i=1}^n f^2(X_i)$ are the random variances.

The inequality is not stated in a 'Bernstein'-way (as in (11) above), nevertheless the exponent has the correct order: if $r/E\Sigma^2$ is moderate (meaning that the variances are not too small) then the tail is Gaussian (i.e., the exponent can be bounded by $r^2/E\Sigma^2K$), whereas if $r/E\Sigma^2$ is large the tail is of a Poissonian form $r\log(1+r)$.

As it stands the inequality may not be useful unless one can control the random variances. If one is only interested in 'Hoeffding' type inequalities, then one can simply estimate $E\Sigma^2 \le n$ and obtain an exponent $-(r^2/Kn)$ in the above inequality. However, taking into account the variances of $f(X)$, in particular in their 'weak' form

$$\sigma^2 \ge \sup_{f\in\mathcal{F}} Ef^2(X), \tag{58}$$

is crucial in many situations. Talagrand showed how to do this in [36], as follows:

$$\begin{aligned} E\Sigma^2 &\leq n\sigma^2 + E\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \left(f^2(X_i) - Ef^2(X)\right)\right| \\ &\leq n\sigma^2 + 4E\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n (f(X_i) - Ef(X))\right| \end{aligned}$$

where the second inequality follows from a contraction inequality for Rademacher processes that we shall detail in Subsection 6.4 below.

Working with this bound for $E\Sigma^2$ one can obtain sharp constants in a Bernstein-type version of Talagrand's inequality. This was initiated by [33]. In [6] and [30] (cf. also [34]) it is proved that

Theorem 15. *Let $\mathcal{F}$ be P-centered, countable and uniformly bounded by one, and let $\sigma^2 \geq Ef^2(X)$. Then*

$$\Pr\left(\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(X_i) \geq E\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(X_i) + r\right) \leq \exp\left\{-\frac{r^2}{2V + \frac{2}{3}r}\right\}$$

as well as

$$\Pr\left(\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(X_i) \leq E\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(X_i) - r\right) \leq \exp\left\{-\frac{r^2}{2V + 2r}\right\}$$

where $V = n\sigma^2 + 2E\sup_{f\in\mathcal{F}}|\sum_{i=1}^n f(X_i)|$.

Note that one can easily show that a bound of the above type (with universal constants) always follows from Theorem 14. Note also that if one prefers to replace $\sum f(X_i)$ by $|\sum f(X_i)|$ one can replace $\mathcal{F}$ by $\mathcal{F} \cup -\mathcal{F}$ to obtain the desired result. Specialized to a singleton class $\mathcal{F}$, the upper deviation result retrieves the one-dimensional Bernstein inequality, so this result is sharp in this sense, and in fact provides a genuine infinite-dimensional version of Bernstein's inequality. For the lower deviations the 'Gaussian' constant is sharp whereas the Poissonian is 2 instead of 2/3 – this remains an open problem at this stage.

Sometimes it is useful to state the last inequalities in the following (essentially equivalent) way:

$$\Pr\left(\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(X_i) \geq E\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(X_i) + \sqrt{2Vr} + \frac{1}{3}r\right) \leq e^{-r} \tag{59}$$

as well as

$$\Pr\left(\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(X_i) \leq E\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(X_i) - \sqrt{2Vr} - r\right) \leq e^{-r}. \tag{60}$$

6.3 A Proof of Talagrand's Inequality

We shall provide here a proof of Theorem 14. We first remark that Talagrand's inequality does not follow from Theorem 13, and refined methods are necessary. Very roughly speaking the idea is prove directly a log-Sobolev type inequality for the function $e^{\lambda Z}$,

$$Z = \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i),$$

with $Pf = 0$. These in turn yield solvable differential inequalities for the Laplace transform of Z whose solutions give exponential inequalities (recall the 'Herbst method').

Obtaining sharp constants in Talagrand's inequality basically requires a very careful (and intricate) study of the differential inequalities involved. This has been done in [33], [6] and [30], but we abstain from it to reduce technicalities. For the upper deviation version in Theorem 15 sharp constants are available, whereas for the lower deviations this is only the case for the 'Gaussian' component, and this remains an open problem.

The proof below follows Ledoux [31] who invented this proof, and borrows from the exposition in Giné [17].

6.3.1 A bound for the Laplace transform of empirical processes

We start with an auxiliary result for empirical processes that take only positive values (i.e. where $f \in \mathcal{F}$ are all positive functions).

Proposition 5. *Let $\mathcal{F}$ be a countable collection of measurable functions on S taking their values in $[0,1]$. Set $Z = \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i)$. Then, for all $\lambda \geq 0$,*

$$Ee^{\lambda Z} \leq e^{(e^{\lambda}-1)EZ}.$$

Proof. It suffices to prove this theorem for $\mathcal{F}$ finite, say $\mathcal{F} = \{f_1, \dots, f_N\}$ (by using approximation arguments as in the proof of Borell's inequality). For $1 \leq i \leq n$ and $1 \leq k \leq N$, set $x_i^k = f_k(X_i)$, $x_i = (f_1(X_i), \dots, f_N(X_i)) \in E := [0,1]^N$, $\mu_i = \mathcal{L}(x_i)$ and $P = \mu_1 \times \cdots \times \mu_n$, a probability measure on $E^n = ([0,1]^N)^n$, and $x = (x_1, \dots, x_n) \in E^n$. Then

$$Ee^{\lambda Z} = \int_{E^n} e^{\lambda Z(x)} dP(x)$$

where

$$Z(x) := \max_{1 \leq k \leq N} \sum_{i=1}^{n} x_i^k, \quad x \in E^n.$$

By Corollary 2

$$\mathrm{Ent}_P(e^{\lambda Z}) \leq \sum_{i=1}^{n} \int \phi\left(\lambda(Z(x) - Z(y_i(x))\right) e^{\lambda Z(x)} dP(x),$$

where $y_i(x) = (x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n)$. Let A_k, $k = 1, \ldots, N$, be a partition of E^n such that

$$A_k \subseteq \{x \in E^n : Z(x) = \sum_{i=1}^{n} x_i^k\}$$

(A_k is contained in the set where the maximum is attained at the k-th function). Let $\tau_k = \tau_k(x) = I_{A_k}(x)$ and $\tau = (\tau_1, \ldots, \tau_N)$. Then, since $x_i^k \geq 0$,

$$0 \leq Z(x) - Z(y_i(x)) = Z(x) - \max_{1\leq k\leq N} \sum_{1\leq j\leq n, j\neq i} x_j^k \leq \sum_{k=1}^{N} \tau_k x_i^k = \tau \cdot x_i \leq 1.$$

Since $\phi(u) := e^{-u} + u - 1$ is convex and 0 at 0, for $\lambda \geq 0$ and $u \in [0,1]$ we have $\phi(\lambda u) \leq u\phi(\lambda)$. We then conclude from the last two inequalities and Corollary 2, that

$$\mathrm{Ent}_P(e^{\lambda Z}) \leq \phi(\lambda) \sum_{i=1}^{n} \int (\tau \cdot x_i) e^{\lambda Z(x)} dP(x) = \phi(\lambda) \int Z(x) e^{\lambda Z(x)} dP(x),$$

since $\sum_{i=1}^n \tau \cdot x_i = \sum_{i=1}^n \sum_{k=1}^N \tau_k x_i^k = \sum_{k=1}^N \tau_k \sum_{i=1}^n x_i^k = Z$. This is a kind of log-Sobolev inequality.

Let $\Lambda(\lambda) = Ee^{\lambda Z}$. By definition of entropy,

$$\mathrm{Ent}_P(e^{\lambda Z}) = E\left(\lambda Z e^{\lambda Z}\right) - \left(Ee^{\lambda Z}\right)\log\left(Ee^{\lambda Z}\right) = \lambda\Lambda'(\lambda) - \Lambda(\lambda)\log\Lambda(\lambda), \quad \lambda \geq 0.$$

Hence, the log Sobolev type inequality gives

$$\lambda\Lambda'(\lambda) - \Lambda(\lambda)\log\Lambda(\lambda) \leq \phi(\lambda)\Lambda'(\lambda), \quad \lambda \geq 0.$$

or

$$(1 - e^{-\lambda})\Lambda'(\lambda) \leq \Lambda(\lambda)\log\Lambda(\lambda), \quad \lambda \geq 0.$$

With $J(\lambda) = \log\Lambda(\lambda)$, this becomes $J' \leq (1 - e^{-\lambda})^{-1} J$, or

$$(\log J)' \leq (\log(e^{\lambda} - 1))', \quad \lambda \geq 0.$$

This can be integrated: for any $\lambda_0 > 0$ and $\lambda > \lambda_0$,

$$\log\frac{J(\lambda)}{J(\lambda_0)} \leq \log\frac{e^{\lambda} - 1}{e^{\lambda_0} - 1},$$

$$J(\lambda) \leq \frac{J(\lambda_0)}{e^{\lambda_0}-1}(e^{\lambda}-1).$$

By l'Hôpital and Taylor development of the logarithm,

$$\lim_{\lambda_0 \to 0} \frac{J(\lambda_0)}{e^{\lambda_0}-1} = EZ$$

(since clearly $\lim_{\lambda \to 0}(\log Ee^{\lambda Z})' = EZ$). □

6.3.2 A Bernstein-type version of Talagrand's inequality

Theorem 16. *Let $\mathcal{F}$ be a countable collection of measurable functions on S uniformly bounded by one, and set $Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)$. Let further $\Sigma^2 := \sup_{f\in\mathcal{F}} \sum_{i=1}^n f^2(X_i)$. Then, for all $r > 0$,*

$$P\{|Z - EZ| \geq r\} \leq 2\exp\left\{-\frac{1}{10}\min\left(\frac{r^2}{3E\Sigma^2}, r\right)\right\}. \tag{61}$$

Proof. Again we can take $\mathcal{F}$ finite, say $\mathcal{F} = \{f_1, \dots, f_N\}$, the case of general $\mathcal{F}$ following from standard approximation arguments. In analogy to the proof of the previous proposition, set $x_i^k = f_k(X_i)$, $x_i = (x_i^1, \dots, x_i^N)$, $x = (x_1, \dots, x_n) \in E^n$, $E = [-1,1]^N$, $\mu_i = \mathcal{L}(f_1(X_i), \dots, f_N(X_i))$, $P = \mu_1 \times \cdots \times \mu_n$, so that $Z = Z(x) = \max_{1\leq k\leq N} \sum_{i=1}^n x_i^k$. Using Lemma 5 and its notation we have

$$\mathrm{Ent}_P(e^{\lambda Z}) = \lambda^2 \sum_{i=1}^n \int\int\int_{\lambda Z_i(x_i) \geq \lambda Z_i(y_i)} [Z_i(x_i) - Z_i(y_i)]^2 e^{\lambda Z_i(x_i)} d\mu_i(x_i) d\mu_i(y_i) dP(x)$$

where we recall that for each i the vector y is such that $y_j = x_j$ for $j \neq i$ and with $y_i \in [-1,1]^N$. [There is a slight abuse of notation as we write y instead of $y^{(i)}$, simply to avoid having to write $y_i^{(i)}$.]

Let us fix i. Let again $A_k \subseteq \{x : Z(x) = \sum_{i=1}^n x_i^k\}$, $k \leq N$, be a partition of E^n, and set $\tau_k(x) = 1_{A_k}(x)$. For $x \in A_k$ we have

$$Z(x) - Z(y) = \sum_{r=1}^n x_r^k - \max_{1\leq \ell\leq N} \sum_{r=1}^n y_r^\ell \leq \sum_{r=1}^n x_r^k - \sum_{r=1}^n y_r^k = x_i^k - y_i^k,$$

since y and x differ only in the i-th coordinate, so that

$$Z(x) - Z(y) \leq \sum_{k=1}^N \tau_k(x)(x_i^k - y_i^k) \tag{62}$$

follows (note the lack of absolute values in this inequality). Furthermore we always have

$$|Z(x) - Z(y)| \leq \max_k |x_i^k - y_i^k| \leq 2. \tag{63}$$

Then, for $\lambda > 0$,

$$\int\int_{\lambda Z_i(x_i)\geq\lambda Z_i(y_i)} [Z_i(x_i) - Z_i(y_i)]^2 e^{\lambda Z_i(x_i)} d\mu_i(x_i) d\mu_i(y_i)$$

$$\leq \int\int \sum_{k=1}^{N} \tau_k(x)(x_i^k - y_i^k)^2 e^{\lambda Z_i(x_i)} d\mu_i(x_i) d\mu_i(y_i)$$

using (62) together with the fact that $Z_i(x_i) \geq Z_i(y_i)$ on the domain of integration. Thus, in this case,

$$\begin{aligned} \mathrm{Ent}_P(e^{\lambda Z}) &\leq \lambda^2 \sum_{i=1}^{n} \int\int \sum_{k=1}^{N} \tau_k(x)(x_i^k - y_i^k)^2 e^{\lambda Z(x)} dP(x) d\mu_i(y_i) \\ &= \lambda^2 \int\int \sum_{i=1}^{n} \sum_{k=1}^{N} \tau_k(x)(x_i^k - y_i^k)^2 e^{\lambda Z(x)} dP(x) d\mu_i(y_i). \end{aligned}$$

Now,

$$\sum_{k=1}^{N} \tau_k(x) \sum_{i=1}^{n} (x_i^k - y_i^k)^2 \leq \max_k \sum_{i=1}^{n} (x_i^k - y_i^k)^2 \leq 2 \max_k \sum_{i=1}^{n} (x_i^k)^2 + 2 \max_k \sum_{i=1}^{n} (y_i^k)^2,$$

and we get, for $\lambda > 0$ as well as trivially for $\lambda = 0$,

$$\mathrm{Ent}_P(e^{\lambda Z}) \leq 2\lambda^2 \left[E(\Sigma^2 e^{\lambda Z}) + E(\Sigma^2) E(e^{\lambda Z}) \right]. \tag{64}$$

For $\lambda < 0$, we use (62) with x and y interchanged, and (63), to obtain, for each i,

$$\int\int_{\lambda Z_i(x_i)\geq\lambda Z_i(y_i)} [Z_i(x_i) - Z_i(y_i)]^2 e^{\lambda Z_i(x_i)} d\mu_i(x_i) d\mu_i(y_i)$$

$$\leq \int\int \sum_{k=1}^{N} \tau_k(y)(x_i^k - y_i^k)^2 e^{\lambda Z_i(y_i) + 2|\lambda|} d\mu_i(x_i) d\mu_i(y_i).$$

Then, adding over i and integrating, as before, we get, for $\lambda < 0$,

$$\mathrm{Ent}_P(e^{\lambda Z}) \leq 2\lambda^2 e^{2|\lambda|} \left[E(\Sigma^2 e^{\lambda Z}) + E(\Sigma^2) E(e^{\lambda Z}) \right]. \tag{65}$$

(64) and (65) give that, for all $|\lambda| \leq \lambda_0$, λ_0 to be specified below,

$$\mathrm{Ent}_P(e^{\lambda Z}) \leq c_0 \lambda^2 \left[E(\Sigma^2 e^{\lambda Z}) + E(\Sigma^2) E(e^{\lambda Z}) \right]. \tag{66}$$

with $c_0 = 2e^{2\lambda_0}$. This is a kind of modified log Sobolev inequality, and what remains now is to integrate the corresponding expression for the Laplace transform of $\tilde{Z} = Z - EZ$. First we transform it.

Set $\tilde{\Lambda}(\lambda) := Ee^{\lambda\tilde{Z}}$. Since $e^{\lambda\tilde{Z}} = e^{-\lambda EZ}e^{\lambda Z}$, homogeneity of Ent and (66) give

$$\lambda\tilde{\Lambda}'(\lambda) - \tilde{\Lambda}(\lambda)\log\tilde{\Lambda}(\lambda) = \mathrm{Ent}_P(e^{\lambda\tilde{Z}}) \le c_0\lambda^2\left[(E\Sigma^2)\tilde{\Lambda}(\lambda) + E(\Sigma^2 e^{\lambda\tilde{Z}})\right], \quad |\lambda| \le \lambda_0. \tag{67}$$

Now we estimate the second summand at the right hand side. Using Young's inquality with $y = e^{\lambda\tilde{Z}} \ge 0$ and $x = \Sigma^2 - (e-1)E\Sigma^2$, we get

$$\begin{aligned} E(\Sigma^2 e^{\lambda\tilde{Z}}) &= (e-1)(E\Sigma^2)Ee^{\lambda\tilde{Z}} + E\left[(\Sigma^2 - (e-1)E\Sigma^2)e^{\lambda\tilde{Z}}\right] \\ &\le (e-1)(E\Sigma^2)Ee^{\lambda\tilde{Z}} + \lambda E(\tilde{Z}e^{\lambda\tilde{Z}}) - Ee^{\lambda\tilde{Z}} + Ee^{\Sigma^2-(e-1)E\Sigma^2}. \end{aligned}$$

By Jensen's inequality, $Ee^{\lambda\tilde{Z}} \ge 1$ (here we are using $E\tilde{Z} = 0$) and by Proposition 5 applied to Σ^2 we also have

$$Ee^{\Sigma^2-(e-1)E\Sigma^2} \le 1,$$

which plugged in the previous inequality give

$$E(\Sigma^2 e^{\lambda\tilde{Z}}) \le (e-1)(E\Sigma^2)\tilde{\Lambda}(\lambda) + \lambda\tilde{\Lambda}'(\lambda), \quad \lambda \in \mathbb{R}.$$

Replacing this in (67) yields a differential inequality that we will be able to integrate:

$$\lambda\tilde{\Lambda}' - \tilde{\Lambda}\log\tilde{\Lambda} \le c_0\lambda^2\left(e(E\Sigma^2)\tilde{\Lambda} + \lambda\Lambda'\right), \quad |\lambda| \le \lambda_0. \tag{68}$$

Set

$$H(\lambda) = \frac{1}{\lambda}\log\tilde{\Lambda}(\lambda)$$

and notice that

$$H(0) := \lim_{\lambda\to 0} H(\lambda) = 0$$

(see the end of the proof of Proposition 5). Since

$$H' = -\frac{1}{\lambda^2}\log\tilde{\Lambda} + \frac{1}{\lambda}\frac{\tilde{\Lambda}'}{\tilde{\Lambda}} = \frac{1}{\lambda^2\tilde{\Lambda}}(\lambda\tilde{\Lambda}' - \tilde{\Lambda}\log\tilde{\Lambda}),$$

inequality (68) becomes

$$H' \le c_0\left(eE\Sigma^2 + \lambda\frac{\tilde{\Lambda}'}{\tilde{\Lambda}} + \log\tilde{\Lambda}(\lambda)\right), \quad H(0) = 0, \quad |\lambda| \le \lambda_0 \tag{69}$$

where we also added $c_0 \log \tilde{\Lambda}(\lambda) \geq 0$ (cf. Jensen above) on the right hand side. Integrating from 0 to λ ($|\lambda| \leq \lambda_0$), and taking into account that $\lim_{\lambda \to 0} \lambda^{-1} \log E e^{\lambda \tilde{Z}} = 0$, we get

$$\log \tilde{\Lambda} \leq c_0 (e\lambda^2 E\Sigma^2 + \lambda^2 \log \tilde{\Lambda}).$$

If $c_0 \lambda_0^2 < 1$, this is $(1 - c_0\lambda^2) \log \tilde{\Lambda} \leq c_0 e \lambda^2 E\Sigma^2$, that is, with $\kappa_0 = c_0 e/(1 - c_0 \lambda_0^2)$,

$$\tilde{\Lambda}(\lambda) \leq e^{\kappa_0 \lambda^2 E\Sigma^2}, \quad |\lambda| \leq \lambda_0.$$

Hence, for $0 \leq \lambda \leq \lambda_0$ and $r > 0$,

$$\mathbb{P}\{Z - EZ \geq r\} = P(e^{\lambda \tilde{Z}} \geq e^{\lambda r}) \leq \exp\left(\kappa_0 \lambda^2 E\Sigma^2 - \lambda r\right),$$

and for $-\lambda_0 \leq \lambda \leq 0$,

$$\mathbb{P}\{Z - EZ \leq -r\} = P(e^{\lambda \tilde{Z}} \geq e^{-\lambda r}) \leq \exp\left(\kappa_0 \lambda^2 E\Sigma^2 - |\lambda| r\right).$$

Make the natural choice $|\lambda| = r/(2\kappa_0 E\Sigma^2)$ as long as this is dominated by λ_0, that is, for $r \leq 2\kappa_0 \lambda_0 E\Sigma^2$, which gives an exponent of $-r^2/(4\kappa_0 E\Sigma^2)$, and choose $|\lambda| = \lambda_0$ for $r \geq 2\kappa_0 \lambda_0 E\Sigma^2$, to get an exponent smaller than or equal to $\lambda_0 r/2$, to get

$$P\{Z - EZ \geq r\} \leq \exp\left\{- \min\left(\frac{r^2}{4\kappa_0 E\Sigma^2}, \frac{\lambda_0 r}{2}\right)\right\}$$

as well as

$$P\{|Z - EZ| \geq r\} \leq 2\exp\left\{- \min\left(\frac{r^2}{4\kappa_0 E\Sigma^2}, \frac{\lambda_0 r}{2}\right)\right\}.$$

Constants: the only constraint is $c_0 \lambda_0^2 = 2e^{2\lambda_0} \lambda_0^2 < 1$. One gets some constants, but not the best. For instance, for $\lambda_0 = 1/5$ we get

$$P\{|Z - EZ| \geq r\} \leq 2\exp\left\{-\frac{1}{10} \min\left(\frac{r^2}{3E\Sigma^2}, r\right)\right\},$$

which is the desired result. □

6.3.3 Completion of the Proof of Theorem 14

To improve the Bernstein-type Theorem 16 to the more general exponent from Theorem 14, one truncates at a certain level, applies Theorem 16 to the truncated variables, and Proposition 5 to the sums of absolute values of the variables truncated away from zero.

Let $0 < \rho \leq 1$ to be chosen below, and define $\mathcal{F}_\rho := \{fI_{|f|\leq\rho} : f \in \mathcal{F}\}$, $Z_\rho := \sup_{f\in\mathcal{F}_\rho} \sum_{i=1}^n f(X_i)$ and $W_\rho = \sup_{f\in\mathcal{F}} \sum_{i=1}^n |f(X_i)|I_{|f(X_i)|>\rho}$. Write

$$P\{|Z - EZ| \geq 4r\} \leq P\{|Z_\rho - EZ_\rho| \geq r\} + \mathbb{P}\{W_\rho + EW_\rho \geq 3r\}.$$

By Theorem 16 and homogeneity

$$P\{|Z_\rho - EZ_\rho| \geq r\} \leq 2\exp\left\{-\frac{1}{10}\min\left(\frac{r^2}{3E\Sigma^2}, \frac{r}{\rho}\right)\right\},$$

for all $r > 0$, and. Note further that Proposition 5 implies by Markov's inequality and a simple optimization in λ

$$\Pr\{Z \geq EZ + r\} \leq \exp\{-E(Z)h(r/E(Z))\}$$

for $h(u) = (1+u)\log(1+u)$. This gives, for $r > EW_\rho$

$$P\{W_\rho + EW_\rho \geq 3r\} \leq P\{W_\rho \geq EW_\rho + r\} \leq \exp\left\{-\frac{r}{2}\log\left(1 + \frac{r}{EW_\rho}\right)\right\}.$$

Choose then

$$\rho = \rho(r) = \min\left(1, \sqrt{\frac{E\Sigma^2}{r}}\right).$$

Either $\rho = 1$ and $W_\rho = 0$ or $\rho \leq 1$ and $r \geq E\Sigma^2$; in this last case, since $W_\rho \leq \Sigma^2/\rho$, we have

$$r \geq \sqrt{rE\Sigma^2} = \frac{E\Sigma^2}{\rho} \geq EW_\rho, \tag{70}$$

so, the inequality for W_ρ applies in this case. Now, for all $u \geq 0$,

$$u/3 > 12^{-1}\log(1 + 4u),$$

and we therefore have, in both cases,

$$r \times \min\left(\frac{1}{\rho}, \frac{r}{3E\Sigma^2}\right) \geq \frac{r}{12}\log\left(1 + \frac{4r}{E\Sigma^2}\right).$$

For the second case, by (70),

$$\log\left(1 + \frac{r}{EW_\rho}\right) \geq \log\left(1 + \sqrt{\frac{r}{E\Sigma^2}}\right) \geq \frac{1}{4}\log\left(1 + \frac{4r}{E\Sigma^2}\right).$$

Combining we obtain

$$P\{|Z - EZ| \geq 4r\} \leq 3\exp\left\{\frac{r}{120}\log\left(1 + \frac{4r}{E\Sigma^2}\right)\right\}.$$

which proves Theorem 14 without absolute values. Applying this theorem to $\mathcal{F} \cup -\mathcal{F}$ completes the proof.

6.4 Moment Bounds via Rademacher Symmetrization

If one wants to apply Talagrand's inequality Theorem 14, some formidable tasks remain: First, there is the task to control

$$E\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}f^2(X_i)\leq n\sigma^2+E\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\left(f^2(X_i)-Ef^2(X)\right)\right| \tag{71}$$

in the case where one needs a result that depends on the weak variances $\sigma^2 \geq \sup_{f\in\mathcal{F}} Ef^2(X)$ and the bound n is too crude. Second – if one needs information on the quantity $\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n} f(X_i)\right|$ rather than on

$$\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}f(X_i)\right|-E\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}f(X_i)\right|$$

then the size of the quantity

$$E\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}f(X_i)\right| \tag{72}$$

has to be estimated. In both cases we have to control the moment of the supremum of a centered empirical process, and we shall discuss in this subsection some tools to do this.

6.4.1 Rademacher Processes

A key technique in the theory of empirical processes is *Rademacher symmetrization.* This was first introduced into empirical processes in the classical paper [21] and we show how this applies in the context of Talagrand's inequality.

Let $\varepsilon_i, i=1,...,n$, be i.i.d. Rademacher random signs (taking values $-1,1$ with probability 1/2), independent of the $X_i's$, defined on a large product probability space with product probability Pr, denote the joint expectation by E, and by E_ε and E_X the corresponding expectations w.r.t. the ε_i's and the X_i's respectively. The following symmetrization inequality holds for random variables in arbitrary normed spaces, but we state it for the supremum norm relevant in empirical process theory: For $\mathcal{F}$ a class of functions on $(S,\mathcal{A})$, define $\|H\|_{\mathcal{F}}=\sup_{f\in\mathcal{F}}|H(f)|$.

Lemma 7. *Let $\mathcal{F}$ be a uniformly bounded P-centered class of functions defined on a measurable space $(S,\mathcal{A})$. Let ε_i be i.i.d. Rademachers as above, and let $a_i, i=1,...,n$ be any sequence of real numbers. Then*

$$\frac{1}{2}E\left\|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right\|_{\mathcal{F}}\leq E\left\|\sum_{i=1}^{n}f(X_i)\right\|_{\mathcal{F}}\leq 2E\left\|\sum_{i=1}^{n}\varepsilon_i(f(X_i)+a_i)\right\|_{\mathcal{F}}. \tag{73}$$

Proof. Let us assume for simplicity that $\mathcal{F}$ is countable (so that we can neglect measurability problems). Since $E_X f(X_i) = 0$ for every f, i, the first inequality follows from

$$\begin{aligned}
&E\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}} = E_\varepsilon E_X \left\|\sum_{i:\varepsilon_i=1} f(X_i) - \sum_{i:\varepsilon_i=-1} f(X_i)\right\|_{\mathcal{F}} \\
&\le E_\varepsilon E_X \left\|\sum_{i:\varepsilon_i=-1} f(X_i) + E_X \sum_{i:\varepsilon_i=1} f(X_i)\right\|_{\mathcal{F}} + E_\varepsilon E_X \left\|\sum_{i:\varepsilon_i=1} f(X_i) + E_X \sum_{i:\varepsilon_i=-1} f(X_i)\right\|_{\mathcal{F}} \\
&\le 2E\left\|\sum_{i=1}^n f(X_i)\right\|_{\mathcal{F}}
\end{aligned}$$

where in the last inequality we have used Jensen's inequality and convexity of the norm. To prove the second inequality, let $X_{n+i}, i = 1, ..., n$ be an independent copy of $X_1, ..., X_n$. Then, proceeding as above,

$$\begin{aligned}
E\left\|\sum_{i=1}^n f(X_i)\right\|_{\mathcal{F}} &= E\left\|\sum_{i=1}^n (f(X_i) - Ef(X_{n+i})\right\|_{\mathcal{F}} \\
&\le E\left\|\sum_{i=1}^n (f(X_i) + a_i) - \sum_{i=1}^n (f(X_{n+i}) + a_i)\right\|_{\mathcal{F}}
\end{aligned}$$

which clearly equals

$$E_\varepsilon E_X \left\| \sum_{i:\varepsilon_i=1} \varepsilon_i(f(X_i) + a_i - f(X_{n+i}) - a_i) - \sum_{i:\varepsilon_i=-1} \varepsilon_i(f(X_i) + a_i - f(X_{n+i}) - a_i))\right\|_{\mathcal{F}}.$$

Now Pr being a product probability measure with identical coordinates, it is invariant by permutations of the coordinates, so that we may interchange $f(X_i)$ and $f(X_{n+i})$ for the $i's$ where $\varepsilon_i = -1$ in the last expectation. This gives that the quantity in the last display equals

$$E_\varepsilon E_X \left\|\sum_{i=1}^n \varepsilon_i(f(X_i) + a_i - f(X_{n+i}) - a_i)\right\|_{\mathcal{F}} \le 2E\left\|\sum_{i=1}^n \varepsilon_i(f(X_i) + a_i)\right\|_{\mathcal{F}}$$

which completes the proof. □

This simple but very useful result says that we can always compare the size of the expectation of the supremum of an empirical process to a symmetrized process. The idea usually is that the symmetrized 'Rademacher process' has, conditional on the X_i's, a very simple structure. One can then derive results for the Rademacher process and integrate the results over the distribution of the X_i's.

Let us illustrate this by an example, which is a contraction principle for Rademacher processes due to Kahane [28] that we state here without proof. See [32], p.112, for a proof.

Theorem 17 (Contraction Principle for Rademacher Processes)**.** *Let $F : \mathbb{R}^+ \to \mathbb{R}^+$ be convex and increasing. Let $\varphi_i : \mathbb{R} \to \mathbb{R}$, $\phi(0) = 0$, be contractions (i.e., such that $|\varphi_i(s) - \varphi_i(t)| \le |s - t|$) and let T be a bounded subset of $\mathbb{R}^n$. Let $(\varepsilon_i)_{i=1}^n$ be i.i.d. Rademachers. Then*

$$EF\left(\frac{1}{2}\sup_{t=(t_1,...,t_n)\in T}\left|\sum_{i=1}^n \varepsilon_i \phi(t_i)\right|\right) \le EF\left(\sup_{t=(t_1,...,t_n)\in T}\left|\sum_{i=1}^n \varepsilon_i t_i\right|\right)$$

Combining this with Lemma 7 we can reduce the moment of the centered random variances in (71) to the usual moment of the empirical process.

Proposition 6. *Let $\mathcal{F}$ be as in Lemma 7 and satisfying in addition that* $\sup_{f\in\mathcal{F}}\|f\|_\infty$ *is bounded by one. Then*

$$E\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \left(f^2(X_i) - Ef^2(X)\right)\right| \le 16E\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n f(X_i)\right|.$$

Proof. We first use Lemma 7 with $a_i = -Ef^2(X_i)$ to obtain

$$E\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \left(f^2(X_i) - Ef^2(X)\right)\right| \le 2E\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f^2(X_i)\right| = 4E\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i (f^2(X_i)/2)\right|.$$

Now apply Theorem 17 with $t_i = f(X_i)$, $T = \{(f(X_i) : i = 1, ..., n) : f \in \mathcal{F}\} \subset \mathbb{R}^n$ and $\varphi_i(s) = \varphi(s) = \min(s^2/2, 1)$ which satisfies $\varphi(0) = 0$ as well as $|\phi(s) - \phi(t)| = |(t - s)(t + s)/2| \le |t - s|$ so that the r.h.s. of the last display is bounded by

$$8E\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right| \le 16E\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n f(X_i)\right|$$

where the last inequality follows from the first part of Lemma 7 ('desymmetrization'). □

6.4.2 Moment Bounds for Empirical Processes

The question then in the application of Talagrand's inequality is how we can estimate the size of $E\sup_{f\in\mathcal{F}}|\sum_{i=1}^n f(X_i)|$? Here the 'geometry' and 'size' of the set $\mathcal{F}$ cannot be neglected.

A classical approach comes from empirical processes and works with 'entropy methods' – entropy here meaning a measure of the size in an infinite-dimensional set: Define $N(\varepsilon, \mathcal{F}, L_2(Q))$ to be the minimal number of balls needed to cover the class $\mathcal{F}$ by balls of radius less than ε in the $L^2(Q)$ norm, where Q is some probability measure on $(S, \mathcal{A})$. The logarithm of this covering number is sometimes called the entropy of $\mathcal{F}$ w.r.t. $L^2(Q)$, the origin of this name being due to Kolmogorov, who was aware of the possible confusion.

This way of measuring the size of classes of functions is at the heart of the study of the central limit theorem for empirical processes, but can be applied to moment bounds as well. The proof uses chaining techniques and Rademacher symmetrization, but does not belong to this course.

Theorem 18. *Suppose $\mathcal{F}$ is P-centered and uniformly bounded by one, and denote by $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ the empirical measure associated with the sample. Suppose*

$$\log N(\varepsilon, \mathcal{F}, L^2(P_n)) \leq H(1/\varepsilon) \tag{74}$$

for every n and some regularly varying function H that is zero on $[0, 1/2]$ and increasing on $[1/2, \infty)$. Denote by $\sigma^2 \geq Ef^2(X)$ the usual weak variances. Then there exists a constant $C(H)$ such that

$$E\left\|\sum_{i=1}^n f(X_i)\right\|_{\mathcal{F}} \leq C(H)\left(\sqrt{n}\sigma\sqrt{H(2/\sigma)} + H(2/\sigma)\right).$$

This theorem has a fairly long history. While it roots go back to Dudley's famous paper [10], with subsequent work in [36], [13], this general version is due to [18]. There is also some work on the constants in this inequality in the appendix of [19]. Remarkably [18] show that the above theorem is sharp at least if the first term in the upper bound is dominant (and if the estimate for H is sharp).

While working with covering numbers with respect to the $L^2(P_n)$ metrics is useful for sharp formulations (Talagrand calls this 'random geometry'), to apply this theorem it is usually more practical to estimate the 'uniform' covering numbers $\sup_Q \log N(\varepsilon, \mathcal{F}, L^2(Q))$. This is usually still a formidable task, but empirical process theory provides several tools for this – from combinatorics, approximation theory, geometry etc. There is a rich and applicable number of examples as well as a general theory, which however, cannot be covered here. We refer to [12] for a very comprehensive account. For instance for all the examples of classes of functions mentioned in the introduction to this section sharp uniform entropy bounds exist in the literature.

6.4.3 A 'statistical version' of Talagrand's Inequality

The moment bounds from the last subsection could be used to estimate the moment $E\|\sum_{i=1}^n f(X_i)\|_{\mathcal{F}}$ by an upper bound that depends only on σ and n. However, what does one do if this bound turns out not to be sharp, or if one does not obtain a bound at all? In statistical applications it is often sufficient to have a good 'random' estimate of $E\left\|\sum_{i=1}^n f(X_i)\right\|_{\mathcal{F}}$. An interesting idea in this direction – due to Koltchinskii (see, e.g., [29] and also his forthcoming St.Flour lecture notes) is the following: Let us assume for the moment that $\mathcal{F}$ is *not* P-centered, so that we want to control

$$E\left\|\sum_{i=1}^n (f(X_i) - Ef(X_i))\right\|_{\mathcal{F}}.$$

The 'statistician' usually does not know this quantity (due to the two expectations). However, by Lemma 7 with $a_i = -Ef(X_i)$ this quantity is less than or equal to

$$2E\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}}, \tag{75}$$

and by virtue of Talagrand's inequality for the two-sample empirical process ε_i, X_i, the random quantity

$$\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}},$$

which is usually computable for the statistician, concentrates around the expectation featuring in (75). Using these ideas we can prove, starting from (59) and (60), the following result. We do not care too much about optimality of the constants here, but it should be clear that they are of a 'small' numerical form.

Proposition 7. *Let $\mathcal{F}$ be a countable class of functions uniformly bounded by $1/2$, and let*

$$\sigma^2 \geq \sup_{f\in\mathcal{F}} Ef^2(X)$$

be the weak variances. We have for every $n \in \mathbb{N}$ and $x > 0$

$$\Pr\left\{\left\|\frac{1}{n}\sum_{i=1}^n (f(X_i) - Ef(X))\right\|_{\mathcal{F}} \geq 6\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}} + 10\sqrt{\frac{(x+\log 2)\sigma^2}{n}} + 22\frac{x+\log 2}{n}\right\} \leq e^{-x}$$

Proof. Let us define $Z = E\left\|\sum_{i=1}^n (f(X_i) - Ef(X))\right\|_{\mathcal{F}}$, then we have from (60) that

$$\begin{aligned}
e^{-x} &\geq \Pr\left\{Z \leq EZ - \sqrt{2x\left(n\sigma^2 + 2EZ\right)} - x\right\} \\
&\geq \Pr\left\{Z \leq EZ - \sqrt{2xn\sigma^2} - \sqrt{4xEZ} - x\right\} \\
&\geq \Pr\left\{Z \leq 0.5EZ - \sqrt{2xn\sigma^2} - 3x\right\} \qquad (76) \\
&= \Pr\left\{\left\|\frac{1}{n}\sum(f(X_i) - Pf)\right\|_{\mathcal{F}} \leq 0.5E\left\|\frac{1}{n}\sum(f(X_i) - Pf)\right\|_{\mathcal{F}} - \sqrt{\frac{2x\sigma^2}{n}} - \frac{3x}{n}\right\}
\end{aligned}$$

where we have used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ as well as $\sqrt{ab} \leq \frac{a+b}{2}$. By the same reasoning we have from (59)

$$\Pr\left\{\left\|\frac{1}{n}\sum(f(X_i) - Pf)\right\|_{\mathcal{F}} \geq 1.5E\left\|\frac{1}{n}\sum(f(X_i) - Pf)\right\|_{\mathcal{F}} + \sqrt{\frac{2x\sigma^2}{n}} + \frac{7x}{3n}\right\}. \qquad (77)$$

To prove the proposition, observe

$$\begin{aligned}
&\Pr\left\{\left\|\frac{1}{n}\sum(f(X_i) - Pf)\right\|_{\mathcal{F}} \geq 6\left\|\frac{1}{n}\sum\varepsilon_i f(X_i)\right\|_{\mathcal{F}} + 10\sqrt{\frac{x\sigma^2}{n}} + \frac{22x}{n}\right\} \\
&\leq \Pr\left\{\left\|\frac{1}{n}\sum(f(X_i) - Pf)\right\|_{\mathcal{F}} \geq 3E\left\|\frac{1}{n}\sum\varepsilon_i f(X_i)\right\|_{\mathcal{F}} + 1.5\sqrt{\frac{x\sigma^2}{n}} + 0.15\frac{22x}{n}\right\} \\
&+ \Pr\left\{6\left\|\frac{1}{n}\sum\varepsilon_i f(X_i)\right\|_{\mathcal{F}} - 3E\left\|\frac{1}{n}\sum\varepsilon_i f(X_i)\right\|_{\mathcal{F}} < -8.5\sqrt{\frac{x\sigma^2}{n}} - 0.85\frac{22x}{n}\right\} \\
&\leq \Pr\left\{\left\|\frac{1}{n}\sum(f(X_i) - Pf)\right\|_{\mathcal{F}} \geq 1.5E\left\|\frac{1}{n}\sum(f(X_i) - Pf)\right\|_{\mathcal{F}} + \sqrt{\frac{2x\sigma^2}{n}} + \frac{7x}{3n}\right\} \\
&+ \Pr\left\{\left\|\frac{1}{n}\sum\varepsilon_i f(X_i)\right\|_{\mathcal{F}} < 0.5E\left\|\frac{1}{n}\sum\varepsilon_i f(X_i)\right\|_{\mathcal{F}} - \sqrt{\frac{2x\sigma^2}{n}} - \frac{3x}{n}\right\}
\end{aligned}$$

where we have used Lemma 7. The first quantity on the r.h.s. of the last inequality is less than or equal to e^{-x} by using the bound (77). For the second term, note that (76) applies to the randomized sums $\sum_{i=1}^n \varepsilon_i f(X_i)$ as well by just taking the class of functions

$$\mathcal{G} = \{g(\tau, x) = \tau f(x) : f \in \mathcal{F}\},$$

$\tau \in \{-1, 1\}$, instead of $\mathcal{F}$ and the probability measure $\bar{P} = 2^{-1}(\delta_{-1} + \delta_1) \times P$ instead of P. It is easy to see that σ can be taken to be the same as for $\mathcal{F}$, so that (76) applies. This gives the overall bound $2e^{-x}$, and a change of variables in x gives the final bound. □

6.5 Sums of i.i.d. Banach-Space valued Random Variables

Similar to Subsection 4.4 above, one can ask how results for empirical processes carry over to sums of i.i.d. Banach space valued random variables. Suppose we are given a sequence of i.i.d. centered random variables $X, X_1, ..., X_n$ taking values in the (for simplicity again separable) Banach space $(B, \|\cdot\|_B)$, and assume furthermore that the variables are bounded by one: $\|X\|_B \leq 1$. [Centered means here $EX = 0$ in the Bochner sense but it suffices by boundedness of X to check that $Ef(X) = 0$ for every $f \in B'$.] Then taking $\mathcal{F}$ to be the unit ball of B' as in Subsection 4.4 above, we see that $\sup_{f\in\mathcal{F}} |f(X)| \leq \|f\|_B' \|X\|_B = 1$, so that the class $\mathcal{F}$ is uniformly bounded by one. Furthermore the norm of the sample mean satisfies

$$\left\|\sum_{i=1}^n X_i\right\|_B = \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n f(X_i)\right|$$

and Talagrand's inequality Theorem 14 gives us, for some universal constant K, that

$$\Pr\left\{\left|\left\|\sum_{i=1}^n X_i\right\|_B - E\left\|\sum_{i=1}^n X_i\right\|_B\right| \geq r\right\} \leq \frac{1}{K}\exp\left\{-\frac{r^2}{Kn}\right\},$$

a genuine analogue of Hoeffding's inequality in infinite dimensions. This can be used to prove, for instance, under sharp conditions (using truncation), the law of the iterated logarithm in Banach spaces. It also implies a law of large numbers in Banach spaces (just use Borell-Cantelli), although clearly the boundedness assumption on X is not necessary.

Moreover, if we take into account the random variances, and using Proposition 6, then

$$\Pr\left\{\left|\left\|\sum_{i=1}^n X_i\right\|_B - E\left\|\sum_{i=1}^n X_i\right\|_B\right| \geq r\right\} \leq \frac{1}{K}\exp\left\{-\frac{r}{K}\log\left(1+\frac{r}{E\Sigma^2}\right)\right\}$$

where

$$E\Sigma^2 \leq n\sigma^2 + 16E\left\|\sum_{i=1}^n X_i\right\|_B$$

with $\sigma^2 \geq \sup_{L:\|L\|_B' \leq 1} EL^2(X)$ the weak variances. The moments $E\left\|\sum_{i=1}^n X_i\right\|_B$ can be directly controlled by using techniques from probability in Banach spaces (taking into account the geometry of B). We refer to [1] and [32] for results of this type. One can alternatively try to use Theorem 18 with $\mathcal{F}$ equal to the unit ball of B' to control moments of norms of sums of i.i.d. Banach space random variables. This will be particularly useful when B is far away from a Hilbert space (and hence has no 'nice' geometry), for instance if B equals the space $C(K)$ of continuous functions on K for K some compact metric space.

6.6 Estimation of a Probability Measure

Talagrand's inequality was partly motivated by concrete problems in statistics (see [2] and in particular [34]), and has since found an enormous amount of applications in statistics, which cannot be summarized here. To give a flavour of some of the results that can be proved using Talagrand's inequality, we shall discuss in this section a theorem that sheds new light on the most fundamental problem of statistics – estimation of the distribution of a random variable.

Suppose we are given a sample $X_1, ..., X_n$ of i.i.d. random variables each of which has law P with distribution $F(t) = P(X \leq t)$. The classical theorems of mathematical statistics attempt to establish that the *empirical distribution function*

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,t]}(X_i)$$

is the *optimal* estimator for F. There are results on this fact from Doob, Kolmogorov, Smirnov, Donsker, Dvoretzky, Kiefer, Wolfowitz, Dudley and many others between the 30s and the 60s. One of the main theorems of this theory is the 'asymptotic minimaxity' of F_n, proved by Dvoretzky, Kiefer and Wolfowitz [9]: Denote by $\mathcal{P}$ the set of all distributions on $\mathbb{R}$, and by $\mathcal{T}$ the set of all 'estimators' (all real-valued measurable functions of the sample $X_1, ..., X_n$ and t). Then

$$\lim_n \frac{\sup_{F \in \mathcal{P}} E_F \sup_{t \in \mathbb{R}} |F_n(t) - F(t)|}{\inf_{T_n \in \mathcal{T}} \sup_{F \in \mathcal{P}} E_F \sup_{t \in \mathbb{R}} |T_n(t) - F(t)|} = 1. \tag{78}$$

The conclusion is that if *nothing is known apriori* about the distribution of F, then the empirical distribution F_n cannot be improved upon as an estimator (at least for large samples). This result is mirrored in Donsker's *functional central limit theorem* which says that

$$\sqrt{n}(F_n - F) \to^d G_P \quad in \quad \ell^\infty(\mathbb{R}) \tag{79}$$

where $\ell^\infty(\mathbb{R})$ is the space of bounded functions on $\mathbb{R}$, and where G_P is the P-Brownian bridge.

However, the sample may contain more information – for instance one may be interested in estimating the density f of F as well. Clearly the statistician usually does not know that F has a density, and indeed even if a density exists the statistical performance of any estimator heavily depends on the regularity properties of f: To understand this phenomenon, define for any nonnegative integer s the spaces $C^s(\mathbb{R})$ of all bounded continuous real-valued functions that are s-times continuously differentiable on $\mathbb{R}$, equipped with the norm

$$\|f\|_{s,\infty} = \sum_{0 \leq \alpha \leq s} \|D^\alpha f\|_\infty .$$

For noninteger $s > 0$, set

$$\|f\|_{s,\infty} := \sum_{0 \le \alpha \le [s]} \|D^\alpha f\|_\infty + \sup_{x \ne y} \frac{\left|D^{[s]}f(x) - D^{[s]}f(y)\right|}{|x-y|^{s-[s]}}$$

and define

$$C^s(\mathbb{R}) = \left\{ f \in \mathsf{C}^{[s]}(\mathbb{R}) : \|f\|_{s,\infty} < \infty \right\}, \tag{80}$$

where $[s]$ denotes the integer part of s.

One can then prove (see [27]) that

$$\lim_n \left(\frac{n}{\log n}\right)^{t/(2t+1)} \inf_{T_n \in \mathcal{T}} \sup_{f : \|f\|_{t,\infty} \le D} E_f \sup_{x \in \mathbb{R}} |T_n(x) - f(x)| \ge c(D) > 0 \tag{81}$$

so that the best rate of consistency (convergence) that *any estimator* for f can achieve depends on t. Now constructing an estimator that achieves this rate of convergence is not simple, and will usually require knowledge of t (which the statistician scarcely has).

Using Talagrand's inequality and methods from nonparametric statistics developed in the last 20 years, one can prove the following result, which shows that one can construct a purely-data-driven estimator $\hat{F}_n$ which is as good as the empirical distribution function for estimating F (cf. (79)) but also optimally estimates the density f of F, if it exists, but without any apriori information required! In this sense $\hat{F}_n$ strictly outperforms the empirical distribution function.

Theorem 19. *Let $X_1, ..., X_n$ be i.i.d. on $\mathbb{R}$ with common law P. For any given T there exists a (purely-data-driven) estimator $\hat{F}_n$ such that*

$$\sqrt{n}\left(\hat{F}_n - F\right) \to^d G_P \quad \text{in } \ell^\infty(\mathbb{R}),$$

the convergence being uniform over the set of all probability measures P on $\mathbb{R}$, in any distance that metrizes convergence in law. If furthermore P possesses a bounded and uniformly continuous density f with respect to Lebesgue measure, then

$$\{\text{the Lebesgue density } \hat{f}_n \text{ of } \hat{F}_n \text{ exists}\}$$

eventually, and

$$\|\hat{f}_n - f\|_\infty = o_P(1).$$

If, in addition, $f \in C^t(\mathbb{R})$ for some $0 < t \le T$, then also

$$\|\hat{f}_n - f\|_\infty = O_P\left(\left(\frac{\log n}{n}\right)^{t/(2t+1)}\right).$$

The paper [19] is entirely devoted to the proof of this theorem, and it would require another lecture course to be covered in all detail. But we should note that it heavily uses Talagrand's inequality at several instances. In [20] a more concrete construction of the estimator $\hat{F}_n$ is given by using wavelet theory.

A Concentration inequalities for Wigner random matrices

The goal of random matrix theory is the study of the statistical properties of the eigenvalues of an $N \times N$ matrix, whose entries are random variables with a given probability law, in the limit $N \to \infty$.

Here, I am going to discuss so called Wigner random matrices, whose entries are, up to some symmetry constraints, independent and identically distributed random variables. In particular, I am going to consider hermitian Wigner matrices (but the results I will present can be easily extended to real symmetric matrices).

Definition 4. *A hermitian Wigner matrix is an $N \times N$ matrix $H = (h_{ij})_{1 \leq i \leq j \leq N}$ such that*

$$\begin{aligned} h_{ij} &= \frac{1}{\sqrt{N}}(x_{ij} + iy_{ij}) \qquad \textit{for all } 1 \leq i < j \leq N \\ h_{ii} &= \frac{x_{ii}}{\sqrt{N}} \qquad \textit{for } 1 \leq i \leq N \end{aligned}$$

where $\{x_{ij}, y_{ij}, x_{ii}\}$ is a collection of real independent identically distributed random variables with $\mathbb{E}\, x_{ij} = 0$ and $\mathbb{E}\, x_{ij}^2 = 1/2$.

Remark 4. *The diagonal element are often assumed to have a different distribution, with $\mathbb{E}\, x_{ii} = 0$ and $\mathbb{E}\, x_{ii}^2 = 1$. All results I will present remain true with this different convention.*

Remark 5. *Note that the entries scale with the dimension N. The scaling is chosen so that, in the limit $N \to \infty$, all eigenvalues of H remain bounded. To see that this is the correct scaling, observe that*

$$\mathbb{E} \sum_{\alpha=1}^{N} \lambda_\alpha^2 = \mathbb{E}\, Tr\, H^2 = \mathbb{E} \sum_{ij} |h_{ij}|^2 = N^2 \mathbb{E}\, |h_{ij}|^2$$

where λ_α, for $\alpha = 1, \dots, N$ are the eigenvalues of H. If all λ_α stay bounded and of order one in the limit of large N, we must have $\mathbb{E}\, Tr H^2 \simeq N$ and therefore $\mathbb{E}\, |h_{ij}|^2 \simeq N^{-1}$.

Example. The Guassian Unitary Ensemble (GUE) has probability density given by

$$P(H)\mathrm{d}H = \mathrm{const}\, e^{-\frac{N}{2}\operatorname{Tr} H^2}\, \mathrm{d}H$$

where $\mathrm{d}H = \prod_{i<j}^{N} \mathrm{d}h_{ij}\mathrm{d}h_{ij}^{*} \prod_{i=1}^{N} \mathrm{d}h_{ii}$ is the Lebesgue measure on $\mathbb{R}^{N^2}$. It is simple to check that the Guassian Unitary Ensemble is an ensemble of hermitian Wigner matrices, where the entries $\{x_{ij}, y_{ij}\}$ have probability density function $g(x) = \text{const} \cdot e^{-x^2}$ and the diagonal entries $\{x_{ii}\}$ have probability density function $\widetilde{g}(x) = \text{const} \cdot e^{-x^2/2}$.

A.1 Wigner's semicircle law

The first rigorous result in random matrix theory was proven by Wigner in 1955. Wigner proved the convergence of the density of the eigenvalues (sometimes called denstiy of states) towards the famous semicircle law. Fix $E \in \mathbb{R}$, and consider the interval of size $\eta > 0$ centered at E, $I_\eta = [E - \eta/2, E + \eta/2]$. Let $\mathcal{N}[I_\eta]$ denote the number of eigenvalues of H inside I_η. The density of eigenvalues in the interval I_η is then given by $\mathcal{N}[I_\eta]/N\eta$ (we divide by N to obtain a quantity of order one, because the typical distance between eigenvalues is of order $1/N$). Wigner showed that, for every $\delta > 0$,

$$\lim_{\eta \to 0} \lim_{N \to \infty} \mathbb{P}\left(\left| \frac{\mathcal{N}[I_\eta]}{N\eta} - \rho_{sc}(E) \right| \geq \delta \right) = 0 \tag{82}$$

where $\rho_{sc}(E) = (2\pi)^{-1}\sqrt{1 - E^2/4}$ for $|E| \leq 2$ and $\rho_{sc}(E) = 0$ otherwise.

Remark 6. *More generally, Wigner proved that, for every fixed $\eta > 0$ and $\delta > 0$,*

$$\lim_{N \to \infty} \mathbb{P}\left(\left| \frac{\mathcal{N}[I_\eta]}{N\eta} - \int_{E-\eta/2}^{E+\eta/2} \rho_{sc}(s)\mathrm{d}s \right| \geq \delta \right) = 0\,.$$

Remark 7. *Note that the semicircle law is independent of the particular law of the matrix entries (under the assumption that $\mathbb{E}x_{ij} = 0$ and $\mathbb{E}x_{ij}^2 = 1/2$).*

Remark 8. *Note that Wigner's result concerns the density of states (density of eigenvalues) in intervals containing typically order N eigenvalues. In (82), the order of the limit is important. It tells us that we let $N \to \infty$ keeping η fixed. In this sense, Wigner's result is about the convergence to the semicircle law on macroscopic intervals.*

Several questions emerge naturally from Wigner's result. Is it possible to show convergence to the semicircle law for the density on smaller intervals, containing less than order N eigenvalues? Is it possible to prove that the density of states concentrates around its mean value? How large are the fluctuation of the density around the semicircle law or around its average value?

A.2 A general concentration inequality for eigenvalues

The next theorem is a first (but, as we will see, not final) answer to these questions.

Theorem 20 (Guionnet, Zeitouni; see [23]). *Suppose that the law of the entries $\{x_{ij}, y_{ij}, x_{ii}\}$ satisfies the logarithmic Sobolev inequality with constant $c > 0$. Then, for any Lipshitz function $f : \mathbb{R} \to \mathbb{C}$, and $\delta > 0$, we have that*

$$\mathbb{P}\left(|Tr f(H) - \mathbb{E}\, Tr f(H)| \geq \delta N\right) \leq 2e^{-\frac{N^2\delta^2}{4c|f|^2_{\mathcal{L}}}}. \tag{83}$$

Moreover, for any $k = 1, \ldots, N$, we have

$$\mathbb{P}\left(|f(\lambda_k(H)) - \mathbb{E}\, f(\lambda_k(H))| \geq \delta\right) \leq 2e^{-\frac{N\delta^2}{4c|f|^2_{\mathcal{L}}}}. \tag{84}$$

In order to prove this theorem, we want to use the observation of Herbst that Lipshitz functions of random matrices satisfying the log-Sobolev inequality exhibit Gaussian concentration. This result was stated and proved in Theorem 9, adn we rephrase it here for the convenience of the reader:

Theorem 21 (Herbst). *Suppose that P satisfies the log-Sobolev inequality on $\mathbb{R}^M$ with constant c. Let $G : \mathbb{R}^M \to \mathbb{R}$ be a Lipshitz function with constant $|G|_{\mathcal{L}}$. Then, for every $\delta > 0$,*

$$\mathbb{P}\left(|G(x) - \mathbb{E}_P\, G(x)| \geq \delta\right) \leq 2e^{-\frac{\delta^2}{2c|G|^2_{\mathcal{L}}}}.$$

Since $\mathrm{Tr} f(H) = \sum_\alpha f(\lambda_\alpha)$, we see that $\mathrm{Tr} f(H)$ is a Lipshitz function of the eigenvalues of H, if f is Lipshitz. The question is whether or not the eigenvalues of H are Lipshitz function of the matrix entries. If yes, then $\mathrm{Tr} f(H)$ is a Lipshitz function of the matrix entries, and concentration follows by Theorem 21. In other words, to complete the proof of Theorem 20, we need to show that the eigenvalues of H are Lipshitz functions of its entries. To this end, we will use the following lemma.

Lemma 8 (Hoffman-Wielandt). *Let A, B be $N \times N$ hermitian matrices, with eigenvalues $\lambda_1^A \leq \lambda_2^A \leq \cdots \leq \lambda_N^A$ and $\lambda_1^B \leq \lambda_2^B \leq \cdots \leq \lambda_N^B$. Then*

$$\sum_{i=1}^N |\lambda_j^A - \lambda_j^B|^2 \leq Tr(A-B)^2.$$

Proof. Since $\sum_{i=1}^N (\lambda_i^A)^2 = \mathrm{Tr}\, A^2$ and $\sum_{i=1}^N (\lambda_i^B)^2 = \mathrm{Tr}\, B^2$, it is enough to show that

$$\mathrm{Tr}\, AB \leq \sum_{i=1}^N \lambda_i^A \lambda_i^B.$$

Suppose now that $A = U_A D_A U_A^*$, $B = U_B D_B U_B^*$, with U_A, U_B unitaries and $D_A = \operatorname{diag}(\lambda_1^A, \ldots, \lambda_N^A)$, and $D_B = \operatorname{diag}(\lambda_1^B, \ldots, \lambda_N^B)$. Then, introducing the unitary matrix $V = U_A^* U_B$, we find that

$$\begin{aligned}
\operatorname{Tr} AB = \operatorname{Tr} D_A V D_B V^* &= \sum_{i,j} \lambda_i^A \lambda_j^B \, |v_{ij}|^2 \\
&\leq \max \left\{ \sum_{i,j} \lambda_i^A \lambda_j^B \, w_{ij} : w_{ij} \geq 0, \quad \sum_i w_{ij} = 1 \text{ for all } j, \sum_j w_{ij} = 1 \text{ for all } i \right\}.
\end{aligned}$$

To conclude the proof of the lemma, it suffices to show that $W = 1$ is a maximizer. To this end, let $W = (w_{ij})$ be any maximizer. If $w_{11} \neq 0$, there exist j, k s.t. $w_{1j} > 0$ and $w_{k1} > 0$. Let $\nu = \min(w_{1j}, w_{k1})$ and define a new matrix $\widetilde{W} = (\widetilde{w}_{ij})$ by

$$\begin{aligned}
\widetilde{w}_{11} &= w_{11} + \nu, \qquad & \widetilde{w}_{kj} &= w_{kj} + \nu, \\
\widetilde{w}_{1j} &= w_{1j} - \nu, \qquad & \widetilde{w}_{k1} &= w_{k1} - \nu
\end{aligned}$$

and $\widetilde{w}_{\ell,m} = w_{\ell,m}$ for all remaining ℓ, m. Then, we observe that

$$\begin{aligned}
\sum_{i,j} \lambda_i^A \lambda_i^B (\widetilde{w}_{ij} - w_{ij}) &= \nu \left(\lambda_1^A \lambda_1^B + \lambda_k^A \lambda_j^B - \lambda_1^A \lambda_j^B - \lambda_k^A \lambda_1^B \right) \\
&= \nu (\lambda_1^A - \lambda_k^A)(\lambda_1^B - \lambda_j^B) \geq 0 \,.
\end{aligned}$$

Hence $\widetilde{W}$ is also maximal. Repeating this procedure at most $2N - 2$ times we arrive at a maximizer with $w_{11} = 1$. Repeating the same procedure for all diagonal elements, we show that $W = 1$ is a maximizer. □

Corollary 3. *Let $X = (\{x_{ij}, y_{ij}, x_{ii}\}) \in \mathbb{R}^{N^2}$ and let $\lambda_\alpha(X)$, $1 \leq \alpha \leq N$ be the eigenvalues of the Wigner matrix $H = H(X)$. Let $g : \mathbb{R}^N \to \mathbb{R}$ be Lipshitz with constant $|g|_{\mathcal{L}}$. Then the map $\mathbb{R}^{N^2} \ni X \to g(\lambda_1(X), \ldots, \lambda_N(X)) \in \mathbb{R}$ is Lipshitz with coefficient $\sqrt{2/N}\,|g|_{\mathcal{L}}$. In particular if $f : \mathbb{R} \to \mathbb{R}$ is Lipshitz with constant $|f|_{\mathcal{L}}$, the map $\mathbb{R}^{N^2} \ni X \to Tr\, f(H)$ is Lipshitz with constant $\sqrt{2}|f|_{\mathcal{L}}$.*

Proof. Let $\Lambda = (\lambda_1, \ldots, \lambda_N)$. Observe that

$$\begin{aligned}
|g(\Lambda(X)) - g(\Lambda(X'))| &\leq |g|_{\mathcal{L}} \, \|\Lambda(X) - \Lambda(X')\|_2 = |g|_{\mathcal{L}} \sqrt{\sum_{i=1}^N |\lambda_i(X) - \lambda_i(X')|^2} \\
&\leq |g|_{\mathcal{L}} \sqrt{\operatorname{Tr}(H(X) - H(X'))^2} = |g|_{\mathcal{L}} \sqrt{\sum_{i,j} |h_{ij}(X) - h_{ij}(X')|^2} \\
&\leq \sqrt{2/N} \, |g|_{\mathcal{L}} \|X - X'\|_{\mathbb{R}^{N^2}} \,.
\end{aligned} \tag{85}$$

Since $g(\Lambda) := \operatorname{Tr} f(H) = \sum_{j=1}^N f(\lambda_j)$ is such that

$$|g(\Lambda) - g(\Lambda')| \leq |f|_{\mathcal{L}} \sum_{j=1}^N |\lambda_j - \lambda'_j| \leq \sqrt{N}\, |f|_{\mathcal{L}} \, \|\Lambda - \Lambda'\|_{\mathbb{R}^N}$$

we see that g is a Lipshitz function on $\mathbb{R}^N$ with constant $\sqrt{N}|f|_{\mathcal{L}}$. Combined with (85), this completes the proof of the corollary. □

We are now ready to show Theorem 20.

Proof of Theorem 20. Let $X = (\{x_{ij}, y_{ij}, x_{ii}\}) \in \mathbb{R}^{N^2}$. Let $G(X) = \operatorname{Tr} f(H(X))$. Then G is Lipshitz with constant $\sqrt{2}|f|_{\mathcal{L}}$. Hence, by Theorem 21, we find

$$\mathbb{P}\left(|\operatorname{Tr} f(H) - \mathbb{E}\operatorname{Tr} f(H)| \geq \delta N\right) \leq 2e^{-\frac{\delta^2 N^2}{4c|f|^2_{\mathcal{L}}}}.$$

To prove (84), we observe that, by Corollary 3, the function $G(X) = f(\lambda_k(X))$ is Lipshitz with constant $\sqrt{2/N}\, |f|_{\mathcal{L}}$. Hence (84) follows by Theorem 21. □

Applications of Theorem 20. From (84), choosing $f(s) = s$, we find immediately that, for any $j = 1, \dots, N$,

$$\mathbb{P}\left(|\lambda_j - \mathbb{E}\,\lambda_j| \geq \delta\right) \leq 2e^{-\frac{N\delta^2}{4c}}\,. \tag{86}$$

This inequality implies that the fluctuations of the j-th eigenvalue around its average value are at most of the order $N^{-1/2}$. Since the distance between eigenvalue is much smaller we actually expect the fluctuations to be much smaller. In fact, it is known that, if λ_N denotes the largest eigenvalue of H,

$$\lim_{N\to\infty} \mathbb{P}\left(N^{2/3}(\lambda_N - 2) \geq s\right) = F_{\mathrm{TW}}(s) \tag{87}$$

where the Tracy-Widom distribution F_{TW} is given by

$$F_{\mathrm{TW}}(s) = e^{-\int_s^\infty \mathrm{d}x\,(x-s)q^2(x)}$$

with q being the solution of the equation $q''(x) = xq(x) + q^3(x)$ with $q(x) \simeq \operatorname{Ai}(x)$ as $x \to \infty$ (Ai is the Airy function). The convergence to the Tracy Widom distribution was first proven by Tracy-Widom for GUE, and then extended by Soshnikov (see [35] and references therein) to a large class of Wigner matrices. Eq. (87) implies in particular that the fluctuations of λ_N are of order $N^{-2/3}$. In the bulk of the spectrum, fluctuations should be even smaller, since eigenvalues are closer. For GUE, it was proven by Gustavsson (see [24]) that, as $N \to \infty$

$$\frac{\lambda_j - t(j)}{\frac{\log^{1/2} N}{4(1-t(j)^2)N}} \to N(0,1)$$

in distribution. Here $t(j)$ is the location, according to the semicircle law, of the j-th eigenvalue of H, and $j = j(N)$ is chosen so that $j(N)/N \to a \in (0,1)$ as $N \to \infty$. Hence in this case the fluctuations of λ_j are of the order $(\log N)^{1/2}/N$.

Theorem 20 can also be used to get concentration of the density of states. There is here a little obstacle, which follows by the observation that

$$\frac{\mathcal{N}[I_\eta]}{N\eta} = \frac{1}{N\eta}\mathrm{Tr}\,\chi(|H-E| \leq \eta)$$

and that the characteristic function $\chi(|x-E| \leq \eta$ is not Lipshitz. To circumvent this problem, it is useful to approximate the density of states by the imaginary part of the trace of the resolvent. The idea here is that $\chi(|x| \leq \eta) \simeq \eta^2/(x^2+\eta^2)$. This leads to

$$\frac{\mathcal{N}[I_\eta]}{N\eta} \simeq \frac{1}{N\eta}\mathrm{Tr}\,\frac{\eta^2}{(H-E)^2+\eta^2} = \frac{1}{N}\mathrm{Im}\ \mathrm{Tr}\,\frac{1}{H-E-i\eta}\,.$$

It is a fact that concentration bounds for the r.h.s. can then be translated into concentration bounds for the density of states $\mathcal{N}[I_\eta]/N\eta$. We will not go into these details; instead, we will look for concentration bounds for the imaginary part of the trace of $(H-E-i\eta)^{-1}$.

Taking $f(s) = (s-E-i\eta)^{-1}$, it follows that f is Lipshitz with constant η^{-2}. Therefore, (83) implies that

$$\mathbb{P}\left(\left|\frac{1}{N}\mathrm{Im}\ \mathrm{Tr}\frac{1}{H-E-i\eta} - \mathbb{E}\frac{1}{N}\mathrm{Im}\ \mathrm{Tr}\frac{1}{H-E-i\eta}\right| \geq \delta\right) \leq 2e^{-\frac{\delta^2N^2\eta^4}{4c}}\,. \qquad (88)$$

This immediately implies that the fluctuations of the imaginary part of the resolvent at distances $\eta \simeq 1$ from the real axis are of order N^{-1} with Gaussian tails (the same bound can then be obtained for the density of states on intervals of size $\eta \simeq 1$). This result is optimal. Note that (88) implies concentration also for the density of states on intervals of size $N^{-1/2} \ll \eta \ll 1$, which do not contain a macroscopic number of eigenvalues. The estimate (88), on the other hand, does not say anything about the density of states on intervals of size $\eta \ll N^{-1/2}$.

A.3 Concentration of eigenvalues at the right scale

This remark leads to two natural questions. Is it possible to establish convergence to the semicircle law for the density of states on scales $N^{-1/2} \ll \eta \ll 1$ for which we know (by (88)) that the density concentrates around its average? Is it possible to get concentration and convergence to the semicircle for $\eta \ll N^{-1/2}$? As long as $\eta \gg N^{-1}$, the interval I_η contains a large number of eigenvalues (which converges to infinity as $N \to \infty$). For this reason, we should expect that, for $\eta \gg N^{-1}$, concentration around the average and convergence to the semicircle law hold. We will see that this is indeed the case. The first ingredient to prove these facts is an upper bound on the density of states on small intervals.

Theorem 22. *Suppose that the random variables $\{x_{ij}, y_{ij}, x_{ii}\}$ have Gaussian decay (in the sense that, for some $\delta > 0$, $\mathbb{E}\, e^{\delta x_{ij}^2} < \infty$). Then there exist constants $K_0 > 0$ and $c, C < \infty$ such that*

$$\mathbb{P}\left(\frac{\mathcal{N}[E - \eta/2, E + \eta/2]}{N\eta} \geq K \right) \leq C e^{c\sqrt{KN\eta}} \tag{89}$$

for all $E \in \mathbb{R}$, $\eta = \eta(N) \geq (\log N)^2/N$, $N \geq 2$, and $K \geq K_0$.

Remark 9. *Under somehow stronger assumptions on the law of $\{x_{ij}, y_{ij}, x_{ii}\}$, one can improve (89) to*

$$\mathbb{P}\left(\frac{\mathcal{N}[E - \eta/2, E + \eta/2]}{N\eta} \geq K \right) \leq C e^{cKN\eta} .$$

Remark 10. *It is possible to extend these bounds to the case $\eta(N) = K/N$ for a large, but fixed K (independent of N). Intervals of size K/N typically contain a finite number of eigenvalues.*

Proof of Theorem 22. We use the inequality (with $I_\eta = [E - \eta/2, E + \eta/2]$)

$$\frac{\mathcal{N}[I_\eta]}{N\eta} \leq \frac{C}{N} \text{Im Tr} \frac{1}{H - E - i\eta} = \frac{C}{N} \text{Im} \sum_{j=1}^{N} \frac{1}{H - E - i\eta}(j,j) . \tag{90}$$

Now observe that, for example for $j = 1$,

$$\frac{1}{H - E - i\eta}(1,1) = \frac{1}{h_{11} - z - a \cdot (B - z)^{-1} a}$$

where $a = (h_{21}, \dots, h_{N-1,1}) \in \mathbb{C}^{N-1}$ is the first column of H, after removing the diagonal element h_{11}, and B is the $(N-1) \times (N-1)$ minor of H, obtained by removing the first line and column. If we denote by μ_α and v_α the eigenvalues and , respectively, the eigenvectors of B, we find that

$$\frac{1}{H - E - i\eta}(1,1) = \frac{1}{h_{11} - z - \sum_\alpha \frac{|a \cdot v_\alpha|^2}{\mu_\alpha - z}} .$$

Note that a is independent of μ_α and v_α. Therefore,

$$\mathbb{E}\, |a \cdot v_\alpha|^2 = \mathbb{E} \sum_{i,j} a_i \bar{a}_j \overline{v_\alpha}(i) v_\alpha(j) = N^{-1} .$$

We define $\xi_\alpha = N|a \cdot v_\alpha|^2$. Then we have $\mathbb{E}\xi_\alpha = 1$ and

$$\frac{1}{H - E - i\eta}(1,1) = \frac{1}{h_{11} - z - \frac{1}{N}\sum_\alpha \frac{\xi_\alpha}{\mu_\alpha - z}} \,. \tag{91}$$

Taking the absolute value, we find that

$$\begin{aligned} \operatorname{Im} \frac{1}{H - E - i\eta}(1,1) &\leq \frac{1}{\eta + \frac{\eta}{N}\sum_\alpha \frac{\xi_\alpha}{(\mu_\alpha - E)^2 + \eta^2}} \\ &\leq \frac{N\eta}{\sum_{\alpha : |\mu_\alpha - E| \leq \eta} \xi_\alpha} \,. \end{aligned} \tag{92}$$

Now we observe that, since the eigenvalues of H are interlaced with the eigenvalues of B, $|\{\alpha : |\mu_\alpha - E| \leq \eta\}| \geq \mathcal{N} - 1$. On the other hand, we observe that, for any $m = 1, \dots, N$,

$$\sum_{j=1}^m \xi_{\alpha_j} = \sum_{j=1}^m |b \cdot v_{\alpha_j}|^2 = \|P_m b\|^2$$

where $b = \sqrt{N}\, a$ is a vector in $\mathbb{C}^{N-1}$ with independent and identically distributed components such that $\mathbb{E}\, b_j = 0$ and $\mathbb{E}\, |b_j|^2 = 1$, and P_m is the orthogonal projection of rank m, projecting into the space spanned by the v_{α_j}, $j = 1, \dots, m$ (remark that the b_j are independent of the v_{α_j}). It is simple to check that $\mathbb{E}\, \|P_m b\|^2 = m$. We need to know that the probability for $\|P_m b\| \leq m/2$ is very small. Concentration estimates for quadratic forms are well-known. To get the desired bound, we can use the following lemma proven by Hanson-Wright, and extended by Wright to variable which are not necessarily symmetric.

Lemma 9 (Hanson-Wright, Wright, see [25, 40])**.** *Let X_i, $i = 1, \dots, N$ be a sequence of independent and identically distributed random variables with $\mathbb{E}X_i = 0$. There exists a constant $C > 0$ such that*

$$\mathbb{P}\left(\sum_{i,j=1}^N a_{ij}\left(X_i X_j - \mathbb{E}X_i X_j\right) \geq \delta\right) \leq e^{-C \min\left(\delta/\|A\|, \delta^2, \|A\|_{HS}^2\right)}$$

where $\|A\|$ denotes the operator norm of the matrix $A = (|a_{ij}|)_{i<j}$ and $\|A\|_{HS} = (\sum_{i,j} |a_{ij}|^2)^{1/2}$ is its Hilbert-Schmidt norm.

Using this lemma, with $A = (|\sum_\alpha v_\alpha(i)\overline{v}_\alpha(j)|)$, we find $\|A\| \leq \sqrt{m}$ and $\|A\|_{\mathrm{HS}}^2 \leq m$, and thus

$$\mathbb{P}\left(\left|\|P_m b\|^2 - m\right| \geq \delta\right) \leq e^{-C \min\left(\frac{\delta}{m}, \frac{\delta^2}{m^2}\right)} .$$

With $\delta = m/2$, we find

$$\mathbb{P}\left(\|P_m b\|^2 \leq m/2\right) \leq e^{-C\sqrt{m}}$$

since we are interested in the regime with $m \gg 1$.

From (90) and (92), we find

$$\frac{\mathcal{N}[I_\eta]}{N\eta} \leq \frac{C}{N} \sum_{j=1}^{N} \frac{N\eta}{\sum_{\alpha:|\mu_\alpha^{(j)}-E|\leq\eta} \xi_\alpha^{(j)}}$$

where $\xi_\alpha^{(j)} = \sqrt{N}|a^{(j)} \cdot v_\alpha^{(j)}|^2$, and $a^{(j)}$ is the j-th column of H without the diagonal entry h_{jj}, $\mu_\alpha^{(j)}$, $v_\alpha^{(j)}$ are the eigenvalues and eigenvectors of the minor $B^{(j)}$ of H obtained by removing the j-th line and the j-th column. Therefore

$$\begin{aligned} \mathbb{P}\left(\frac{\mathcal{N}[I_\eta]}{N\eta} \geq K\right) &\leq \mathbb{P}\left(\mathcal{N}[I_\eta] \geq KN\eta \quad \text{and } \exists j = 1, \dots, N : \sum_{\alpha:|\mu^{(j)}-E|\leq\eta} \xi_\alpha^{(j)} \leq \frac{KN\eta}{2}\right) \\ &= N\mathbb{P}\left(\mathcal{N}[I_\eta] \geq KN\eta \quad \text{and} \sum_{\alpha:|\mu^{(1)}-E|\leq\eta} \xi_\alpha^{(1)} \leq \frac{KN\eta}{2}\right) \\ &\leq Ne^{-C\sqrt{KN\eta}} \leq Ce^{-c\sqrt{KN\eta}} \end{aligned}$$

where we used the fact that $KN\eta \geq K(\log N)^2$, for a sufficiently large K, to absorb the factor of N in the exponential. □

Using the upper bound on the density given by Theorem 22, it is immediately possible to improve the concentration estimate (88). Let

$$f\left(\{x_{ij}, y_{ij}, x_{ii}\}\right) = \frac{1}{N} \mathrm{Tr}\, \frac{1}{H - E - i\eta}\,.$$

Assuming that the entries satisfy a Poincaré inequality, we have that

$$\begin{aligned}
\text{var }(f) &= \mathbb{E}\left|\frac{1}{N}\text{Tr}\,\frac{1}{H-E-i\eta} - \mathbb{E}\frac{1}{N}\text{Tr}\,\frac{1}{H-E-i\eta}\right|^2 \\
&\le C\sum_{i<j}\left\{\left|\frac{\partial}{\partial x_{ij}}\left(\frac{1}{N}\sum_\alpha \frac{1}{\lambda_\alpha - E - i\eta}\right)\right|^2 + \left|\frac{\partial}{\partial y_{ij}}\left(\frac{1}{N}\sum_\alpha \frac{1}{\lambda_\alpha - E - i\eta}\right)\right|^2\right\} \\
&\quad + \sum_i \left|\frac{\partial}{\partial x_{ii}}\left(\frac{1}{N}\sum_\alpha \frac{1}{\lambda_\alpha - E - i\eta}\right)\right|^2 \\
&= \frac{C}{N^2}\sum_{\alpha,\beta}\frac{1}{(\lambda_\alpha - E - i\eta)^2(\lambda_\beta - E + i\eta)^2} \\
&\qquad \times \left\{\sum_{i<j}\frac{\partial\lambda_\alpha}{\partial x_{ij}}\overline{\frac{\partial\lambda_\beta}{\partial x_{ij}}} + \frac{\partial\lambda_\alpha}{\partial y_{ij}}\overline{\frac{\partial\lambda_\beta}{\partial y_{ij}}} + \sum_i \frac{\partial\lambda_\alpha}{\partial x_{ii}}\overline{\frac{\partial\lambda_\beta}{\partial x_{ii}}}\right\}.
\end{aligned}$$

We find

$$\sum_{i<j}\frac{\partial\lambda_\alpha}{\partial x_{ij}}\overline{\frac{\partial\lambda_\beta}{\partial x_{ij}}} + \frac{\partial\lambda_\alpha}{\partial y_{ij}}\overline{\frac{\partial\lambda_\beta}{\partial y_{ij}}} + \sum_i \frac{\partial\lambda_\alpha}{\partial x_{ii}}\overline{\frac{\partial\lambda_\beta}{\partial x_{ii}}} = \frac{1}{N}\sum_{i,j} v_\alpha(i)\overline{v}_\alpha(j)v_\beta(j)\overline{v}_\alpha(i) = \frac{\delta_{\alpha,\beta}}{N}$$

and therefore

$$\text{var }(f) \le \frac{C}{N^3}\sum_\alpha \frac{1}{|\lambda_\alpha - E - i\eta|^4}\,. \tag{93}$$

Using the trivial bound $|\lambda_\alpha - E - i\eta| \ge \eta$, we find

$$\text{var }(f) \le \frac{C}{N^2\eta^4} \tag{94}$$

which implies that the fluctuations of the density of states are small as long as $\eta \gg N^{-1/2}$; this is the same result as in (88). But now, using the upper bound on the density, we can improve (94). In fact, we only have to use the bound $|\lambda_\alpha - E - i\eta| \ge \eta$ for those α for which $|\lambda_\alpha - E| \le \eta$; the upper bound implies that, with very high probability, there are not more than $CN\eta$ such α's. Hence, making use of the upper bound, we find

$$\text{var }(f) \le \frac{CN\eta}{N^3\eta^4} = \frac{C}{N^2\eta^3} \tag{95}$$

which is small, for $\eta \gg N^{-2/3}$. This argument can be made rigorous using a dyadic decomposition (see Theorem 3.1 in [14]). Note, also, that the bound (95) can be exponentiated, if we assume that the entries satisfy a log-Sobolev inequality (the simplest proof consists in applying an inequality by Bobkov-Götze, see [3]). We find (see Theorem 3.1 in [14])

$$\mathbb{P}\left(\left|\frac{1}{N}\mathrm{Tr}\,\frac{1}{H-E-i\eta}-\mathbb{E}\,\frac{1}{N}\mathrm{Tr}\,\frac{1}{H-E-i\eta}\right|\geq\varepsilon\right)\leq e^{-cN\eta\varepsilon\,\min\{(\log N)^{-1},N\eta^2\varepsilon\}} \tag{96}$$

for every $\varepsilon>0$ which implies concentration for all $\eta\gg N^{-2/3}$.

The concentration bound (96) is still not optimal. As mentioned above, one expect concentration of the density of states (and convergence to the semicircle law) for arbitrary intervals of size $\eta\gg N^{-1}$. It turns out that the right approach to get concentration on these small scales consists in proving first the convergence to the semicircle law. For intervals away from the edges, the convergence to the semicircle law on these small scales is established in the following theorem.

Theorem 23. *[Theorem 4.1 in [15]] Assume that the random variables $\{x_{ij},y_{ij},x_{ii}\}$ have Gaussian decay at infinity. Let*

$$m_N(z)=\frac{1}{N}Tr\frac{1}{H-z}$$

and

$$m_{sc}(z)=\int\mathrm{d}y\,\frac{\rho_{sc}(y)}{y-z}\,.$$

Then, for any $\kappa>0$ there exist $C,c<\infty$ such that

$$\mathbb{P}\left(\sup_{E\in(-2+\kappa,2-\kappa)}|m_N(E+i\eta)-m_{sc}(E+i\eta)|\geq\delta\right)\leq Ce^{-c\delta\sqrt{N\eta}}$$

for all $\delta\leq C\kappa$, $(\log N)^4/N\leq\eta\leq 1$, $N\geq 2$.

Remark 11. *m_N, m_{sc} are known as the Stieltjes transforms of the empirical eigenvalue measure*

$$\mu_N(x)=\frac{1}{N}\sum_\alpha\delta(\lambda_\alpha-x)$$

and, respectively, of the semicircle law. The convergence of the Stieltjes transform implies convergence for the density of states. We find (see Theorem 4.1 in [15])

$$\mathbb{P}\left(\sup_{E\in(-2+\kappa,2-\kappa)}\left|\frac{\mathcal{N}[E-\eta/2,E+\eta/2]}{N\eta}-\rho_{sc}(E)\right|\geq\delta\right)\leq Ce^{-c\delta^2\sqrt{N\eta}}\,.$$

Remark 12. *In particular, Theorem 23 implies that*

$$|\mathbb{E}m_N(z) - m_{sc}(z)| \le \frac{C}{\sqrt{N\eta}}$$

for all $z = E + i\eta$ *with* $E \in (-2+\kappa, 2-\kappa)$ *and* $(\log^4 N)/N \le \eta \le 1$*. Therefore, we obtain the concentration bound*

$$\mathbb{P}\left(\sup_{E \in (-2+\kappa, 2-\kappa)} |m_N(E+i\eta) - \mathbb{E}m_N(E+i\eta)| \ge \delta \right) \le Ce^{-c\delta\sqrt{N\eta}}\,. \tag{97}$$

Remark 13. *It is possible, with some more work, to extend the result of Theorem 23 and the concentration bound (97) to the microscopic scale* $\eta = K/N$*, with large but fixed* K*; see Theorem 3.1 in [15]. This scale is then optimal; for* $\eta \le N^{-1}$*, the typical number of eigenvalues in the interval* $I_\eta = [E - \eta/2, E + \eta/2]$ *is very small and the fluctuations of the density are certainly important.*

A complete proof of Theorem 23 can be found in [15]. The main idea of the proof is that the Stieltjes transform of the semicircle law satisfy a self-consistent equation

$$m_{\text{sc}}(z) + \frac{1}{z + m_{\text{sc}}(z)} = 0$$

which is stable away from the edges. This implies that to prove that $|m_N(z) - m_{\text{sc}}(z)|$ is small , it is enough to show that $|m_N(z) + (z + m_N(z))^{-1}|$ is small. This follows making use of expressions like (91). An important ingredient of this proof is the upper bound on the density of states obtained in Theorem 22.

References

[1] A. Araujo, E. Giné. *The central limit theorem for real and Banach-valued random variables.* Wiley. (1980).

[2] Barron, A.; Birgé, L.; Massart, P. Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301-413 (1999).

[3] Bobkov, S. G., Götze, F.: Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *J. Funct. Anal.* **163** (1999), no. 1, 1–28.

[4] V. Bogachev. *Gaussian Measures.* American Mathematical Society, Mathematical Surveys and Monographs 62 (1998).

[5] C. Borell. The Brunn-Minkowski inequality in Gauss space. *Invent. Math.* 30 207-216 (1975).

[6] Bousquet, O. Concentration inequalities for sub-additive functions using the entropy method. In: *Stochastic inequalities and applications., Progr. Probab.* **56**, E. Giné, C. Houdré, D. Nualart, eds., Birkhäuser, Boston, 213-247. (2003).

[7] I. Chavel. *Eigenvalues in Riemaniann geometry.* Academic press. (1984).

[8] I. Chavel. *Isoperimetric inequalities.* Cambridge University Press, Cambridge UK. (2001).

[9] A. Dvoretzky, J. Kiefer; J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **27** 642-669 (1956)

[10] R. M. Dudley. Central limit theorems for empirical measures. *Ann. Probab.* **6** 899-929 (1978).

[11] R. M. Dudley. *Real analysis and probability.* Cambridge University Press, Cambridge UK. (2002).

[12] R.M. Dudley. *Uniform Central Limit Theorems.* Cambridge University Press, Cambridge UK (1999).

[13] Einmahl, U. and Mason, D. M. An empirical process approach to the uniform consistency of kernel-type function estimators. *J. Theoret. Probab.* **13** 1-37. (2000)

[14] Erdős, L., Schlein, B., Yau, H.-T.: Semicircle law on short scales and delocalization of eigenvectors for Wigner random matrices. *Ann. Probab.* **37**, No. 3, 815–852 (2008)

[15] Erdős, L., Schlein, B., Yau, H.-T.: Wegner estimate and level repulsion for Wigner random matrices. To appear in Int. Math. Res. Notices (2008). Preprint arxiv.org/abs/0811.2591

[16] E. Friedgut, G. Kalai, Every monotone graph property has a sharp threshold. *Proc. Amer. Math. Soc.* 124, 1017-1054 (1996).

[17] E. Giné. *Empirical processes and some of their applications.* lecture notes (2007).

[18] Giné, E. and Koltchinskii, V. Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.* **34** 1143-1216. (2006)

[19] Giné, E. and Nickl, R. An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probab. Theory Related Fields* **143** 569-596. (2009).

[20] Giné, E. and Nickl, R. Uniform limit theorems for wavelet density estimators. *Ann. Probab.* 37 1605-1646 (2009).

[21] E. Giné and J. Zinn. Some limit theorems for empirical processes. *Ann. Probab.* **12** 929-989 (1984).

[22] L. Gross. Logarithmic Sobolev inequalities. *Amer. J. Math.* 97 1061-1083 (1975).

[23] Guionnet, A., Zeitouni, O.: Concentration of the spectral measure for large matrices. *Electronic Comm. in Probability* **5** (2000) Paper 14.

[24] Gustavsson, J.: Gaussian fluctuations of eigenvalues in the GUE. *Ann. Inst. Henri Poincaré (B)* **41** (2005), no. 2, 151-178.

[25] Hanson, D.L., Wright, F.T.: A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Math. Stat.* **42** (1971), no.3, 1079-1083.

[26] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13-30 (1963).

[27] I.A. Ibragimov and R.Z. Khasminski (1980). On estimation of distribution density. *Zapiski Nauchnyh Seminarov LOMI.* **98** 61-86.

[28] J.P. Kahane. *Some random series of functions.* Cambridge Univ. Press (1985), first edition 1968.

[29] Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593-2656.

[30] Klein, T. and Rio, E. Concentration around the mean for maxima of empirical processes. *Ann. Probab.* **33** 1060-1077. (2005)

[31] M. Ledoux. *The concentration of measure phenomenon.* American Mathematical Society, Mathematical Surveys and Monographs 89 (2001).

[32] M. Ledoux, M. Talagrand. *Probability in Banach Spaces.* Springer (1991).

[33] P. Massart. About the constants in Talagrand's deviation inequalities for empirical processes. *Ann. Probab.* **28** 863–884 (2000).

[34] P. Massart. *Concentration inequalities and model selection.* Springer (2007).

[35] Soshnikov, A.: Universality at the edge of the spectrum in Wigner random matrices. *Comm. Math. Phys.* **207** (1999), no.3. 697-733.

[36] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Ann. Probab.* **22** 28-76 (1994).

[37] M. Talagrand. A new look at independence. *Ann. Probab.* 24, 1-34 (1996)

[38] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.* 125, 505-563 (1996).

[39] M. Talagrand. *The generic chaining.* Springer (2005).

[40] Wright, F.T.: A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric. *Ann. Probab.* **1** No. 6. (1973), 1068-1070.