Sketch notes on concentration of measure

Quantitative Linear Algebra Tutorials, March 2018,
IPAM, UCLA

I have written these notes to accompany my first two tutorial lectures on concentration of measure given at IPAM during March 20–23, 2018. I intend the lectures to be quite basic, so they will miss out many large and important topics related to concentration. These notes are a guide to my lectures and to further reading, so only a few proofs are given in full here. I am grateful to a number of participants in the tutorials for pointing out corrections.

The two sections below roughly reflect the contents of my first and second lectures, respectively. I also gave a third lecture on some applications of concentration in classical ergodic theory. That subject is not treated here. A nice introduction can be found in [Shi96, Chapters III and IV].

1 Concentration for product measures

1.1 Point of departure: exponential bounds in the LLN

Let $X_1, X_2, \ldots$ be i.i.d. RVs, and suppose for simplicity that

$$\|X_i\|_\infty \leq 1. \tag{1}$$

Let $m = \mathbf{E}[X_i]$, and let $S_n = X_1 + \cdots + X_n$. According to the weak law of large numbers,

$$\mathbf{P}\big(|S_n/n - m| \geq \varepsilon\big) \longrightarrow 0 \quad \forall \varepsilon > 0$$

as $n \longrightarrow \infty$. More is true: given assumption (1), the convergence to zero of these probabilities is exponentially fast. That is, for every $\varepsilon > 0$, there exists $c > 0$ such that

$$\mathbf{P}\big(|S_n/n - m| \geq \varepsilon\big) \leq \mathrm{e}^{-cn}$$

for all sufficiently large n. Thus, a 'macroscopic' deviations of S_n/n from its mean is exponentially unlikely as $n \longrightarrow \infty$. In the presence of (1), such a c can be chosen depending only on ε: for instance, any $c < \varepsilon^2/2$ will do. Such basic quantitative forms of the law of large numbers go back to work of Bernstein and Chernoff.

Remark. In this setting, one can actually show that

$$\mathbf{P}\big(S_n/n - m \geq \varepsilon\big) = \mathrm{e}^{-cn+o(n)} \tag{2}$$

for some particular choice of $c > 0$, which depends more finely on the common distribution of the X_is. This calculation is Cramér's theorem, one of the starting points of large deviations theory. But in these notes our concern is only with upper bounds, and with settings in which no such precise asymptotic as (2) is known.

In general, if $(Y_n)_{n\geq 1}$ is any sequence of $\mathbb{R}$-valued RVs, then they are said to exhibit **concentration** if there are constants $(c_n)_{n\geq 1}$ such that

$$\mathbf{P}(|Y_n - c_n| \geq \varepsilon) \longrightarrow 1 \quad \forall \varepsilon > 0.$$

Often one may take $c_n = \mathbf{E}[Y_n]$, but not always. If in fact there are positive constants $c(\varepsilon)$ such that

$$\mathbf{P}(|Y_n - c_n| \geq \varepsilon) \leq \mathrm{e}^{-c(\varepsilon)n+o(n)}, \tag{3}$$

then one speaks of **exponential concentration**. The inequalities above assert that the running averages S_n/n concentrate exponentially fast around their mean m.

There are many other settings in which it is useful to know that some sequence of RVs exhibits concentration, and a whole branch of probability theory has grown around the phenomenon.

1.2 Other functionals of i.i.d. RVs

Consider again the setting above. For each n, we may write S_n/n as $f(X_1, \ldots, X_n)$, where $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is the averaging function

$$f(x_1, \ldots, x_n) = \frac{x_1 + \cdots + x_n}{n}.$$

It turns out that some quite general properties of this function are enough to imply that the RV $f(X_1, \ldots, X_n)$ is highly concentrated once n is large. This extends the phenomenon of concentration far beyond the particular example of the averaging function.

Heuristically, the key realization in this direction is the following:

One should expect concentration of the RV $f(X_1, \ldots, X_n)$ whenever f has a sufficiently small dependence on each individual coordinate.

In the example of the averaging function, suppose for simplicity that each X_i takes values in $[0, 1]$, so f may be restricted to a function $[0, 1]^n \longrightarrow \mathbb{R}$. If we alter a single coordinate of $(x_1, \ldots, x_n) \in [0, 1]^n$, then the value of $f(x_1, \ldots, x_n)$ changes by at most $1/n$.

It is valuable think about this feature of the averaging function in the following geometric way. First, recall that a map

$$f : (X, d_X) \longrightarrow (Y, d_Y)$$

between metric spaces is L**-Lipschitz** if

$$d_Y(f(x), f(x')) \leq L \cdot d_X(x, x') \quad \forall x, x' \in X.$$

The least constant L for which is holds is the **Lipschitz constant** of f. In case $Y = \mathbb{R}$ with its usual metric, we denote the this Lipschitz constant by $\|f\|_{\mathrm{L}}$. In that case one can check easily that this is a seminorm on the linear space of all Lipschitz functions, and that it vanishes precisely on the constant functions.

Next, for any nonempty set K and $n \in \mathbb{N}$, let us endow the product space K^n with the **normalized Hamming metric**:

$$d_n(\boldsymbol{x}, \boldsymbol{y}) := \frac{|\{i = 1, 2, \ldots, n : \ x_i \neq y_i\}|}{n}$$

for $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ in K^n. With this metric, a function $f : K^n \longrightarrow \mathbb{R}$ is L-Lipschitz provided $|f(\boldsymbol{x} - f(\boldsymbol{y})| \leq L/n$ whenever $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in only one coordinate. Thus, our discussion above concludes that the averaging function is 1-Lipschitz for the normalized Hamming metric on $[0, 1]^n$.

According to the next theorem, this last fact is enough to imply concentration in general.

Theorem 1 (McDiarmid's inequality)**.** *Let $(K, \mathcal{A}, \mu)$ be a probability space, let $L > 0$ be a constant, and let $f : K^n \longrightarrow \mathbb{R}$ be a measurable function for the product σ-algebra $\mathcal{A}^{\otimes n}$ which is L-Lipschitz for the normalized Hamming metric*[1]*. Then*

$$\mu^{\times n}\left\{f \geq \int f \, \mathrm{d}\mu^{\times n} + \varepsilon\right\} \leq \mathrm{e}^{-2\varepsilon^2 n/L^2}.$$

[1]If K is not finite, then the normalized Hamming metric is not separable, and it usually does not have $\mathcal{A}^{\otimes n}$ for its Borel σ-algebra — this is why we must assume measurability separately.

Remark. More probabilistically, this theorem asserts that, if $\boldsymbol{X} = (X_1, \dots, X_n)$ has i.i.d.(μ) coordinates, then

$$\mathbf{P}\big(f(\boldsymbol{X}) - \mathbf{E}[f(\boldsymbol{X})] \geq \varepsilon\big) \leq \mathrm{e}^{-2\varepsilon^2 n/L^2}.$$

By applying Theorem 1 to $-f$, it gives also

$$\mathbf{P}\big(f(\boldsymbol{X}) - \mathbf{E}[f(\boldsymbol{X})] \leq -\varepsilon\big) \leq \mathrm{e}^{-2\varepsilon^2 n/L^2},$$

and hence

$$\mathbf{P}\big(\big|f(\boldsymbol{X}) - \mathbf{E}[f(\boldsymbol{X})]\big| \geq \varepsilon\big) \leq 2\mathrm{e}^{-2\varepsilon^2 n/L^2}.$$

To prove Theorem 1, we bound the probability of interest by bounding an exponential moment and applying Markov's inequality. Specifically, we show that

$$\int \mathrm{e}^f \,\mathrm{d}\mu^{\times n} \leq \mathrm{e}^{\langle f\rangle + \|f\|_{\mathrm{L}}^2/8n} \tag{4}$$

for any measurable and Lipschitz function f on K^n, where $\langle f\rangle$ is a short-hand for $\int f \,\mathrm{d}\mu^{\times n}$. To turn (4) into Theorem 1, let f be as in that theorem, and first observe that we can normalized f and so assume that $\langle f\rangle = 0$ and $\|f\|_{\mathrm{L}} \leq 1$. Having done so, we apply (4) to the function $g := \lambda n f$ for some $\lambda > 0$, which has $\|g\|_{\mathrm{L}} \leq \lambda n$. Combined with Markov's inequality, this results in

$$\mu^{\times n}\big\{f \geq \varepsilon\big\} = \mu^{\times n}\big\{\mathrm{e}^g \geq \mathrm{e}^{\varepsilon\lambda n}\big\} \leq \mathrm{e}^{-\varepsilon\lambda n}\int \mathrm{e}^g \,\mathrm{d}\mu^{\times n} \leq \mathrm{e}^{-\varepsilon\lambda n}\mathrm{e}^{\lambda^2 n/8} = \mathrm{e}^{(-\varepsilon\lambda+\lambda^2/8)n}.$$

Optimizing, we are led to use $\lambda = 4\varepsilon$. This choice gives the desired bound $\mathrm{e}^{-2\varepsilon^2 n}$.

Inequality (4) is proved by induction on n. First, define the operator P_n from functions on K^n to functions on K^{n-1}:

$$P_n f(x_1, \dots, x_{n-1}) := \int_K f(x_1, \dots, x_{n-1}, x')\,\mu(\mathrm{d}x'). \tag{5}$$

This has a simple probabilistic interpretation. For $0 \leq i \leq n$, let $\mathcal{F}_i$ be the sigma-subalgebra of sets in $\mathcal{A}^{\otimes n}$ which depend on only the first i coordinates. Then the function

$$(x_1, \dots, x_n) \mapsto P_n f(x_1, \dots, x_{n-1})$$

is a version of the conditional expectation $E_{\mu^{\times n}}(f \,|\, \mathcal{F}_{n-1})$. In particular, we clearly have

$$\int P_n f \,\mathrm{d}\mu^{\times(n-1)} = \int f \,\mathrm{d}\mu^{\times n}.$$

Lemma 2 (Decay in Lipschitz constant)**.** *We have*

$$\|P_n f\|_{\mathrm{L}} \le \frac{n-1}{n}\|f\|_{\mathrm{L}}.$$

Proof. If $\boldsymbol{y}_1, \boldsymbol{y}_2 \in K^{n-1}$ differ in d coordinates, then their distance in K^{n-1} is $d/(n-1)$. On the other hand, for any fixed $x' \in K$, the distance in K^n between $(\boldsymbol{y}_1, x')$ and $(\boldsymbol{y}_2, x')$ is d/n. Therefore

$$\begin{aligned}|P_n f(\boldsymbol{y}_1) - P_n f(\boldsymbol{y}_2)| &\le \int |f(\boldsymbol{y}_1, x') - f(\boldsymbol{y}_2, x')|\,\mu(\mathrm{d}x') \\ &\le \|f\|_{\mathrm{L}} \cdot \frac{d}{n} = \frac{n-1}{n}\|f\|_{\mathrm{L}} \cdot d_{n-1}(\boldsymbol{y}_1, \boldsymbol{y}_2).\end{aligned}$$

□

Our next ingredient is a classical inequality of Hoeffding.

Lemma 3 (Hoeffding's basic inequality)**.** *Let $(K, \mathcal{A}, \mu)$ be a probability space, let $a < b$, and let $f : K \longrightarrow [a, b]$ be measurable. Then*

$$\int \mathrm{e}^f \,\mathrm{d}\mu \le \exp\Big(\int f \,\mathrm{d}\mu + (b-a)^2/8\Big). \tag{6}$$

Proof (of slightly weaker result). By subtracting $\int f \,\mathrm{d}\mu$, we may assume that this average is zero. Let $c := b - a$.

The proof of (6) is elementary, but requires some delicate calculus. Here we give the simpler proof of the slightly weaker inequality

$$\int \mathrm{e}^f \,\mathrm{d}\mu \le \mathrm{e}^{c^2/2}$$

(that is, the factor of $1/8$ is replaced by $1/2$). For this, it suffices to make the weaker assumptions that $\int f \,\mathrm{d}\mu = 0$ and

$$-c \le f \le c.$$

Given these, for any $x \in K$ we may write $f(x)$ as a weighted average of the values $\pm c$: say

$$f(x) = t(x) \cdot c + (1 - t(x)) \cdot (-c), \quad \text{where } t(x) = \frac{1}{2}(f(x) + 1).$$

Since $\exp$ is convex, it follows that

$$\mathrm{e}^{f(x)} = \mathrm{e}^{t(x)c+(1-t(x))(-c)} \leq t(x)\mathrm{e}^{c} + (1-t(x))\mathrm{e}^{-c},$$

and hence

$$\int \mathrm{e}^{f} \leq \mathrm{e}^{c} \int t + \mathrm{e}^{-c} \int (1-t) = \frac{1}{2}(\mathrm{e}^{c} + \mathrm{e}^{-c}) \leq \mathrm{e}^{c^2/2}. \tag{7}$$

The last inequality here follows by a term-by-term comparison of Taylor series. □

Cheat: We have proved only a weaker bound than the statement of Lemma 3, but I will use the full version of that lemma in the sequel. The only difference this makes is slightly improved constants.

Corollary 4 (Decay in MGF controlled by Lipschitz constant)**.** *We have*

$$\int \mathrm{e}^{f} \, \mathrm{d}\mu^{\times n} \leq \mathrm{e}^{\|f\|_{\mathrm{L}}^2/8n^2} \int \mathrm{e}^{P_n f} \, \mathrm{d}\mu^{\times(n-1)}$$

Proof. Observe that

$$\int \mathrm{e}^{f} \, \mathrm{d}\mu^{\times n} = \int_{K^{n-1}} \Big(\int_{K} \mathrm{e}^{f(\boldsymbol{y},x) - P_n f(\boldsymbol{y})} \, \mu(\mathrm{d}x) \Big) \, \mathrm{e}^{P_n f(\boldsymbol{y})} \, \mu^{\times(n-1)}(\mathrm{d}\boldsymbol{y}).$$

Consider the inner integral on the right. For each fixed $\boldsymbol{y}$, the function

$$f(\boldsymbol{y}, \, \cdot \,) - P_n f(\boldsymbol{y})$$

has integral zero with respect to μ (by the definition of P_n) and takes values in an interval of length at most $\|f\|_{\mathrm{L}}/n$ (because we are allowing only one coordinate to vary). Therefore, by Lemma (6), that inner integral is at most $\mathrm{e}^{\|f\|_{\mathrm{L}}^2/8n^2}$. Now substitute this bound into the right-hand side above. □

Proof of Theorem 1. When $n = 1$ this is a special case of Lemma 3. So now suppose that $n \geq 2$ and that (4) is already known for functions on K^{n-1}. In light of Lemma 2, we may apply this inductive hypothesis to the function $P_n f$ to obtain

$$\begin{aligned} \int \mathrm{e}^{P_n f} \, \mathrm{d}\mu^{\times(n-1)} &\leq \exp\Big(\int P_n f \, \mathrm{d}\mu^{\times(n-1)} + \frac{\|P_n f\|_{\mathrm{L}}^2}{8(n-1)} \Big) \\ &\leq \exp\Big(\langle f \rangle + \frac{n-1}{n} \frac{\|f\|_{\mathrm{L}}^2}{8n} \Big). \end{aligned}$$

Now Corollary 4 turns this into

$$\int \mathrm{e}^{f} \, \mathrm{d}\mu^{\times n} \leq \exp\Big(\langle f \rangle + \frac{1}{n} \frac{\|f\|_{\mathrm{L}}^2}{8n} + \frac{n-1}{n} \frac{\|f\|_{\mathrm{L}}^2}{8n} \Big) = \mathrm{e}^{\langle f \rangle + \|f\|_{\mathrm{L}}^2/8n}.$$

This is (4), which we have seen implies Theorem 1. □

1.3 Generalization: the Azuma–Hoeffding theorem

Actually, the argument above applies in an even more general setting with almost no change, and it is often presented that way. Recall that we can identify $P_n f$ with $E_{\mu^{\times n}}(f \mid \mathcal{F}_{n-1})$, where $\mathcal{F}_i$ is the sigma-subalgebra of sets in K^n that depend only on the first i coordinates. It turns out that we can forget the product structure of K^n, and assume only a filtration $(\mathcal{F}_i)_{i=1}^n$ and a certain bound on how the conditional expectations $E_{\mu^{\times n}}(f \mid F_i)$ changes as i increases. The more general theorem which results is one of the oldest and most popular methods for proving concentration. It is easiest to state in more probabilistic language.

Theorem 5. *Let $L > 0$ be a constant, and consider a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a filtration*

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \leq \mathcal{F}_1 \leq \mathcal{F}_2 \leq \cdots \leq \mathcal{F}_n = \mathcal{F}.$$

Suppose the RV X has the following property: for each $i = 1, 2, \ldots, n$, there are $\mathcal{F}_{i-1}$-measurable RVs Y_i and Z_i such that

$$Y_i \leq \mathbf{E}[X \mid \mathcal{F}_i] \leq Z_i \quad \textit{and} \quad Z_i \leq Y_i + L/n \quad \textit{a.s.}. \tag{8}$$

Then

$$\mathbf{P}\big(X - \mathbf{E}[X] > \varepsilon\big) \leq \mathrm{e}^{-2\varepsilon^2 n/L^2}.$$

This is essentially a version of the Azuma–Hoeffing theorem. That theorem is often stated with the slightly different assumption that

$$\|\mathbf{E}[X \mid F_i] - \mathbf{E}[X \mid \mathcal{F}_{i-1}]\|_\infty \leq L/n.$$

In this case one can apply Theorem 5 with

$$Y_i := \mathbf{E}[X_i \mid \mathcal{F}_{i-1}] - L, \quad Z_i := \mathbf{E}[X \mid \mathcal{F}_{i-1}] + L,$$

and the resulting probability bound is $\mathrm{e}^{-2\varepsilon^2 n/(2L)^2} = \mathrm{e}^{-\varepsilon^2 n/2L^2}$.

Thus, to prove concentration for a RV, one can look for a filtration with respect to which X satisfies (8). Intuitively, this condition (8) asserts that, conditionally on $\mathcal{F}_{i-1}$, the next conditional expectation $\mathbf{E}[X \mid \mathcal{F}_i]$ has essential range of length at most L/n: the auxiliary RVs Y_i and Z_i give an $\mathcal{F}_{i-1}$-measurable choice of an interval of that length which contains $\mathbf{E}[X \mid \mathcal{F}_i]$.

The conditional expectations $\mathbf{E}[X \mid \mathcal{F}_i]$ form a martingale, so this method for proving concentration is often called the **method of bounded marginale differences**. To derive Theorem 1 from Theorem 5, we take $(\Omega, \mathcal{F}, \mathbf{P})$ to be $(K^n, \mathcal{A}^{\otimes n}, \mu^{\times n})$,

and let $\mathcal{F}_i$ be the σ-subalgebra of $\mathcal{A}^{\otimes n}$ generated by the first i coordinates in K^n. Then the Lipschitz condition on f translates into (8). In other applications of Theorem 5 the choice of filtration can be more subtle. These applications include many in the setting of random optimization problems: see, for instance, [McD98, GS01].

Here is another valuable example. The symmetric group $\mathrm{Sym}(n)$ has a natural normalized Hamming metric of its own:

$$d_{\mathrm{Sym}(n)}(\sigma, \tau) := \frac{|\{i = 1, 2, \ldots, n : \ \sigma(i) \neq \tau(i)|}{n}.$$

It also has a natural filtration $(\mathcal{F}_i)_{i=1}^n$, where $\mathcal{F}_i$ is the sigma-algebra of sets that depend on only the restriction $\sigma|_{\{1,2,\ldots,i\}}$. This metric and filtration interact similarly to the case of product spaces: an L-Lipschitz function on $\mathrm{Sym}(n)$ satisfies (8) except that the constant is doubled to $2L$ (exercise!). Therefore Theorem 5 has the following corollary.

Theorem 6. *If $f : \mathrm{Sym}(n) \longrightarrow \mathbb{R}$ is L-Lipschitz for $d_{\mathrm{Sym}(n)}$, and μ is the uniform distribution on $\mathrm{Sym}(n)$, then*

$$\mu\Big\{ f \geq \int f \, \mathrm{d}\mu + \varepsilon \Big\} \leq \mathrm{e}^{-\varepsilon^2 n / 2L^2}.$$

□

1.4 Isoperimetry in product spaces

A second simple application of Theorem 1 connects to another valuable point of view on measure concentration.

Let A be a finite alphabet. Let us apply Theorem 1 to 1-Lipschitz functions on A^n of the following kind. Consider a nonempty subset $U \subseteq A^n$, and its associated distance function:

$$f(\boldsymbol{x}) := \min\{d_n(\boldsymbol{x}, \boldsymbol{y}) : \ \boldsymbol{y} \in U\}.$$

The function f vanishes on U, and it is 1-Lipschitz as an easy consequence of the triangle inequality.

Therefore, letting μ be the uniform distribution on A^n, Theorem 1 gives

$$\mu\{f \geq c + \varepsilon\} \leq \mathrm{e}^{-2\varepsilon^2 n} \quad \text{and} \quad \mu\{f \leq c - \varepsilon\} \leq \mathrm{e}^{-2\varepsilon^2 n}, \tag{9}$$

where $c = \int f \, \mathrm{d}\mu$.

Corollary 7. *If $\mu(U) > \mathrm{e}^{-2\varepsilon^2 n}$, then*

$$\mu\{\boldsymbol{x} : \ f(\boldsymbol{x}) < 2\varepsilon\} = \mu\big(B_{2\varepsilon}(U)\big) \geq 1 - \mathrm{e}^{-2\varepsilon^2 n}.$$

Proof. By the second part of (9), we have

$$\mu\{f \leq c - \varepsilon\} \leq \mathrm{e}^{-2\varepsilon^2 n}.$$

In view of our assumption on U, this means that

$$\exists \boldsymbol{x} \in U \setminus \{f \leq c - \varepsilon\}.$$

But $U = \{f = 0\}$, so this tells us that $c < \varepsilon$. Now the first part of (9) gives

$$\mu\{f < 2\varepsilon\} \geq \mu\{f \leq c + \varepsilon\} \geq 1 - \mathrm{e}^{-\varepsilon^2 n/2}.$$

□

Thus, if U is any subset of A^n which is bigger than $\mathrm{e}^{-2\varepsilon^2 n}$ (which is very small), and we expand U by the (small) distance 2ε, then it fills up all but an exponentially small portion of the whole space A^n! This curious phenomenon is a special feature of various 'high-dimensional' settings for probability theory. It turns out that, up to the values of various constants, it is equivalent to concentration of Lipschitz functions as expressed in Theorem 1. (Exercise: prove this by applying this expansion phenomenon of sets to level-sets of 1-Lipschitz functions).

This expansion phenomenon of subsets is another natural way to introduce the idea of measure concentration. It is an even more classical result in the case of high-dimensional unit spheres. For those it follows from a famous theorem of Lévy showing that the isoperimetrc problem on spheres is solved by intersections with half-spaces. In the next lecture we turn to concentration of Lipschitz functions on spheres, and give an approach to the basic result there which is analogous to our proof of Theorem 1 above.

2 Concentration and positive curvature

2.1 High-dimensional spheres

Let S^n be the unit sphere in $\mathbb{R}^{n+1}$, and let μ be its surface-area measure, normalized to have total mass 1.

Theorem 8 (Concentration on spheres)**.** *Any* $f \in \mathrm{Lip}(\mathrm{S}^n)$ *satisfies*

$$\int \mathrm{e}^f \, \mathrm{d}\mu \leq \mathrm{e}^{\langle f \rangle + \|f\|_{\mathrm{L}}^2/2(n-1)}.$$

This theorem can be proved using a Markov process on the sphere, following analogous steps to the proof of Theorem 1. This is explained in [Led92], and we give an intuitive sketch of that proof here. The main complication is that now one must use more advanced probability theory to set up a suitable Markov process in the first place. The right process to use is Brownian motion on the sphere. We do not introduce this rigorously here: see, for instance, [RW00, Chapter V]. We shall simply cite a few basic facts.

For each $x \in \mathrm{S}^n$, one can construct a **Brownian motion on the sphere started at** x. This is a continuous-time Markov process $(X_t)_{t \geq 0}$ on S^n which has continuous sample-paths, satisfies $X_0 = x$, and is well-approximated over short timescales by Brownian motions on tangent spaces. Using this process, we can define the associated expectation operators on $C(\mathrm{S}^n)$ by

$$P_t f(x) := \mathbf{E}^x f(X_t) \quad \text{for } t \geq 0,$$

where $\mathbf{E}^x$ denotes expectation with respect to the Brownian motion started at x. These operators form a semigroup:

$$P_{t+s} = P_t \circ P_s \quad \text{whenever } t, s \geq 0.$$

This follows from the Markov property of the process. This semigroup of operators satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t} P_t f(x) = \frac{1}{2} \Delta P_t f(x) \quad \text{for } f \in C^2(\mathrm{S}^n), \tag{10}$$

where Δ is the Laplace–Beltrami operator of the sphere. That operator may be written

$$\Delta f(x) = \sum_{i=1}^{n} \partial_{e_i} \partial_{e_i} f,$$

where $e_1, \ldots, e_n$ are any orthonormal basis for the tangent space to S^n at x, and ∂_{e_i} denotes differentiation in direction e_i. The relation (10) is described by saying that Δ is the **generator** of the semigroup $(P_t)_{t \geq 0}$ (where I'm suppressing some fiddly issues about the correct choice of domain for Δ).

Two basic facts that I will assume without proof:

Lemma 9 (Integration by parts)**.** *For any $f, h \in C^2(\mathrm{S}^n)$, we have*

$$\int \Delta f \cdot h \, \mathrm{d}\mu = -\int \langle \operatorname{grad} f, \operatorname{grad} h \rangle \, \mathrm{d}\mu.$$

□

With these preparations in hand, the proof of Theorem 8 has two basic steps. They mirror Lemma 2 and Corollary 4 in the proof of Theorem 1.

Lemma 10 (Decay in Lipschitz constant)**.** *If f is Lipschitz then*

$$\|P_t f\|_{\mathrm{L}} \leq \mathrm{e}^{-(n-1)t/2} \|f\|_{\mathrm{L}} \quad \forall t \geq 0.$$

Remark. In the original presentation of Theorem 8 in [Led92], Ledoux uses the semigroup $Q_t := P_{2t}$: that is, he lets Brownian motion run at twice its usual speed. This removes some tedious factors of $1/2$ during the course of the proof: for instance, the above lemma becomes

$$\|Q_t f\|_{\mathrm{L}} \leq \mathrm{e}^{-(n-1)t} \|f\|_{\mathrm{L}}.$$

I have stuck to the usual definition of Brownian motion and its semigroup for consistency with the bulk of the probability theory literature.

Sketch proof. One can prove this quite quickly via a differential inequality for the function $t \mapsto \|P_t f\|_{\mathrm{L}}$. That, in turn, can be obtained using the calculus of differential operators on the sphere: in particular, a standard identity called *Bochner's formula*. However, let us sketch a more 'visual' proof.

Consider two non-antipodal points on the sphere, say x and y. Let the Riemannian distance between them be ℓ, and let $\gamma : [0, \ell] \longrightarrow \mathrm{S}^n$ be the speed-one geodesic from x to y.

There is a unique rotation T of $\mathbb{R}^{n+1}$ such that (i) $Tx = y$ and (ii) the subspace orthogonal to x and y is pointwise fixed by T. One can think of T as an isometry of S^n which moves x along the geodesic γ to y, and fixes the rest of the sphere as far as possible.

Now let x' be another point very close to x, say with $d(x, x') = \delta$, and let $y' := Tx'$. We can now state the key geometric feature of the sphere which underlies Theorem 8: after constructing y' as above, we have

$$d(x', y') = \cos\theta \cdot d(x, y) = (1 - \theta^2/2 + o(\theta^2)) \cdot d(x, y), \tag{11}$$

where θ is the orthogonal distance from x to the geodesic γ. The negative sign in front of $\theta^2/2$ here is a basic manifestation of the positive curvature of the sphere.

We apply (11) together with the following fact: if $(X_t^x)_t$ is a Brownian motion on S^n started at x, then $(T(X_t^x))_t$ is a Brownian motion on S^n started at y. This is because the law of Brownian motion on S^n is respected by the isometries of the sphere. This defines a canonical coupling of those two Brownian motions. Using that coupling, we obtain

$$\begin{aligned}|P_t f(x) - P_t f(y)| = \big|\mathbf{E}^x[f(X_t)] - \mathbf{E}^y[f(X_t)]\big| &\leq \mathbf{E}^x\big[|f(X_t) - f(T(X_t))|\big] \\ &\leq \|f\|_{\mathrm{L}} \cdot \mathbf{E}^x[d(X_t, T(X_t))].\end{aligned}$$

By (11), this upper bound equals

$$\|f\|_{\mathrm{L}} \cdot d(x,y) \cdot \Big(1 - \mathbf{E}^x\big[d(\gamma, X_t)^2/2 + o(d(\gamma, X_t)^2)\big]\Big) \quad \text{as } t \downarrow 0.$$

In this expression, $d(\gamma, X_t)$ is the distance from X_t to γ. The leading-order behaviour of this quantity as $t \downarrow 0$ is the same as the distance between a Brownian motion in $\mathbb{R}^n$ and a fixed line through the origin: in particular, its variance behaves the same as a sum of $n-1$ Gaussians of variance t:

$$\mathbf{E}^x[d(\gamma, X_t)^2/2] = (n-1)t/2 + o(t). \tag{12}$$

Inserting this into the upper bound above, we obtain

$$\|f\|_{\mathrm{L}} \cdot d(x,y) \cdot \big(1 - (n-1)t/2 + o(t)\big) \quad \text{as } t \downarrow 0.$$

Taking the supremum over x and y, we have shown that

$$\|P_t f\|_{\mathrm{L}} \leq (1 - (n-1)t/2 + o(t))\|f\|_{\mathrm{L}} \quad \text{as } t \downarrow 0.$$

Finally, for any fixed t, the semigroup property of $(P_s)_{s \geq 0}$ lets us write

$$P_t f = (P_{t/m})^m f$$

for any $m \in \mathbb{N}$. Applying the above inequality m times, we obtain

$$\begin{aligned}\|P_t f\|_{\mathrm{L}} &\leq \Big(1 - (n-1)\frac{t}{2m} + o\Big(\frac{t}{m}\Big)\Big)\|(P_{t/m})^{m-1} f\|_{\mathrm{L}} \\ &\leq \cdots \leq \Big(1 - (n-1)\frac{t}{2m} + o\Big(\frac{t}{m}\Big)\Big)^m \|f\|_{\mathrm{L}}.\end{aligned}$$

Sending $m \longrightarrow \infty$, this upper bound converges to $\mathrm{e}^{-(n-1)t/2}\|f\|_{\mathrm{L}}$. □

We also need the following corollary of Lemma 10.

Corollary 11 (Ergodicity of Brownian motion)**.** *For any $f \in C(\mathrm{S}^n)$ we have*

$$P_t f(x) \longrightarrow \int f \,\mathrm{d}\mu \quad \textit{as } t \longrightarrow \infty,$$

uniformly in the choice of start-point x. □

We now turn to the second step in the proof of Theorem 8.

Lemma 12 (Decay in MGF controlled by Lipschitz constant)**.** *If $f \in C^2(\mathrm{S}^n)$, then*

$$\frac{\mathrm{d}}{\mathrm{d}t} \log \int \mathrm{e}^{P_t f} \,\mathrm{d}\mu \geq -\frac{1}{2} \|P_t f\|_{\mathrm{L}}^2 \quad \forall t \geq 0.$$

Proof. Differentiating under the integral sign gives

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \mathrm{e}^{P_t f} \,\mathrm{d}\mu = \int \frac{\mathrm{d}}{\mathrm{d}t} (P_t f) \cdot \mathrm{e}^{P_t f} \,\mathrm{d}\mu = \frac{1}{2} \int \Delta(P_t f) \cdot \mathrm{e}^{P_t f} \,\mathrm{d}\mu.$$

By Lemma 9, this equals

$$-\frac{1}{2} \int \langle \operatorname{grad} P_t f, \operatorname{grad} (\mathrm{e}^{P_t f}) \rangle \,\mathrm{d}\mu = -\frac{1}{2} \int \langle \operatorname{grad} P_t f, \operatorname{grad} P_t f \rangle \mathrm{e}^{P_t f} \,\mathrm{d}\mu,$$

and so it is bounded below by

$$-\frac{1}{2} \|P_t f\|_{\mathrm{L}}^2 \int \mathrm{e}^{P_t f} \,\mathrm{d}\mu.$$

Dividing by $\int \mathrm{e}^{P_t f} \,\mathrm{d}\mu$, this becomes the desired lower bound. □

Proof of Theorem 8. Since $C^2(\mathrm{S}^n)$ is uniformly dense in $\operatorname{Lip}(\mathrm{S}^n)$, it suffices to prove the result when f is twice differentiable.

Consider the integral

$$\int \mathrm{e}^{P_t f} \,\mathrm{d}\mu$$

as a function of $t \in [0, \infty)$. At $t = 0$ it equals $\int \mathrm{e}^f \,\mathrm{d}\mu$, and as $t \longrightarrow \infty$ it tends to $\exp \int f \,\mathrm{d}\mu$, by Lemma 11. It is also easily checked to be differentiable in t, so

$$\log \int \mathrm{e}^f \,\mathrm{d}\mu = \langle f \rangle - \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t} \log \Big(\int \mathrm{e}^{P_t f} \,\mathrm{d}\mu \Big) \,\mathrm{d}t.$$

By Lemma 12, this is at most

$$\langle f \rangle + \frac{1}{2} \int_0^\infty \|P_t f\|_{\mathrm{L}}^2 \, \mathrm{d}t.$$

By Lemma 10, this is at most

$$\langle f \rangle + \frac{1}{2} \int_0^\infty \mathrm{e}^{-(n-1)t} \|f\|_{\mathrm{L}}^2 \, \mathrm{d}t = \langle f \rangle + \frac{1}{2(n-1)} \|f\|_{\mathrm{L}}^2.$$

Exponentiating, we arrive at the desired inequality. □

2.2 Generalization: positively curved manifolds

Now consider a compact Riemannian manifold (M, g). Let μ be its volume measure, normalized to be a probability measure. Let Δ_g be the Laplace–Beltrami operator. The operator $\frac{1}{2}\Delta_g$ generates a Feller semi-group $(P_t)_{t \geq 0}$ acting on $C(M)$, which is associated to Brownian motion on M.

Using these constructs, one can try to generalize the argument about the sphere above. Inevitably, one needs some kind of geometric assumption about the manifold (M, g). The specific need for it appears when we try to generalize Lemma 10.

For a general manifold, one cannot argue about great circles and rigid isometries as we did in the proof of that theorem for sphere. However, substitutes are available. Given the geodesic γ from x to y and a point x' very close to x, there is a unique unit vector $v(0) \in T_x M$ such that x' is reached by following a geodesic from x of length δ and initial direction $v(0)$. Now the general construction of *parallel transport* produces allows us to move the tangent vector $v(0)$ along the geodesic γ, producing a continuous family of unit tangent vectors $v(t) \in T_{\gamma(t)} M$. Parallel transport is the unique way to do this so that these vectors exhibit no 'infinitessimal rotation' as t increases. In particular, we arrive at $v(\ell) \in T_y M$, and then we can produce y' by following a geodesic from y of length δ and initial direction $v(\ell)$.

Having made this construction, the key to completing the proof is understanding the correct substitute for (11). This is provided by a standard formula from differential geometry: the formula for the *second variation of arc length* (see, for instance, [dC76, Section 5-4, Proposition 4] and [dC92, Chapter 9, Proposition 2.8]). It gives

$$d(x', y') = d(x, y) - \int_0^\ell \theta^2 \cdot K(v(t), \gamma'(t)) \, \mathrm{d}t + o(\theta^2),$$

where once again θ is the orthogonal distance from x' to the (extension of the) geodesic γ. The new quantity appearing in this integrand, $K(v(t), \gamma'(t))$, is the *sectional curvature* (see [dC92, Section 4.3]) of the infinitessimal surface in M passing through the point $\gamma(t)$ and defined by the unit tangent vectors $v(t)$ and $\gamma'(t)$. The sectional curvature is a fundamental descriptor of curvature in differential geometry, and is computed from the more informative, rank-four *curvature tensor* (see [dC92, Section 4.2]).

Applying the above calculation with $x' = X_t^x$, similarly to the sphere, one first checks that y' has the same law as X_t^y. On more general manifolds this holds only up to some higher-order error terms as $t \longrightarrow \infty$, but this still suffices. Then one takes expectation to obtain

$$\mathbf{E}[d(X_t^x, X_t^y)] = d(x, y) - \int_0^\ell \mathbf{E}[\theta^2 K(v(t), \gamma'(t))] \, \mathrm{d}t + \mathbf{E}[o(\theta^2)].$$

Here, θ is now the perpendicular distance from X_t^x to the geodesic γ, and $v(t)$ is the unit vector at $T_{\gamma(t)}M$ obtained by parallel transport of the direction in which X_t^x lies from x. Thus, θ and $v(t)$ are both random, depending on X_t^x.

By some slighly more involved calculations along the lines of (12), the expectation involving these quantities has leading-order behaviour equal to

$$(n-1)t \int_{\mathrm{S}(T_{\gamma(t)}M) \cap (\gamma'(t))^\perp} K(v, \gamma'(t)) \, \mathrm{d}v = t \cdot \mathrm{Ric}_{\gamma(t)}(\gamma'(t)),$$

where the integral is over the unit vectors in $T_{\gamma(t)}M$ which are orthogonal to the direction $\gamma'(t)$, and where $\mathrm{Ric}_{\gamma(t)}(\gamma'(t))$ is the *Ricci curvature* of M at $\gamma(t)$ in the direction $\gamma'(t)$ (see [dC92, Section 4.4])[2].

Filling in the details of this argument, one can generalize Lemma 10 and hence Theorem 8 assuming a lower bound on the Ricci curvature.

Theorem 13 (Concentration on positively curved manifolds)**.** *Let (M, g) be a Riemannian manifold with normalized volume measure μ, and assume that the Ricci curvature is at least $c > 0$ in all unit tangent directions on M. Then any $f \in \mathrm{Lip}(M)$ satisfies*

$$\int \mathrm{e}^f \, \mathrm{d}\mu \leq \mathrm{e}^{\langle f \rangle + \|f\|_{\mathrm{L}}^2 / 2c}.$$

[2]Beware that my conventions differ from [dC92]: his Ricci curvature is our Ric divided by $n-1$. Our convention matches most of the literature on concentration.

This theorem is proved in [Led92] following the lines used above for the case of the sphere. A previous proof used more delicate results on isoperimetry and comparison results between positively-curved manifolds and corresponding spheres; see [GM83].

Much as we saw how the Azuma–Hoeffding theorem allows us to extend McDiarmid's inequality to many other discrete metric probability spaces, Theorem 13 extends Theorem 8 to many other natural manifolds of interest. Of particular interest in random matrix theory are the unitary and orthogonal groups in high dimensions and various related subugroups. For instance, a standard calculation shows that if we give the special unitary group $\mathrm{SU}(n)$ the Riemannian structure defined by the normalized Frobenius inner product

$$\langle A, B \rangle := \frac{1}{n} \mathrm{Tr}(AB^*),$$

then its dimension as a real maniforld is $n^2 - 1$ and its Ricci curvature is $n^2/2$ in all unit tangent directions. (See, for instance, [AGZ10, Appendix F].) Thus:

Theorem 14 (Concentration on unitary groups)**.** *If $f \in \mathrm{Lip}(\mathrm{SU}(n))$ for the normalized Frobenius inner product, then*

$$\int \mathrm{e}^f \, \mathrm{d}\mu \leq \mathrm{e}^{\langle f \rangle + \|f\|_{\mathrm{L}}^2 / n^2}.$$

□

This applies, for instance, to any function of unitary matrices of the form

$$U \mapsto \frac{1}{n} \mathrm{Tr}(F(U)),$$

where F is a Lipschitz function from the unit circle to $\mathbb{R}$ and $F(U)$ is defined via the spectral theorem. Concentration for this kind of functional of matrices is the basis of many proofs that large random matrices have approximately a fixed spectral distribution with high probability, not just one average.

A further generalization of Theorem 13 can be obtained for a re-weighted measure on M of the form

$$\frac{\mathrm{e}^{-V(x)} \mu(\mathrm{d}x)}{\text{normalizing constant}}.$$

Now one must replace the lower bound on Ricci curvature by a lower bound on

$$\text{Ricci curvature} + \text{Hessian of } V,$$

and then use a more general construction of diffusion processes on manifolds to complete the proof. The resulting theorem also encompasses another important classical example of measure concentration: Gaussian distributions in high-dimensional Euclidean spaces. In random matrix theory, this is the source of many concentration results for the GUE and GOE.

Many good sources explore applications of concentration inequalities such as Theorem 1 and Theorem 13 to random matrices. Often more specialized concentration inequalities than those can be used to obtain sharper results. See, for instance, [GZ00] and [Led07, Section 3].

Notes and remarks

Lemma 3 is due to Hoeffding [Hoe63]. Its use to prove Theorem 1 is due to McDiarmid: see [McD98, Lemma 2.6] and the arguments that follow it. The method of bounded martingale differences has been studied in depth by McDiarmid, and his surveys [McD89, McD98] are a good place to learn about a wide range of refinements and applications.

A standard treatment of the Azuma–Hoeffding theorem, including the application to bin-packing and several others, can be found in [GS01, Section 12.2]

An insightful geometric point of view on measure concentration is offered in [Gro01, Chapter $3\frac{1}{2}$], which considers a wide variety of examples. Other good introductory references include [MS86, Led01, Ver].

References

[AGZ10] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2010.

[dC76] Manfredo P. do Carmo. *Differential geometry of curves and surfaces*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1976. Translated from the Portuguese.

[dC92] Manfredo Perdigão do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser Boston, Inc., Boston, MA, 1992. Translated from the second Portuguese edition by Francis Flaherty.

[GM83] M. Gromov and V. D. and Milman. A topological application of the isoperimetric inequality. *Amer. J. Math.*, 105(4):843–854, 1983.

[Gro01] Mikhail Gromov. *Metric Structures for Riemannian and Non-Riemannian Spaces*. Birkhäuser, Boston, 2nd edition, 2001.

[GS01] Geoffrey R. Grimmett and David R. Stirzaker. *Probability and random processes*. Oxford University Press, New York, third edition, 2001.

[GZ00] A. Guionnet and O. Zeitouni. Concentration of the spectral measure for large matrices. *Electron. Comm. Probab.*, 5:119–136, 2000.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[Led92] M. Ledoux. A heat semigroup approach to concentration on the sphere and on a compact Riemannian manifold. *Geom. Funct. Anal.*, 2(2):221–224, 1992.

[Led01] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[Led07] M. Ledoux. Deviation inequalities on largest eigenvalues. In *Geometric aspects of functional analysis*, volume 1910 of *Lecture Notes in Math.*, pages 167–219. Springer, Berlin, 2007.

[McD89] Colin McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Math. Soc. Lecture Note Ser.*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.

[McD98] Colin McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, volume 16 of *Algorithms Combin.*, pages 195–248. Springer, Berlin, 1998. Free preprint available online at `http://cgm.cs.mcgill.ca/~breed/conc/colin.pdf`.

[MS86] Vitali D. Milman and Gideon Schechtman. *Asymptotic theory of finite-dimensional normed spaces*, volume 1200 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1986. With an appendix by M. Gromov.

[RW00] L. C. G. Rogers and David Williams. *Diffusions, Markov processes, and martingales. Vol. 2*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000. Itô calculus, Reprint of the second (1994) edition.

[Shi96] Paul Shields. *The Ergodic Theory of Discrete Sample Paths*, volume 13 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, 1996.

[Ver] Roman Vershynin. *High-dimensional probability: An Introduction with Applications in Data Science*. Book to appear from Cambridge University Press; draft available online at `https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.html`.

TIM AUSTIN

Email: `tim@math.ucla.edu`
URL: `math.ucla.edu/~tim`