NOTES ON TALAGRAND'S ISOPERIMETRIC INEQUALITY

NICK COOK

Abstract. These are notes on a well-known concentration of measure inequality of Talagrand, prepared for the participating analysis seminar at UCLA in the Fall quarter of 2011, and slightly revised for the concentration of measure reading group (organized by Marek Biskup and Oren Louidor) in Spring 2012. The material here is based largely on the excellent treatments in the books [Led01], [AS08], [Ste97] and [Tao12], as well as the papers [Tal96], [KV02].

Contents

1. Two of Talagrand's Inequalities

The aim of these notes is to prove the following theorem and corollary, which first appeared in [Tal95] (see also [Tal96]), and to present a selection of applications. The bound (2) can be regarded as a sort of isoperimetric inequality, while (3) has the form of what is commonly referred to in probability theory as a "concentration inequality". All three of (1), (2) and (3) are commonly referred to as "Talagrand's inequality" in the literature; somewhat confusingly, Talagrand has a few other well-known inequalities, unrelated to these, that are also commonly called "Talagrand's inequality"!

Date: May 7, 2012, minor corrections June 22, 2017.

Theorem 1. *Let $(\Omega_1, \mathcal{F}_1, \mu_1), \dots, (\Omega_n, \mathcal{F}_n, \mu_n)$ be probability spaces and consider the probability space $\Omega = \Omega_1 \times \cdots \times \Omega_n$ equipped with the product sigma algebra and probability measure $P = \mu_1 \otimes \cdots \otimes \mu_n$. For all nonempty measurable subsets $A \subset \Omega$,*

$$\int_\Omega e^{d_C(x,A)^2/4} dP(x) \leq 1/P(A) \tag{1}$$

where $d_C(x, A)$ is the **convex distance** *(defined in Section 2.1) from x to A. As a consequence, by Chebyshev's inequality we have*

$$P(A_t^c) \leq \frac{1}{P(A)} e^{-t^2/4} \tag{2}$$

where $A_t = \{x \in \Omega : d_C(x, A) \leq t\}$.

Corollary 2. *Let $X = (X_1, \dots, X_n)$ be a random variable with independent components taking values in $[0, 1]$. Let $F : \mathbf{R}^n \to \mathbf{R}$ be a convex 1-Lipschitz function. Let $MF(X)$ be a median for $F(X)$. Then for all $t \geq 0$*

$$P(|F(X) - MF(X)| \geq t) \leq 4e^{-t^2/4} \tag{3}$$

We first present a nice immediate application of the corollary.

Let M be an $n \times n$ Hermitian matrix. Then the largest eigenvalue $\lambda_1(M) = \|M\|_{op}$. Considering the operator norm of M as a function of the n^2 components of the entries (the n real diagonal entries and the $n(n-1)$ real and imaginary parts of the entries above the diagonal), we see that it is a convex and 1-Lipschitz function from $\mathbf{R}^{n^2}$ with Euclidean distance to $\mathbf{R}_+$. Indeed, convexity follows from the triangle inequality for the operator norm, and 1-Lipschitz follows from $\|M\|_{op} \leq \|M\|_2$ (where the Frobenius norm on $n \times n$ matrices is the Euclidean norm on $\mathbf{R}^{n^2}$).

Hence, if X is a random Hermitian matrix, where the diagonal entries and the real and imaginary parts of the strict upper-triangle entries are independent bounded scalar random variables, and we identify the space of Hermitian matrices with $\mathbf{R}^{n^2}$, then by Talagrand's inequality we have that the random variable $\lambda_1(X)$ is concentrated around its mean with sub-Gaussian tails independent of n. This is especially interesting as it turns out that $\lambda_1(X)$ is of order $\sqrt{n}$.

2. Talagrand's proof

2.1. **The convex distance.** First we must define the convex distance, which is a primitive notion of distance between a point and a set that is somewhat opaque at first. It does not, in fact, come from a metric in the usual way.

Let Ω as in the theorem, and let $A \subset \Omega$, $x \in \Omega$. We define $U_A(x) \subset \{0, 1\}^n$

$$U_A(x) = \{s \in \{0, 1\}^n : \exists y \in A \text{ with } y_i = x_i \text{ whenever } s_i = 0\}. \tag{4}$$

If Ω is such that we are allowed to subtract elements, we can slightly more intuitively say that a vector $s = (s_1, \dots, s_n)$ in the binary cube *supports* a vector $z \in \Omega$ if $z_i \neq 0$ only when $s_i = 1$, and define $U_A(x)$ to be the set of vectors in the binary cube that support some element of $A - x$. (Indeed when we prove the corollary we will have $\Omega = [0,1]^n$, so we will have vector space structure.) One can internalize the elements $s \in U_A(x)$ as (rather coarse) travel plans: if a $0-1$ vector s is in $U_A(x)$, it means that starting at x, to get to A it is sufficient to vary only the coordinates i for which s_i is 1.

Now let $V_A(x) \subset \mathbf{R^n}$ be the convex hull in of $U_A(x)$. We define $d_C(x,A) = d_E(0, V_A(x))$.

2.2. **Proof of Corollary 2.**

The corollary uses the notation of probability, while the theorem uses the integral symbol. We'll use the former here.

We first recall that a median MX for a random variable X is a real number satisfying the inequalities $P(X \geq MX) \geq 1/2$ and $P(X \leq MX) \geq 1/2$.

Key in the passage from Theorem 1 to Corollary 2 is the observation that for the special case of A convex in $[0,1]^n$, the convex distance controls the Euclidean distance.

Lemma 3. *Let A convex in $[0,1]^n$ and $x \in [0,1]^n$. Then $d_E(x,A) \leq d_C(x,A)$.*

We postpone the proof of this lemma, and note that the lemma and the theorem imply

$$Ee^{d_E(X,A)^2/4} \leq 1/P(X \in A)$$

for any convex subset A of $[0,1]^n$. From here it is a short walk to Corollary 2. Indeed, let $a \geq 0$ and take $A = \{F \leq a\}$. Observe that by the Lipschitz property, if $X \in \{F \geq a+t\}$ for some $t \geq 0$, then $d_E(X,A) \geq t$. Then by applying Chebyshev's inequality to the LHS of (7), we have

$$P(F(X) \geq a+t)e^{t^2/4} \leq 1/P(F(X) \leq a).$$

Now taking $a = MF(X)$ we get the upper tail estimate

$$P(F(X) - MF(X) \geq t) \leq 2e^{-t^2/4}$$

and taking $a = MF(X) - t$ we get the lower tail estimate

$$P(F(X) - MF(X) \leq -t) \leq 2e^{-t^2/4}$$

(where the definition of median has given us the prefactors 2). The desired result follows from union bound. $\square$

Proof of Lemma 3. Suppose $d_C(x,A) \leq t$. Then by definition of convex distance, there exists a convex combination $w = \sum_{i=1}^m \lambda_i \vec{s}_i$ of vectors $\vec{s}_i \in U_A(x)$ $1 \leq i \leq m$ such that $\|w\|_E \leq t$. Now for each i, $\vec{s}_i \in U_A(x)$ means there exists $\vec{z}_i \in A - x$ supported by $\vec{s}_i$. Let $z = \sum_{i=1}^m \lambda_i \vec{z}_i$. Then $z \in A - x$ by convexity. $z_i \in [0,1]^n$ implies that each component of $\vec{z}_i$ is bounded by 1, and since each component is only nonzero when the corresponding component of $\vec{s}_i$ nonzero, we have that each component of $\vec{z}_i$ is bounded by the corresponding component of $\vec{s}_i$. Thus, $d_E(x,A) \leq \|z\|_E \leq \|w\|_E \leq t$, and the claim follows. $\square$

2.3. **Proof of the isoperimetric inequality.** We now present Talagrand's proof of Theorem 1 by induction on n. For the case $n = 1$, note that $d_C(x, A) = 0$ if $x \in A$ and $d_C(x, A) = 1$ if x is not in A. Now we must show

$$e^{1/4}(1 - P(A)) + P(A) \leq 1/P(A)$$

which follows from $e^{1/4}(1 - u) + u \leq 1/u$ for all $u \in [0, 1]$. The case $n = 1$ is established.

For the inductive step we need a Lemma:

Lemma 4. *a) For all u in $(0, 1]$ we have*

$$\inf_{\lambda \in [0,1]} e^{(1-\lambda)^2/4} u^{-\lambda} \leq 2 - u. \tag{5}$$

b) $e^{1/4} \leq 2$ (Pf: $e \leq 2^4$)

Proof. If $u \geq e^{-1/2}$ take $\theta = 1 + 2 \log u$. Otherwise take $\theta = 0$. Then do calculus. □

Assume the result holds for n. Let $\Omega' = \Omega_1 \times \cdots \times \Omega_n$ a product probability space with product measure P, and let Ω_{n+1} be another probability space with measure μ_{n+1}. Let $\Omega = \Omega' \times \Omega_{n+1}$. Let $A \subset \Omega$ and $x \in \Omega$. The proof of the result for $n + 1$ follows these steps:

1. Obtain an inequality for $d_C(x, A)$ from consideration of "slices" and the "projection" of A in Ω', and convexity.
2. Apply Hölder's inequality and the induction hypothesis.
3. Optimize using the Lemma.
4. Fubini.

For a point $z \in \Omega$ we write $z = (z', \omega)$, $z' \in \Omega', \omega \in \Omega_{n+1}$. Let $A(\omega) = \{z' \in \Omega^n : (z', \omega) \in A\}$ be the ω-slice of A, and $B = \bigcup_{\omega \in \Omega} A(\omega)$ be the projection of A to Ω^n. Let $x = (x', \omega) \in \Omega^{n+1}$. The key observation that gets the proof rolling is that we can bound the convex distance $d_C(x, A)$ in terms of the distances to the sections A_ω and the projection B.

We again think of the $s \in U_A(x)$ as listing the coordinates it is sufficient to vary along in order to get from x to A. We see two ways to get to A from $x = (x', \omega)$: if $A(\omega)$ is nonempty then we can get there by leaving the $n+1$ coordinate ω fixed and varying the other coordinates (for which we would consult the travel plans in $U_{A(\omega)}(x')$); we can also leave ω fixed and walk into B, which is the shadow of A in Ω', and then move up or down to get into A.

Phrasing these two travel plans in terms of $U_A(x), U_{A(\omega}(x')$, and $U_B(x')$, we have that if $s \in U_{A(\omega)}(x')$ then $(s, 0) \in U_A(x)$. If $t \in U_B(x')$ then $(t, 1) \in U_A(x)$.

Taking convex hulls, if $\xi \in V_{A(\omega)}(x')$ and $\zeta \in V_B(x')$, then $(\xi, 0)$ and $(\zeta, 1)$ are in $V_A(x)$, and by convexity, $\lambda(\xi, 0) + (1 - \lambda)(\zeta, 1) \in V_A(x)$ for all $\lambda \in [0, 1]$. Hence we have for all $\lambda \in [0,1]$ and conditional on $A(\omega)$ nonempty

$$\begin{aligned} d_C(x,A) &\le |\lambda(\xi,0) + (1-\lambda)(\zeta,1)|_E \\ &= |(\lambda\xi + (1-\lambda)\zeta, 1-\lambda)|_E \end{aligned}$$

and by convexity of $u \mapsto |u|_E^2$ on $\mathbf{R}$,

$$d_C(x,A)^2 \le |\lambda\xi + (1-\lambda)\zeta|_E^2 + (1-\lambda)^2 \le \lambda|\xi|_E^2 + (1-\lambda)|\zeta|_E^2 + (1-\lambda)^2.$$

Since ξ and ζ were arbitrary we have

$$d_C(x,A)^2 \le \lambda d_C(x', A(\omega))^2 + (1-\lambda) d_C(x',B)^2 + (1-\lambda)^2.$$

In the case that $x = (x',\omega)$ is such that $A(\omega)$ is empty, the only plans to get from x to A are of the form $(t,1)$ for some $t \in U_B(x')$, and we get the same estimate as above with $\lambda = 0$. (Note that B nonempty since A is assumed to be nonempty.)

Talagrand notes that the main trick of the proof is to resist the temptation to optimize in λ at this point! Instead we fix ω, exponentiate the inequality and integrate over Ω'. Conditional on $A(\omega)$ nonempty we have

$$\begin{aligned} &\int_{\Omega^n} e^{d_C((x',\omega),A)^2/4} dP(x') \\ &\quad \le e^{(1-\lambda)^2/4} \int_{\Omega^n} e^{\lambda d_C(x',A(\omega))^2/4} e^{(1-\lambda)d_C(x',B)^2/4} dP(x') \\ &\quad \le e^{(1-\lambda)^2/4} \left(\int_{\Omega^n} e^{d_C(x',A(\omega))^2/4} dP(x') \right)^{\lambda} \left(\int_{\Omega^n} e^{d_C(x',B)^2/4} dP(x') \right)^{1-\lambda} \\ &\quad \le e^{(1-\lambda)^2/4} \left(\frac{1}{P(A(\omega))} \right)^{\lambda} \left(\frac{1}{P(B)} \right)^{1-\lambda} \\ &\quad = \frac{1}{P(B)} e^{(1-\lambda)^2/4} \left(\frac{P(A(\omega))}{P(B)} \right)^{-\lambda}, \end{aligned}$$

where we have applied Hölder's inequality and the induction hypothesis. In the case that $A(\omega)$ is empty we have

$$\int_{\Omega'} e^{d_C(x',\omega),A)^2/4} dP(x') \le \frac{1}{P(B)} e^{(1-\lambda)^2/4}.$$

We now optimize λ by applying the lemma. For $A(\omega)$ nonempty we apply part (a) of the lemma with $u = P(A(\omega))/P(B)$ (which is in $[0,1]$ by monotonicity of P) to obtain

$$\int_{\Omega^n} e^{d_C((x',\omega),A)^2/4} dP(x') \le \frac{1}{P(B)} \left(2 - \frac{P(A(\omega))}{P(B)} \right). \tag{6}$$

For $A(\omega)$ empty, applying part (b) of the lemma gives the same estimate. Now we have

$$\begin{aligned}\int_\Omega \int_{\Omega^n} e^{d_C((x',\omega),A)^2/4} dP(x')d\mu(\omega) &\leq \frac{1}{P(B)}\left(2 - \frac{P\otimes\mu(A)}{P(B)}\right) \\ &= \frac{1}{P\otimes\mu(A)}\frac{P\otimes\mu(A)}{P(B)}\left(2 - \frac{P\otimes\mu(A)}{P(B)}\right) \\ &\leq \frac{1}{P\otimes\mu(A)}\end{aligned}$$

where the last inequality follows from $u(2-u) \leq 1$ for all $u \in \mathbf{R}$. □

Remark 5. One might wonder what rationale would lead one to seek and apply a bound like that of Lemma 4. A clear motivation is that to apply Fubini's theorem to integrate out the ω dependence we would like the right hand side of (6) to be a linear function of $P(A(\omega))$.

3. The Herbst argument for the upper tail

We present here a different argument of Ledoux for the upper tail

$$P(F(X) - EF(X) \geq t) \leq Ce^{-ct^2} \tag{7}$$

of Corollary 2 (but with the median replaced by the expectation) based on a log-Sobolev inequality. The presentation here closely follows that of [Tao12]. First by a standard regularization argument we may assume that F is smooth. We first establish the following log-Sobolev inequality:

Lemma 6. *Let $F : \mathbf{R}^n \to \mathbf{R}$ be a smooth convex function. Then*

$$E(F(X)e^{F(X)}) \leq (Ee^{F(X)})(\log Ee^{F(X)}) + CE(e^{F(X)}|\nabla F(X)|^2) \tag{8}$$

for some absolute constant C independent of n.

Proof. We begin with the $n = 1$ case. The trick is to introduce an independent copy Y of the random variable X in order to take advantage of the Lipschitz hypothesis. Indeed, for such a Y we note that

$$E(F(X) - F(Y))(e^{F(X)} - e^{F(Y)}) = 2EF(X)e^{F(X)} - 2EF(X)Ee^{F(X)}$$

so that the left hand side of (8) can be rewritten

$$EF(X)Ee^{F(X)} + \frac{1}{2}E(F(X) - F(Y))(e^{F(X)} - e^{F(Y)}).$$

The first term can be bounded by the first term on the right hand side of (8) with Jensen's inequality. For the second term, note that by F Lipschitz and X, Y bounded we have

$$F(X) - F(Y) = O(|\nabla F(X)|)$$

and

$$e^{F(X)} - e^{F(Y)} = O(e^{F(X)}|\nabla F(X)|)$$

which give the claim.

Now we assume the claim for $n-1$, and write $X = (X', X_n)$, where $X' = (X_1, \ldots, X_{n-1})$. Conditioning on X_n we can apply the induction hypothesis to get

$$E(F(X)e^{F(X)}|X_n) \leq f(X_n)e^{f(X_n)} + CE(e^{F(X)}|\nabla' F(X)|^2|X_n) \tag{9}$$

where $f(X_n) := \log E(e^{F(X)}|X_n)$, and ∇' is the $n-1$-dimensional gradient. For the expectation of first term on the right hand side, we can apply the $n = 1$ case to get

$$Ef(X_n)e^{f(X_n)} \leq Ee^{f(X_n)} \log Ee^{f(X_n)} + C'E(e^{f(X_n)}|f'(X_n)|^2). \tag{10}$$

The first term on the right is just $Ee^{F(X)} \log Ee^{F(X)}$. As for the second term, we note that by the chain rule

$$f'(X_n) = e^{-f(X_n)}E(e^{F(X)}F_{x_n}(X)|X_n).$$

Squaring this and applying the Cauchy-Schwarz inequality to the conditional expection gives

$$\begin{aligned}|f'(X_n)|^2 &\leq e^{-2f(X_n)}E(e^{F(X)}|X_n)E(e^{F(X)}|F_{x_n}(X)|^2|X_n)\\ &= e^{-f(X_n)}E(e^{F(X)}|F_{x_n}(X)|^2|X_n).\end{aligned}$$

Inserting this in (10) and combining our estimates in (9) completes the induction. □

To deduce the upper tail estimate, we apply the above lemma to sF for $s > 0$:

$$EsF(X)e^{sF(X)} \leq (Ee^{sF(X)})(\log Ee^{sF(X)}) + Cs^2Ee^{sF(X)}.$$

If we let $H(s) = Ee^{sF(X)}$, we can rewrite the above as the differential inequality

$$\frac{d}{ds}(\frac{1}{s}\log H(s)) \leq C.$$

On the other hand, from Taylor expansion we see that

$$\frac{1}{s}\log H(s) \to EF(X)$$

as $s \to 0$. Combining these estimates we have

$$\frac{1}{s}\log H(s) \leq EF(X) + Cs$$

which we rewrite as

$$Ee^{sF(X)} \leq e^{sEF(X)+Cs^2}.$$

(7) now follows from Markov's inequality and optimizing in s.

4. Necessity of the convexity hypothesis

The convexity hypothesis in Corollary 2 cannot be dropped, as demonstrated by the following examples. The first was posted in the comments of [Tao], and the second is given in [Led01].

Example 7. Consider the discrete cube $\{0,1\}^n$ with Hamming (ℓ^1) distance and uniform product measure. Let

$$F(x) = \max(\min(|x|_1, n/2+\sqrt{n}), n/2-\sqrt{n})/n^{1/4}.$$

This function is $O(1)$-Lipschitz with respect to Euclidean norm. Indeed, since F only has range $O(n^{1/4})$, it suffices to verify the Lipschitz property for $|x-y|_2 = O(n^{1/4})$. Then we have $|x-y|_1 = |x-y|_2^2 \ll n^{1/4}|x-y|_2$.

The function can be extended to a $O(1)$-Lipschitz function on $\mathbf{R}^n$ which is not convex, and concentrates at scale $O(n^{1/4})$ rather than $O(1)$.

Example 8. With the same metric measure space as above, let A be the "hereditary set" $\{y : \sum_{i=1}^n y_i \le n/2\}$ (such sets $\{y : \sum_{i=1}^n y_i \le \lambda\}$ are the isoperimetric analogues of unit balls in Euclidean space with Lebsgue measure). Let $F(x) = d(x,A)$ be the Euclidean distance from $x \in \mathbf{R}^n$ to A. F is 1-Lipschitz but not convex.

Now consider the set $B_\delta = \{x : \sum_{i=1}^n x_i - n/2 \ge \delta\sqrt{n}\} \subset \{0,1\}^n$. The we have $B_\delta \subset \{x : d(x,A) \ge \sqrt{\delta}n^{1/4}\}$. Indeed, for any $x \in B_\delta$ and any $y \in A$,

$$\delta\sqrt{n} \le \sum_{i=1}^n x_i - n/2 \le \sum_{i=1}^n (x_i - y_i) \le \sum_{i=1}^n |x_i - y_i|^2.$$

Hence by monotonicity of the uniform product measure P and the central limit theorem,

$$P(\{F(x) \ge \sqrt{\delta}n^{1/4}\}) \ge P(\{\frac{\sum_{i=1}^n x_i - n/2}{\sqrt{n}} \ge \delta\}) \ge 1/10$$

for δ sufficiently small independent of n. Here again, F concentrates at scale $\gg n^{1/4}$.

5. Applications

In this section we present applications of Theorem 1 to prove concentration for eigenvalues of random matrices and for the length of the longest increasing subsequence of a random sequence. Theorem 1 and Corollary 2 have a vast number of applications; below we list a few of these with some references for the interested reader.

Applications of Corollary 2 in random matrix theory (see [Tao12]):

- Can be used to control the distance of a random vector to a hyperplane. This is related to studying the probability of singularity of (say) a $0-1$ matrix, since singularity occurs when one of the rows is in the span of the other rows.

- Concentration of Schatten norms $\|M\|_{S^k} = (\mathrm{tr}(M^k))^{1/k}$, which can be used to the problem of finding almost sure limits of moments of the (random) empirical spectral distribution to finding the limits of their expectations.
- Concentration for convex functions on the spectrum [GZ00].

Applications of Theorem 1 in combinatorial optimization (see [Tal95,Tal96,Ste97]):

- Concentration for the length of the longest increasing subsequence of a random permutation (or equivalently, of the coordinates of point selected uniformly at random from $[0,1]^n$).
- Stochastic bin packing (Talagrand's first use of his inequality): draw n numbers uniformly at random from $[0,1]$, and consider the smallest number of bins of size 1 needed to store n objects with sizes given by the n numbers. Whatever it is, it concentrates sharply around its median for n large!
- Traveling salesman problem and minimal spanning tree.
- First time passage percolation.

5.1. **An equivalent definition of convex distance.** There is an equivalent definition of the convex distance $d_C(x,A)$ that is used for applications (the above definition is only used for the proof!). Recall that the Hamming distance between points x and y in a product space $\Omega_1 \times \cdots \times \Omega_n$ is given by

$$d_H(x,y) := \#\{i : x_i \neq y_i, 1 \leq i \leq n\} = \sum_{i=1}^{n} 1_{\{x_i \neq y_i\}}.$$

For any $\alpha \in \mathbf{R}_+^{\mathbf{n}}$, consider the weighted Hamming metric

$$d_\alpha(x,y) := \sum_{i=1}^{n} \alpha_i 1_{\{x_i \neq y_i\}}$$

and let

$$D(x,A) = \sup_{|\alpha|=1} d_\alpha(x,A), \tag{11}$$

where $d_\alpha(x,A) = \inf_{y \in A} d_\alpha(x,y)$ and $|\cdot|$ denotes Euclidean norm.

Proposition 9. *$D(x,A) = d_C(x,A)$.*

Proof. We first show $D(x,A) \leq d_C(x,A)$. Let $\alpha \in \mathbf{R}_+^{\mathbf{n}}$, $|\alpha| = 1$. Then by first applying the definition of $U_A(x)$, then using the fact that the minimum of linear functional on a convex set is equal to the minimum over the set of extreme points, and finally applying the Cauchy-Schwarz inequality, we have

$$\begin{aligned} d_\alpha(x,A) &= \inf_{y\in A}\sum_{i=1}^n \alpha_i 1_{\{x_i\neq y_i\}} \\ &= \inf_{s\in U_A(x)} \alpha\cdot s \\ &= \inf_{v\in V_A(x)} \alpha\cdot v \\ &\le \inf_{v\in V_A(x)} |v| = d_C(x,A). \end{aligned}$$

Taking the supremum over α, we have the desired inequality.

For the sake of completeness we prove the reverse inequality, though the above is sufficient for our purposes. Let $z\in V_A(x)$ such that $|z| = d_C(x,A)$. If $|z|=0$ we are done. Otherwise, let $\alpha = z/|z|$ and let v be another point in $V_A(x)$. Then by convexity, $\lambda v + (1-\lambda)z \in V_A(x)$ for any $\lambda\in[0,1]$, so

$$|z|^2 \le |z+\lambda(y-z)|^2 = |z|^2 + 2\lambda z\cdot(v-z) + \lambda^2|v-z|^2.$$

Hence $0\le 2z\cdot(v-z)+\lambda|v-z|^2$, and setting $\lambda=0$ implies that $z\cdot(v-z)\ge 0$. It follows that $\alpha\cdot v\ge |z| = d_C(x,A)$ for all $v\in V_A(x)$. Then we have

$$D(x,A)\ge d_\alpha(x,A) = \inf_{v\in V_A(x)} \alpha\cdot v \ge d_C(x,A).$$

and we are done. □

The power of this formulation of convex distance is the freedom we have with the weights α_i. Indeed, note that in the convex distance definition (11), the supremum over α is taken at each point $x\in\Omega$. Hence, α can be allowed to vary with x. We now see how the game of applying Talagrand's inequality reduces to the problem of finding an appropriate weight function $\alpha(x)$. Indeed, it follows from the above proposition and Theorem 1 that

$$P(d_{\alpha(x)}(x,A)\ge t) \le \frac{1}{P(A)} e^{-t^2/4}$$

for any weight function α.

While the full convex distance is not so easy to apply directly as, say, Euclidean distance (a reason for the appeal of Corollary 2), we will in applications that a weight function often suggests itself.

For subsets A,B of a product space Ω, if we suppose that $B\subset A_t^c = \{x : d_c(x,A)\ge t\}$, then Theorem 1 and monotonicity imply $P(A)P(B)\le e^{-t^2/4}$. The following phrases the hypothesis $B\subset A_t^c$ in terms of the Hamming definition (11) of the convex hull distance.

Corollary 10. *Let $\Omega_1,\dots,\Omega_n$ be probability spaces and let Ω be the product space with product probability measure P. Let A and B be measurable subsets of Ω.*

Suppose that for every $x = (x_1, \dots, x_n) \in B$ *there is* $\alpha = \alpha(x) \in \mathbf{R}_+^{\mathbf{n}}$ *such that for all* $y = (y_1, \dots, y_n) \in A$,

$$\sum_{i=1}^{n} \alpha_i 1_{\{x_i \neq y_i\}} \geq t \left(\sum_{1}^{n} \alpha_i^2 \right)^{1/2} \tag{12}$$

for some $t > 0$. *Then*

$$P(A)P(B) \leq e^{-t^2/4}. \tag{13}$$

As before A and B will typically come to us as sets of the form $\{X \leq a\}, \{X \geq b\}$ for some random variable X, and concentration around a median MX would again be demonstrated by alternately taking a and b as the median or shifts of the median by t. We demonstrate the use of this formulation with examples.

5.2. **The top eigenvalue, take 2.** In this section we present an application the above formulation of Talagrand's inequality for the top eigenvalue of a random matrix, as originally given in [KV02], as an instructive (though lengthier) alternative path to the result obtained from Corollary 2. This is a warmup to a proof along the same lines of concentration for other eigenvalues, as given in [AKV02], which cannot follow directly from Corollary 2.

For $1 \leq i \leq j \leq n$ let x_{ij} be independent (real) random variables, $|x_{ij}| \leq 1$. Let for each $1 \leq j < i \leq n$ set $x_{ji} = x_{ij}$ and let X be the symmetric $n \times n$ matrix with entries x_{ij}. Let $\lambda_1(X) \geq \cdots \geq \lambda_n(X)$ denote the eigenvalues of X.

Theorem 11. *(Krivelevich, Vu '00)* $P(|\lambda_1(X) - M\lambda_1(X)| \geq t) \leq 4e^{-t^2/32}$.

(This theorem was proved as part of a paper on algorithms for approximating statistics of random graphs.)

Proof. First the product space with product measure: let $m = n(n+1)/2$, $\Omega^m = [-1, 1]^m$, and P be the joint distribution of the entries on Ω^m.

Let $M \in \mathbf{R}$, $t > 0$. Let $A = \{X : \lambda_1(X) \leq M\}$ and $B = \{Y : \lambda_1(Y) \geq M + t\}$. We'll occationally abuse notation by writing matrices as vectors with m entries and sometimes as doubly-indexed matrices or triangular arrays.

If we can get an inequality like (13) for this choice of A, B, we'll be able to deduce concentration around a median in the usual way. So our job now is to find, for each $Y \in B$, a vector of weights $\alpha \in \mathbf{R}_+^{\mathbf{m}}$ (alternatively written $\alpha = (\alpha_{ij})_{1 \leq i \leq j \leq n}$), such that for all $X \in A$, we have

$$\sum_{1 \leq i \leq j \leq n}^{n} \alpha_{ij} 1_{\{x_{ij} \neq y_{ij}\}} \geq ct \left(\sum_{1 \leq i \leq j \leq n} \alpha_{ij}^2 \right)^{1/2} \tag{14}$$

for some constant $c > 0$ (from the statement of the theorem one sees that we have $c = 1/2\sqrt{2}$).

So how can we get $\alpha(Y)$ that will accomplish (14))? We'll get it from the eigenvector of Y associated to its largest eigenvalue.

> **The Answer:**
> For $Y \in B$, let v be the unit eigenvector of Y associated to the eigenvalue $\lambda_1(Y)$. For $1 \le i \le n$, let $\alpha_{ii} = v_i^2$, and for $1 \le i < j \le n$, let $\alpha_{ij} = 2|v_i||v_j|$.

Let's proceed as if we didn't know this. For given X and Y, we want to translate the inequalities $\lambda_1(X) \le M$, $\lambda_1(Y) \ge M + t$ into an inequality like (14) involving the entries of the matrices. We can use v to do this: we have $v^tYv = \lambda_1(Y) \ge M + t$ and $v^tXv \le \lambda_1(X) \le M$. Now we follow our noses:

$$\begin{aligned} t \le v^t(Y - X)v &= \sum_{1 \le i,j \le n} v_i(y_{ij} - x_{ij})v_j \\ &\le 2 \sum_{1 \le i,j \le n} |v_i||v_j| 1_{\{x_{ij} \ne y_{ij}\}} \\ &= 2 \sum_{1 \le i \le j \le n} \alpha_{ij} 1_{\{x_{ij} \ne y_{ij}\}} \end{aligned}$$

where α_{ij} is as defined in **The Answer**.

It only remains to get a bound on the Euclidean norm of α.

$$\sum_{1 \le i \le j \le n} \alpha_{ij}^2 \quad = \textstyle\sum_{i=1}^n v_i^4 + 4 \sum_{1 \le i < j \le n} |v_i|^2 |v_j|^2 \tag{15}$$

$$\le 2 \left(\textstyle\sum_{i=1}^n v_i^2 \right)^2 = 2. \tag{16}$$

So for our choice of α we have

$$\sum_{1 \le i \le j \le n}^{n} \alpha_{ij} 1_{\{x_{ij} \ne y_{ij}\}} \ge t/2 \ge \frac{t}{2\sqrt{2}} \left(\sum_{1 \le i \le j \le n} \alpha_{ij}^2 \right)^{1/2}.$$

The result follows from Corollary 10 and substituting $M = M\lambda_1(X)$ and $M = M\lambda_1(X) - t$. $\square$

Note that we can get the same concentration result for the bottom eigenvalue λ_n by taking $A = \{X : \lambda_n(X) \ge M + t\}, B = \{Y : \lambda_n(Y) \le M\}$ (but still using the eigenvector of Y!).

Remark 12. Notice the slight inefficiency in passing from (15) to (16): we could let the diagonal entries vary on a set of diameter $2\sqrt{2}$ instead of 2 and get the same estimate. This is an artifact of double counting the off diagonal entries while only counting the diagonal ones once.

5.3. **Other eigenvalues.** The above approach can actually be extended to obtain concentration for other eigenvalues, which are non-convex statistics and hence do not admit immediate application of Corollary 2. (We note however that with some additional effort a strengthening of the following result was actually obtained by Meckes using Corollary 2 [Mec04].)

Theorem 13. *(Alon, Krivelevich, Vu '02) (Same setup as before.) For each* $1 \leq s \leq n$,

$$P(|\lambda_s(X) - M\lambda_s(X)| \geq t) \leq 4e^{-t^2/32s^2}. \tag{17}$$

Proof. With the same setup as before, we now fix $1 \leq s \leq n$, let $M \in \mathbf{R}$, $t > 0$, and let $A = \{X : \lambda_s(X) \leq M\}, B = \{Y : \lambda_s(Y) \geq M + t\}$.

Now what are the proper weights α? Here we construct them from the s unit eigenvectors of Y associated to its s largest eigenvalues.

For fixed Y, let $v^1, \dots v^s$ be those eigenvectors of Y, and for $1 \leq p \leq s$ we write $v^p = (v^p_1, \dots, v^p_n)$.

The Answer:
For $1 \leq i \leq n$, let $\alpha_{ii} = \sum_{p=1}^s (v_i^p)^2$, and for $1 \leq i < j \leq n$, let $\alpha_{ij} = 2\sqrt{\sum_{p=1}^s (v_i^p)^2}\sqrt{\sum_{p=1}^s (v_j^p)^2}$. Then

$$\sum_{1 \leq i \leq j \leq n} \alpha_{ij}^2 \leq 2s^2 \tag{18}$$

and $\forall X \in A$,

$$\sum_{1 \leq i \leq j \leq n} \alpha_{ij} 1_{\{x_{ij} \neq y_{ij}\}} \geq t/2 \tag{19}$$

.

As before we first show (19)). The key intuition comes from the Courant-Fischer minimax characterization of eigenvalues: we can find a unit vector u in the span of the first s eigenvectors of Y that is orthogonal to the span of the largest $s-1$ eigenvectors of X. This is how we narrow in on λ_s.

Let u be such a vector and write $u = \sum_{p=1}^s c_p v^p$. We have $u^t X u \leq \lambda_s(X) \leq M$ and $u^t Y u \geq \lambda_s(Y) \geq M + t$. Hence,

$$\begin{aligned}
t \leq u^t(Y - X)u &= \sum_{1 \leq i,j \leq n} u_i(y_{ij} - x_{ij})u_j \\
&= \sum_{1 \leq i,j \leq n} (y_{ij} - x_{ij})(\sum_{p=1}^s c_p v_i^p)(\sum_{q=1}^s c_q v_j^q) \\
&\leq 2 \sum_{1 \leq i,j \leq n} |\sum_{p=1}^s c_p v_i^p||\sum_{q=1}^s c_q v_j^q| 1_{\{y_{ij} \neq x_{ij}\}} \\
&\leq 2 \sum_{1 \leq i,j \leq n} (\sum_{p=1}^s c_p^2)^{1/2}(\sum_{p=1}^s (v_i^p)^2)^{1/2}(\sum_{p=1}^s c_p^2)^{1/2}(\sum_{p=1}^s (v_i^p)^2)^{1/2} \\
&= 2 \sum_{1 \leq i,j \leq n} (\sum_{p=1}^s (v_i^p)^2)^{1/2}(\sum_{p=1}^s (v_i^p)^2)^{1/2}.
\end{aligned}$$

Letting α as in the answer we have (19)).

Now for (18)):

$$\sum_{1\le i\le j\le n}\alpha_{ij}^2=\sum_{i=1}^n(\sum_{p=1}^s(v_i^p)^2)^2+4\sum_{1\le i<j\le n}(\sum_{p=1}^s(v_i^p)^2)(\sum_{p=1}^s(v_j^p)^2)$$
$$\le 2(\sum_{i=1}^n\sum_{p=1}^s(v_i^p)^2)^2=2(\sum_{p=1}^s\sum_{i=1}^n(v_i^p)^2)^2=2s^2.$$

The same estimate for $\lambda_{n-s+1}(X)$ follows as before by letting $B=\{Y:\lambda_{n-s+1}(Y)\le M\}$, $A=\{X:\lambda_{n-s+1}(X)\ge M+t\}$ and finding a u in the span of the eigenvectors of Y associated to its smallest s eigenvectors, and orthogonal to the corresponding $s-1$ eigenvectors of X. $\square$

Remark 14. Our concentration gets worse as we head deeper into the "bulk" of the spectrum. However, we have reason to believe concentration in the bulk should be the same, as this can be verified for Gaussian entries using Gaussian concentration estimates for Lipschitz functions (which do not require the hypothesis of convexity). This holds more generally for any measure that obeys a log-Sobolev inequality (of which the Gaussian is the model example). Our functional logarithmic Sobolev inequality in section 3 was obtained with the added assumption that the function is convex, and no hypotheses on the measure apart from boundedness.

5.4. **The longest increasing subsequence.** Let $X=(x_1,\dots,x_n)$ be uniformly distributed in $\Omega=[0,1]^n$, let $J(x)$ be the longest increasing subsequence of $\{x_1,\dots,x_n)$, and let $F_n(x)=|J(x)|$ be its length. We will use Corollary 10 to show that F_n concentrates tightly around its median MF_n.

Note that we cannot apply Corollary 2 as F_n is not convex. For instance, with $n=3$, taking $x=(0,1,.6)$ and $y=(.8,0,.6)$ we have that $F_3(x)=2=F_3(y)$, but $F_3(\frac{x+y}{2})=F_3((.4,.5,.6))=3$. However, F_n is 1-Lipschitz with respect to the Hamming metric, which is essentially what makes it ammenable to analysis with the deeper Theorem 1.

We let $a>0$ and $A=\{F(y)\le a\}$. Our construction of the weights α begins with a simple observation: for any $x,y\in\Omega$,

$$F_n(y)\ge F_n(x)-\sum_{i=1}^n 1_{\{x_i\in J(x),x_i\ne y_i\}}. \tag{20}$$

Indeed, if $J(x)$ is the longest increasing subsequence of x, then by taking those elements of $J(x)$ shared by y we have an increasing subsequence of y, the length of which is bounded by $F_n(y)$. From this we have

$$F_n(x)-F_n(y)\le\sum_{i=1}^n 1_{\{x_i\in J(x)\}}1_{\{x_i\ne y_i\}}$$

which is starting to look like (12)). If we let

$$\alpha(x)=\frac{1}{\sqrt{F_n(x)}}1_{J(x)} \tag{21}$$

we have

$$d_{\alpha(x)}(x,y) = \frac{1}{\sqrt{F_n(x)}} \sum_{i=1}^{n} 1_{\{x_i \in J(x)\}} 1_{\{x_i \neq y_i\}}$$
$$\geq \frac{1}{\sqrt{F_n(x)}} (F_n(x) - F_n(y)).$$

For the convex distance from x to A we have

$$d_C(x,A) \geq \frac{F_n(x) - a}{\sqrt{F_n(x)}}.$$

For $t \geq a$ the function $g(t) = (t-a)/\sqrt{t}$ is monotone increasing. From this and Theorem 1 it follows that

$$P(F_n(x) \geq a+t) \leq P(\frac{F_n(x) - a}{\sqrt{F_n(x)}} \geq \frac{t}{a+t})$$
$$\leq P(d_C(x,A) \geq \frac{t}{a+t})$$
$$\leq \frac{1}{P(A)} e^{-t^2/4(a+t)}.$$

Taking $a = M_n := MF_n(x)$ we get the upper tail estimate

$$P(F_n(x) \geq M_n + t) \leq 2\exp(-\frac{t^2}{4(M_n+t)})$$

and taking $a = M_n - t$ we get the lower tail estimate

$$P(F_n(x) \leq M_n - t) \leq 2\exp(-\frac{t^2}{4M_n}).$$

Hence, for large n $F_n(x)$ is roughly $M_n + O(\sqrt{M_n})$. It can be shown by elementary means (see e.g. [Ste97]) that $M_n = O(\sqrt{n})$, so the above concentration estimates are enough to prove a strong law of the form $(F_n(x) - M_n)/\sqrt{n} \to 0$ a.s.

Remark 15. Much more is known about the asymptotics of F_n. For instance, it has been shown by a variety of methods that $F_n/\sqrt{n} \to 2$ almost surely. Furthermore, in [BDJ99] it was shown that $n^{-1/6}(F_n - 2\sqrt{n})$ converges in distribution to the Tracy-Widom law, putting it in the same universality class as our first example, the top eigenvalue of a random Hermitian matrix. For an entertaining survey of these results and their relation to *patience sorting* of cards, see [AD99].

References

[AD99] David Aldous and Persi Diaconis. Longest increasing subsequences: from patience sorting to the Baik-Deift-Johansson theorem. *Bull. Amer. Math. Soc. (N.S.)*, 36(4):413–432, 1999.

[AKV02] Noga Alon, Michael Krivelevich, and Van H. Vu. On the concentration of eigenvalues of random symmetric matrices. *Israel J. Math.*, 131:259–267, 2002.

[AS08] Noga Alon and Joel H. Spencer. *The probabilistic method*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Inc., Hoboken, NJ, third edition, 2008. With an appendix on the life and work of Paul Erdos.

[BDJ99] Jinho Baik, Percy Deift, and Kurt Johansson. On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.*, 12(4):1119–1178, 1999.

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

[KV02] Michael Krivelevich and Van H. Vu. Approximating the independence number and the chromatic number in expected polynomial time. *J. Comb. Optim.*, 6(2):143–155, 2002.

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

[Mec04] Mark W. Meckes. Concentration of norms and eigenvalues of random matrices. *J. Funct. Anal.*, 211(2):508–524, 2004.

[Ste97] J. Michael Steele. *Probability theory and combinatorial optimization*, volume 69 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[Tal95] Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Inst. Hautes Études Sci. Publ. Math.*, (81):73–205, 1995.

[Tal96] Michel Talagrand. A new look at independence. *Ann. Probab.*, 24(1):1–34, 1996.

[Tao] Terence Tao. http://terrytao.wordpress.com/2010/01/03/254a-notes-1-concentration-of-measure/.

[Tao12] Terence Tao. *Topics in random matrix theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.