Concentration of Measure and the Compact Classical Matrix Groups

Elizabeth Meckes

Program for Women and Mathematics 2014

Institute for Advanced Study
and
Princeton University

Lecture 1

Introduction to the compact classical matrix groups

1.1 What is an orthogonal/unitary/symplectic matrix?

The main question addressed in this lecture is "what is a random orthogonal/unitary/symplectic matrix?", but first, we must address the preliminary question: "What is an orthogonal/unitary/sympletic matrix?"

Definition.

1. An $n \times n$ matrix U over $\mathbb{R}$ is **orthogonal** if
$$UU^T = U^T U = I_n, \tag{1.1}$$
where I_n denotes the $n \times n$ identity matrix, and U^T is the transpose of U. The set of $n \times n$ orthogonal matrics over $\mathbb{R}$ is denoted $\mathbb{O}(n)$.
2. An $n \times n$ matrix U over $\mathbb{C}$ is **unitary** if
$$UU^* = U^* U = I_n, \tag{1.2}$$
where U^* denotes the conjugate transpose of U. The set of $n \times n$ unitary matrices over $\mathbb{C}$ is denoted $\mathbb{U}(n)$.
3. An $2n \times 2n$ matrix U over $\mathbb{C}$ is symplectic[1] if $U \in \mathbb{U}(2n)$ and
$$UJU^* = U^* JU = J, \tag{1.3}$$
where
$$J := \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}.$$
The set of $2n \times 2n$ symplectic matrices over $\mathbb{C}$ is denoted $\mathbb{Sp}(2n)$.

[1] Alternatively, you can define the symplectic group to be $n \times n$ matrices U with quaternionic entries, such that $UU^* = I_n$, where U^* is the (quaternionic) conjugate transpose. You can represent quaternions as 2×2 matrices over $\mathbb{C}$, and then these two definitions should be the same. Honestly, I got sick of it before I managed to grind it out, but if you feel like it, go ahead.

Note that it is immediate from the definitions that U is orthogonal if and only if U^T is orthogonal, and U is unitary or symplectic if and only if U^* is.

The algebraic definitions given above are the most standard and the most compact. However, it's often more useful to view things more geometrically. (Incidentally, from now on in these lectures, we'll mostly follow the nearly universal practice in this area of mathematics of ignoring the symplectic group most, if not all, of the time.)

One very useful viewpoint is the following.

Lemma 1.1. *Let M be an $n \times n$ matrix over $\mathbb{R}$. Then M is orthogonal if and only if the columns of M form an orthonormal basis of $\mathbb{R}^n$. Similarly, if M is an $n \times n$ matrix over $\mathbb{C}$, then M is unitary if and only if the columns of M form an orthonormal basis of $\mathbb{C}^n$.*

Proof. Note that the $(i,j)^{th}$ entry of U^TU (if U has real entries) or U^*U (if U has complex entries) is exactly the inner product of the ith and jth columns of U. So $U^TU = I_n$ or $U^*U = I_n$ is exactly the same thing as saying the columns of U form an orthonormal basis of $\mathbb{R}^n$ or $\mathbb{C}^n$. □

If we view orthogonal (resp. unitary) matrices as maps on $\mathbb{R}^n$ (resp. $\mathbb{C}^n$), we see even more important geometric properties.

Lemma 1.2.

1. *For U an $n \times n$ matrix over $\mathbb{R}$, $U \in \mathbb{O}(n)$ if and only if U acts as an isometry on $\mathbb{R}^n$; that is,*
$$\langle Uv, Uw \rangle = \langle v, w \rangle \qquad for\, all\, v, w \in \mathbb{R}^n.$$
2. *For U an $n \times n$ matrix over $\mathbb{C}$, $U \in \mathbb{U}(n)$ if and only if U acts as an isometry on $\mathbb{C}^n$:*
$$\langle Uv, Uw \rangle = \langle v, w \rangle \qquad for\, all\, v, w \in \mathbb{C}^n.$$

Proof. Exercise. □

Another important geometric property of matrices in $\mathbb{O}(n)$ and $\mathbb{U}(n)$ is the following.

Lemma 1.3. *If U is an orthogonal or unitary matrix, then $|\det(U)| = 1$.*

Proof. If $U \in \mathbb{O}(n)$, then

$$1 = \det(I) = \det(UU^T) = \det(U)\det(U^T) = \big[\det(U)\big]^2.$$

If $U \in \mathbb{U}(n)$, then

$$1 = \det(I) = \det(UU^*) = \det(U)\det(U^*) = \big|\det(U)\big|^2.$$

□

We sometimes restrict our attention to the so-called "special" counterparts of the orthogonal and unitary groups, defined as follows.

Definition. The set $\mathbb{SO}(n) \subseteq \mathbb{O}(n)$ of **special orthogonal matrices** is defined by

$$\mathbb{SO}(n) := \{U \in \mathbb{O}(n) : \det(U) = 1\}.$$

The set $\mathbb{SU}(n) \subseteq \mathbb{U}(n)$ of **special unitary matrices** is defined by

$$\mathbb{SU}(n) := \{U \in \mathbb{U}(n) : \det(U) = 1\}.$$

A final (for now) important observation is that the sets $\mathbb{O}(n)$, $\mathbb{U}(n)$, $\mathbb{Sp}(2n)$, $\mathbb{SO}(n)$, and $\mathbb{SU}(n)$ are *compact Lie groups*; that is, they are groups (with matrix multiplication as the operation), and they are manifolds. For now we won't say much about their structure as manifolds, but right away we will need to see that they can all be seen as subsets of Euclidean space – $\mathbb{O}(n)$ and $\mathbb{SO}(n)$ can be thought of as subsets of $\mathbb{R}^{n^2}$; $\mathbb{U}(n)$ and $\mathbb{SU}(n)$ can be seen as subsets of $\mathbb{C}^{n^2}$ and $\mathbb{Sp}(2n)$ can be seen as a subset of $\mathbb{C}^{(2n)^2}$. You could see this by just observing that there are n^2 entries in an $n \times n$ matrix and leave it at that, but it's helful to say a bit more. No matter how we organize the entries of, say, a matrix $A \in \mathbb{O}(n)$ in a vector of length n^2, it will be the case that if $A, B \in \mathbb{O}(n)$ with entries $\{a_{ij}\}$ and $\{b_{ij}\}$, and $\vec{A}$ and $\vec{B}$ denote the vectors in $\mathbb{R}^{n^2}$ to which we have associated A and B, then

$$\left\langle \vec{A}, \vec{B} \right\rangle = \sum_{i,j=1}^{n} a_{ij} b_{ij} = \mathrm{Tr}(AB^T).$$

So rather than fuss about how exactly to write a matrix as a vector, we often talk instead about the Euclidean spaces $M_n(\mathbb{R})$ (resp. $M_n(\mathbb{C})$) of $n \times n$ matrices over $\mathbb{R}$ (resp. $\mathbb{C}$), with inner products

$$\langle A, B \rangle := \mathrm{Tr}(AB^T)$$

for $A, B \in M_n(\mathbb{R})$, and

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

for $A, B \in M_n(\mathbb{C})$. These inner products are called the **Hilbert-Schmidt** inner products on matrix space. The norm induced by the Hilbert-Schmidt inner product is sometimes called the Frobenius norm or the Schatten 2-norm.

Notice that the discussion above presents us with two ways to talk about distance within the compact classical matrix groups: we can use the Hilbert-Schmidt inner product and define the distance between two matrices A and B by

$$d_{HS}(A, B) := \|A - B\|_{HS} := \sqrt{\langle A - B, A - B \rangle_{HS}} = \sqrt{\mathrm{Tr}\left[(A - B)(A - B)^*\right]}. \quad (1.4)$$

On the other hand, since for example $A, B \in \mathbb{U}(n)$ can be thought of as living in a submanifold of Euclidean space $M_n(\mathbb{C})$, we could consider the *geodesic distance* $d_g(A, B)$ between A and B; that is, the length, as measured by the Hilbert-Schmidt metric, of the shortest path lying entirely in $\mathbb{U}(n)$ between A and B. In the case of $\mathbb{U}(1)$, this is arc-length distance, whereas the Hilbert-Schmidt distance defined in Equation (1.4) is the straight-line distance between two points on the circle. It doesn't make much difference which of these two distances you use, though:

Lemma 1.4. *Let $A, B \in \mathbb{U}(n)$. Then*

$$d_{HS}(A,B) \leq d_g(A,B) \leq \frac{\pi}{2} d_{HS}(A,B).$$

Exercise 1.5. Prove Lemma 1.4:

1. Observe that $d_{HS}(A,B) \leq d_g(A,B)$ trivially.
2. Show that $d_g(A,B) \leq \frac{\pi}{2} d_{HS}(A,B)$ for $A, B \in \mathbb{U}(1)$; that is, that arc-length on the circle is bounded above by $\frac{\pi}{2}$ times Euclidean distance.
3. Show that both $d_{HS}(\cdot,\cdot)$ and $d_g(\cdot,\cdot)$ are translation-invariant; that is, if $U \in \mathbb{U}(n)$, then
$$d_{HS}(UA,UB) = d_{HS}(A,B) \qquad \text{and} \qquad d_g(UA,UB) = d_g(A,B).$$
4. Show that it suffices to assume that $A = I_n$ and B is diagonal.
5. If $A = I_n$ and $B = \left[diag(e^{i\theta_1}, \ldots, e^{i\theta_n})\right]$, compute the length of the geodesic from A to B given by $U(t) := \left[diag(e^{it\theta_1}, \ldots, e^{it\theta_n})\right]$, for $0 \leq t \leq 1$.
6. Combine parts 2, 4, and 5 to finish the proof.

We observed above that orthogonal and unitary matrices act as isometries on $\mathbb{R}^n$ and $\mathbb{C}^n$; it is also true that they act as isometries on their respective matrix spaces, via matrix multiplication.

Lemma 1.6. *If $U \in \mathbb{O}(n)$ (resp. $\mathbb{U}(n)$), then the map $T_U : M_n(\mathbb{R}) \to M_n(\mathbb{R})$ (resp. $T_U : M_n(\mathbb{C}) \to M_n(\mathbb{C})$) defined by*

$$T_U(M) = UM$$

is an isometry on $M_n(\mathbb{R})$ (resp. $M_n(\mathbb{C})$) with respect to the Hilbert-Schmidt inner product.

Proof. Exercise. □

1.2 What is a *random* $\mathbb{O}(n)$/$\mathbb{U}(n)$/$\mathbb{S}\mathbb{p}(2n)$ matrix?

The most familiar kind of random matrix is probably one described as something like: "take an empty $n \times n$ matrix, and fill in the entries with independent random variables, with some prescribed distributions". Thinking of matrices as the collections of their entries is very intuitive and appealing in some contexts, but less so in ours. Since orthogonal matrices are exactly the linear isometries of $\mathbb{R}^n$, they are inherently geometric objects, and the algebraic conditions defining orthogonality, etc., are about the relationships among the entries that create that natural geometric property.

The situation is analogous to thinking about, say, a point on the circle in $\mathbb{R}^2$ (it's a generalization of that, actually, since $\mathbb{S}^1 \subseteq \mathbb{C}$ is exactly $\mathbb{U}(1)$). We can think of a point on the circle as $z = x + iy$ with the condition that $x^2 + y^2 = 1$, but that's a bit unwieldy, and definitely doesn't lead us directly to any ideas about how to describe a "uniform random

point" on the circle. It's much more intuitive to think about the circle as a geometric object: what we should mean by a "uniform random point on the circle" should be a complex random variable taking values in $\mathbb{S}^1 \subseteq \mathbb{C}$, whose distribution is *rotation invariant*; that is, if $A \subseteq \mathbb{S}^1$, the probability of our random point lying in A should be the same as the probability that it lies in $e^{i\theta}A := \{e^{i\theta}a : a \in A\}$.

The story with the matrix groups is similar: if G is one of the matrix groups defined in the last section, a "uniform random element" of G should be a random $U \in G$ whose distribution is *translation invariant*; that is, if $M \in G$ is any fixed matrix, then we should have the equality in distribution

$$MU \stackrel{d}{=} UM \stackrel{d}{=} U.$$

Alternatively, the distribution of a uniform random element of G should be a translation invariant probability measure μ on G: for any subset $\mathcal{A} \subseteq G$ and any fixed $M \in G$,

$$\mu(M\mathcal{A}) = \mu(\mathcal{A}M) = \mu(\mathcal{A}),$$

where $M\mathcal{A} := \{MU : U \in \mathcal{A}\}$ and $\mathcal{A}M := \{UM : U \in \mathcal{A}\}$.

It turns out that there is one, and only one, way to do this.

Theorem 1.7. *Let G be any of $\mathbb{O}(n)$, $\mathbb{SO}(n)$, $\mathbb{U}(n)$, $\mathbb{SU}(n)$, or $\mathbb{Sp}(2n)$. Then there is a unique translation-invariant probability measure (called **Haar measure**) on G.*

The theorem is true in much more generality (in particular, any compact Lie group has a Haar probability measure), but we won't worry about that, or the proof of the theorem. Also, in general one has to worry about invariance under left-translation or right-translation, since they could be different. In the case of compact Lie groups, left-invariance implies right-invariance and vice versa, so I will rather casually just talk about "translation-invariance" without specifying the side, and using both sides if it's convenient.

Exercise 1.8.

1. Prove that a translation-invariant probability measure on $\mathbb{O}(n)$ is invariant under transposition: if U is Haar-distributed, so is U^T.
2. Prove that a translation-invariant probability measure on $\mathbb{U}(n)$ is invariant under transposition and under conjugation: if U is Haar-distributed, so are both U^T and U^*.

The theorem above is an existence theorem which doesn't itself tell us how to describe Haar measure in specific cases. In the case of the circle, you already are very familiar with the right measure: (normalized) arc length. That is, we measure an interval on the circle by taking its length and dividing by 2π, and if we're feeling fussy we turn the crank of measure theory to get a *bona fide* probability measure.

In the case of the matrix groups, we will describe three rather different-seeming constructions that all lead back to Haar measure. We'll specialize to the orthogonal group for simplicity, but the constructions for the other groups are similar.

The Riemannian approach

We've already observed that $\mathbb{O}(n) \subseteq \mathbb{R}^{n^2}$, and that it is a compact submanifold.

Quick Exercise 1.9. What is $\mathbb{O}(1)$?

Incidentally, as you've just observed in the $n = 1$ case, $\mathbb{O}(n)$ is not a connected manifold – it splits into two pieces: $\mathbb{SO}(n)$ and what's sometimes called $\mathbb{SO}^{-}(n)$, the set of matrices $U \in \mathbb{O}(n)$ with $\det(U) = -1$.

Exercise 1.10. Describe both components of $\mathbb{O}(2)$.

Because $\mathbb{O}(n)$ sits inside of the Euclidean matrix space $M_n(\mathbb{R})$ (with the Hilbert-Schmidt inner product), it has a Riemannian metric that it inherits from the Euclidean metric. Here's how that works: a Riemannian metric is a gadget that tells you how to take inner products of two vectors which both lie in the tangent space to a manifold at a point. For us, this is easy to understand, because our manifold $\mathbb{O}(n)$ lies inside Euclidean space: to take the inner products of two tangent vectors at the same point, we just shift the two vectors to be based at the origin, and then take a dot product the usual way. An important thing to notice about this operation is that it is invariant under multiplication by a fixed orthogonal matrix: if $U \in \mathbb{O}(n)$ is fixed and I apply the map T_U from Lemma 1.6 (i.e., multiply by U) to $M_n(\mathbb{R})$ and then take the dot product of the images of two tangent vectors, it's the same as it was before. The base point of the vectors changes, but the fact that T_U is an isometry exactly means that their dot product stays the same. (Incidentally, this is essentially the solution to the second half of part 3 of Exercise 1.5.)

Now, on any Riemannian manifold, you can use the Riemannian metric to define a natural notion of volume, which you can write a formula for in coordinates if you want. The discussion above means that the volume form we get on $\mathbb{O}(n)$ is translation-invariant; that is, it's Haar measure.

An explicit geometric construction

Recall that Lemma 1.1 said that U was an orthogonal matrix if and only if its columns were orthonormal. One way to construct Haar measure on $\mathbb{O}(n)$ is to add entries to an empty matrix column by column (or row by row), as follows. First choose a random vector U_1 uniformly from the sphere $\mathbb{S}^{n-1} \subseteq \mathbb{R}^n$ (that is, according to the probability measure defined by normalized surface area). Make U_1 the first column of the matrix; by construction, $\|U_1\| = 1$. The next column will need to be orthogonal to U_1, so consider the unit sphere in the orthogonal complement of U_1; that is, look at the submanifold of $\mathbb{R}^n$ defined by

$$\left(U_1^{\perp}\right) \cap \mathbb{S}^{n-1} = \left\{x \in \mathbb{R}^n : \|x\| = 1, \langle x, U_1 \rangle = 0\right\}.$$

Q1.9: $\{\mp 1\}$

This is just a copy of the sphere $\mathbb{S}^{n-2}$ sitting inside a (random) $n-1$-dimensional subspace of $\mathbb{R}^n$, so we can choose a random vector $U_2 \in \left(U_1^{\perp}\right) \cap \mathbb{S}^{n-1}$ according to normalized surface area measure, and let this be the second column of the matrix. Now we continue in the same way; we pick each column to be uniformly distributed in the unit sphere of vectors which are orthogonal to each of the preceding columns. The resulting matrix $\begin{bmatrix} | & & | \\ U_1 & \dots & U_n \\ | & & | \end{bmatrix}$ is certainly orthogonal, and moreover, it turns out that is also Haar-distributed.

Observe that if M is a fixed orthogonal matrix, then

$$M \begin{bmatrix} | & & | \\ U_1 & \dots & U_n \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ MU_1 & \dots & MU_n \\ | & & | \end{bmatrix}.$$

So the first column of $M \begin{bmatrix} | & & | \\ U_1 & \dots & U_n \\ | & & | \end{bmatrix}$ is constructed by choosing U_1 uniformly from $\mathbb{S}^{n-1}$ and then multiplying by M. But $M \in \mathbb{O}(n)$ means that M acts as a linear isometry of $\mathbb{R}^n$, so it preserves surface area measure on $\mathbb{S}^{n-1}$. (If you prefer to think about calculus, if you did a change of variables $y = Mx$, where $x \in \mathbb{S}^{n-1}$, then you would have $y \in \mathbb{S}^{n-1}$ too and the Jacobian of the change of variables is $|\det(M)| = 1$.) That is, the distribution of MU_1 is exactly uniform on $\mathbb{S}^{n-1}$.

Now, since M is an isometry, $\langle MU_2, MU_1 \rangle = 0$ and, by exactly the same argument as above, MU_2 is *uniformly distributed* on

$$(MU_1)^{\perp} \cap \mathbb{S}^{n-1} := \{x \in \mathbb{R}^n : |x| = 1, \langle MU_1, x \rangle = 0\}.$$

So the second column of $M\left[U_1 \dots U_n\right]$ is distributed uniformly in the unit sphere of the orthogonal complement of the first column.

Continuing the argument, we see that the distribution of $M\left[U_1 \dots U_n\right]$ is exactly the same as the distribution of $\left[U_1 \dots U_n\right]$, so this construction is left-invariant. By uniqueness of Haar-measure, this means that our construction *is* Haar measure.

The Gaussian approach

This is probably the most commonly used way to describe Haar measure, and also one that's easy to implement on a computer.

We start with an empty $n \times n$ matrix, and fill it with independent, identically distributed (i.i.d.) standard Gaussian entries $\{x_{i,j}\}$ to get a random matrix X. That is, the joint density (with respect to $\prod_{i,j=1}^{n} dx_{ij}$) of the n^2 entries of X is given by

$$\frac{1}{(2\pi)^{n^2}} \prod_{i,j=1}^{n} e^{-\frac{x_{ij}^2}{2}} = \frac{1}{(2\pi)^{n^2}} \exp\left\{-\frac{1}{2}\sum_{i,j=1}^{n} x_{i,j}^2\right\}.$$

The distribution of X is invariant under multiplication by an orthogonal matrix: by the change of variables $y_{ij} := \left[MX\right]_{ij} = \sum_{k=1}^{n} M_{ik}x_{kj}$, the density of the entries of MX with respect to $\prod dy_{ij}$ is

$$\frac{|\det(M^{-1})|}{(2\pi)^2}\exp\left\{-\frac{1}{2}\sum_{i,j=1}^{n}[M^{-1}y]_{ij}^2\right\} = \frac{1}{(2\pi)^{n^2}}\exp\left\{-\frac{1}{2}\sum_{i,j=1}^{n}y_{i,j}^2\right\},$$

since M^{-1} is an isometry.

So filling a matrix with i.i.d. standard Gaussians gives us something invariant under left-multiplication by an orthogonal matrix, but this isn't a Haar-distributed orthogonal matrix, because it's not orthogonal! To take care of that, we make it orthogonal: we use the Gram-Schmit process. Fortunately, performing the Gram-Schmidt process commutes with multiplication by a fixed orthogonal matrix M: let X_i denote the columns of X. Then, for example, when we remove the X_1 component from X_2, we replace X_2 with $X_2 - \langle X_1, X_2\rangle X_2$. If we then multiply by M, the resulting second column is

$$MX_2 - \langle X_1, X_2\rangle MX_2.$$

If, on the other hand, we first multiply X by M, we have a matrix with columns $MX_1, \ldots, MX_n$. If we now remove the component in the direction of column 1 from column 2, our new column 2 is

$$MX_2 - \langle MX_1, MX_2\rangle MX_1 = MX_2 - \langle X_1, X_2\rangle MX_2,$$

since M is an isometry.

What we have then, is that if we fill a matrix X with i.i.d. standard normal random variables, perform the Gram-Schmidt process, and then multiply by M, that is the same as applying the Gram-Schmidt process to MX, which we saw above has the same distribution as X itself. In other words, the probability measure constructed this way gives us a random orthogonal matrix whose distribution is invariant under left-multiplication by a fixed orthogonal matrix: we have constructed Haar measure (again).

Note: if you're more familiar with the terminology, it may help to know that what we checked above is that if you fill a matrix with i.i.d. standard Gaussian entries and write its QR-decomposition, the Q part is exactly a Haar-distributed random orthogonal matrix.

Haar measure on $\mathbb{SO}(n)$ and $\mathbb{SO}^-(n)$

The constructions above describe how to choose a uniform random matrix from $\mathbb{O}(n)$, but as we noted above, $\mathbb{O}(n)$ decomposes very neatly into two pieces, those matrices with determinant 1 ($\mathbb{SO}(n)$) and those with determinant -1 ($\mathbb{SO}^-(n)$). Theorem 1.7 says that $\mathbb{SO}(n)$ has a unique translation-invariant probability measure; it's easy to see that it's exactly what you get by restricting Haar measure on $\mathbb{O}(n)$.

There is also a measure that we call Haar measure on $\mathbb{SO}^-(n)$, which is what you get by restricting Haar measure from $\mathbb{O}(n)$. The set $\mathbb{SO}^-(n)$ isn't a group, it's a coset of the subgroup $\mathbb{SO}(n)$ in the group $\mathbb{O}(n)$; we continue to use the name Haar measure on $\mathbb{SO}^-(n)$ even though we don't have a translation-invariant measure on a group, because what we have instead is a probability measure which is invariant under translation within $\mathbb{SO}^-(n)$ by any matrix from $\mathbb{SO}(n)$. There is a neat connection between Haar measure on $\mathbb{SO}(n)$ and Haar measure on $\mathbb{SO}^-(n)$: if U is Haar-distributed in $\mathbb{SO}(n)$ and $\tilde{U}$ is any fixed matrix in $\mathbb{SO}^-(n)$, then $\tilde{U}U$ is Haar-distributed in $\mathbb{SO}^-(n)$.

Exercise 1.11. Carefully check the preceding claim.

1.3 Who cares?

"Random matrices" are nice buzz-words, and orthogonal or unitary matrices sound natural enough, but why invest one's time and energy in developing a theory about these objects?

Much of the original interest in random matrices from the compact classical groups (mainly the unitary group) stems from physics. One big idea is that, in quantum mechanics, the energy levels of a quantum system are described by the eigenvalues of a Hermitian operator (the Hamiltonian of the system). If we can understand some important things about the operator by considering only finitely many eigenvalues (that is, working on a finite-dimensional subspace of the original domain), we're led to think about matrices. These matrices are too complicated to compute, but a familiar idea from statistical physics was that under such circumstances, you could instead think probabilistically – consider a random matrix assumed to have certain statistical properties, and try to understand what the eigenvalues are typically like, and hope that this is a good model for the energy levels of quantum systems. Some of the most obvious statistical models for random matrices lacked certain symmetries that seemed physically reasonable, so Freeman Dyson proposed considering random unitary matrices. They were not meant to play the role of Hamiltonians, but rather to encode the same kind of information about the quantum system, at least in an approximate way.

A somewhat more recent and initially rather startling connection is between random unitary matrices and number theory, specifically, properties of the zeroes of the Riemann zeta function. Remarkably enough, this connection was noticed through physics. The story goes that Hugh Montgomery gave a talk at Princeton about some recent work on conjectures about the pair correlations of the zeroes of the Riemann zeta function. Dyson couldn't attend, but met Montgomery for tea, and when Montgomery started to tell Dyson his conjecture, Dyson said, "Do you think it's this?" and proceeded to write Montgomery's conjecture on the board. He explained that this was what you would get if the eigenvalues of a large random unitary matrix could model the zeroes of zeta. Since then, a huge literature has been built up around understanding the connection and using it and facts about random unitary matrices in order to understand the zeta zeroes. There's been far too much activity to pretend to survey here, but some notable developments were Odlyzko's compu-

tational work, which shows that the connection between zeta zeroes and eigenvalues passes every statistical test anyone's thrown at them (see in particular the paper of Persi Diaconis and Marc Coram), Keating and Snaith's suggestion that the characteristic polynomial of a random unitary matrix can model zeta itself, which has led to a remarkable series of conjectures on the zeta zeroes, and Katz and Sarnak's discovery (and rigorous proof!) of the connection between the eigenvalues of random matrices from the compact classical groups and other L-functions.

Finally, we've talked already about some geometric properties of orthogonal and unitary matrices; they encode orthonormal bases of $\mathbb{R}^n$ and $\mathbb{C}^n$. As such, talking about random orthogonal and unitary matrices lets us talk about random bases and, maybe more importantly, random projections onto lower-dimensional subspaces. This leads to beautiful results about the geometry of high-dimensional Euclidean space, and also to important practical applications. In Lecture 4, we'll see how deep facts about Haar measure on the orthogonal group yield powerful randomized algorithms in high-dimensional data analysis.

Lecture 2

Some properties of Haar measure on the compact classical matrix groups

2.1 Some simple observations

There are a few useful and important properties of Haar measure on $\mathbb{O}(n)$, etc., that we can get easily from translation invariance and the orthogonality (unitarity? unitariness?) of the matrices themselves. The first was actually Exercise 1.8, which said that the distribution of a Haar random matrix is invariant under taking the transpose (and conjugate transpose). The next is an important symmetry of Haar measure which will come up constantly.

Lemma 2.1. *Let U be distributed according to Haar measure in G, where G is one of $\mathbb{O}(n)$, $\mathbb{U}(n)$, $\mathbb{SO}(n)$, and $\mathbb{SU}(n)$. Then all of the entries of U are identically distributed.*

Proof. Recall that permutations can be encoded by matrices: to a permutation $\sigma \in S_n$, associate the matrix M_σ with entries in $\{0, 1\}$, such that $m_{ij} = 1$ if and only if $\sigma(i) = j$. Such a permutation matrix M_σ is in G (check!). Moreover, multiplication on the left by M_σ permutes the rows by σ and multiplication on the right by M_σ permutes the columns by σ^{-1}. We can thus move any entry of the matrix into, say, the top-left corner by multiplication on the right and/or left by matrices in G. By the translation invariance of Haar measure, this means that all entries have the same distribution. □

Exercise 2.2. If U is Haar-distributed in $\mathbb{U}(n)$, the distributions of $\mathrm{Re}(U_{11})$ and $\mathrm{Im}(U_{11})$ are identical.

In addition to making the distribution of our random matrices reassuringly symmetric, the lemma makes some computations quite easy. For example, now that we know that the entries of U all have the same distributions, a natural thing to do is to try to calculate a few things like $\mathbb{E}u_{11}$ and $\mathrm{Var}(u_{11})$. We could use one of the contructions from the last lecture, but that would be overkill; putting the symmetries we have to work is much easier, as in the following example.

Example. Let U be Haar distributed in G, for G as above.

1. $\mathbb{E}[u_{11}] = 0$: note that Haar measure is invariant under multiplication on the left by

$$\begin{bmatrix} -1 & 0 & & 0 \\ 0 & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{bmatrix};$$

   doing so multiplies the top row (so in particular u_{11}) of U by -1, but doesn't change the distribution of the entries. So $u_{11} \stackrel{d}{=} -u_{11}$ ($\stackrel{d}{=}$ means "equals in distribution"), and thus $\mathbb{E}[u_{11}] = 0$.
2. $\mathbb{E}|u_{11}|^2 = \frac{1}{n}$: because $U \in G$, we know that $\sum_{j=1}^n |u_{1j}|^2 = 1$, and because all the entries have the same distribution, we can write

$$\mathbb{E}|u_{11}|^2 = \frac{1}{n}\sum_{j=1}^n \mathbb{E}|u_{1j}|^2 = \frac{1}{n}\mathbb{E}\left(\sum_{j=1}^n |u_{1j}|^2\right) = \frac{1}{n}.$$

Exercise 2.3. For $U = \left[u_{ij}\right]_{j=1}^n$, compute $\mathrm{Cov}\left(u_{ij}, u_{k\ell}\right)$ and $\mathrm{Cov}\left(u_{ij}^2, u_{k\ell}^2\right)$ for all i, j, k, ℓ.

Understanding the asymptotic distribution of the individual entries of Haar-distributed matrices is of course more involved than just calculating the first couple of moments, but follows from classical results. Recall that our geometric construction of Haar measure on $\mathbb{O}(n)$ involves filling the first column with a random point on the sphere (the same construction works for $\mathbb{U}(n)$, filling the first column with a uniform random point in the complex sphere $\{z \in \mathbb{C}^n : |z| = 1\}$.) That is, the distribution of u_{11} is *exactly* that of x_1, where $x = (x_1, \ldots, x_n)$ is a uniform random point of $\mathbb{S}^{n-1} \subseteq \mathbb{R}^n$. The asymptotic distribution of a single coordinate of a point on the sphere has been known for over a hundred years; the first rigorous proof is due to Borel in 1906, but it was recognized by Maxwell and others decades earlier. It is also often referred to as the "Poincaré limit", although apparently without clear reasons (Diaconis and Freedman's paper which quantifies this result has an extensive discussion of the history.)

Theorem 2.4 (Borel's lemma)**.** *Let $X = (X_1, \ldots, X_n)$ be a uniform random vector in $\mathbb{S}^{n-1} \subseteq \mathbb{R}^n$. Then*

$$\mathbb{P}\left[\sqrt{n}X_1 \leq t\right] \xrightarrow{n\to\infty} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^t e^{-\frac{x^2}{2}}dx;$$

that is, $\sqrt{n}X_1$ converges weakly to a Gaussian random variable, as $n \to \infty$.

There are various ways to prove Borel's lemma; one way is by the method of moments. The following proposition taken from [**?**] gives a general formula for integrating polynomials over spheres.

Proposition 2.5. *Let $P(x) = |x_1|^{\alpha_1}|x_2|^{\alpha_2}\cdots|x_n|^{\alpha_n}$. Then if X is uniformly distributed on $\sqrt{n}S^{n-1}$,*

$$\mathbb{E}\big[P(X)\big] = \frac{\Gamma(\beta_1)\cdots\Gamma(\beta_n)\Gamma(\frac{n}{2})n^{\left(\frac{1}{2}\sum\alpha_i\right)}}{\Gamma(\beta_1+\cdots+\beta_n)\pi^{n/2}},$$

where $\beta_i = \frac{1}{2}(\alpha_i+1)$ for $1\le i\le n$ and

$$\Gamma(t) = \int_0^\infty s^{t-1}e^{-s}ds = 2\int_0^\infty r^{2t-1}e^{-r^2}dr.$$

(The proof is essentially a reversal of the usual trick for computing the normalizing constant of the Gaussian distribution – it's not a bad exercise to work it out.)

Proof of Borel's lemma by moments. To prove the lemma, we need to show that if we consider the sequence of random variables Y_n distributed as the first coordinate of a uniform random point on $\sqrt{n}\mathbb{S}^{n-1}$, that for m fixed,

$$\lim_{n\to\infty}\mathbb{E}\big[Y_n^m\big] = \mathbb{E}\big[Z^m\big], \tag{2.1}$$

where Z is a standard Gaussian random variable. Recall that the moments of the standard Gaussian distribution are

$$\mathbb{E}\big[Z^m\big] = \begin{cases}(m-1)(m-3)(m-5)\dots(1), & m=2k;\\ 0, & m=2k+1.\end{cases} \tag{2.2}$$

The expression $(m-1)(m-3)\dots(1)$ is sometimes called "$(m-1)$ skip-factorial" and denoted $(m-1)!!$.

To prove (2.1), first note that it follows by symmetry that $\mathbb{E}[X_1^{2k+1}] = 0$ for all $k\ge 0$. Next, specializing Proposition 2.5 to $P(X) = X_1^{2k}$ gives that the even moments of X_1 are

$$\mathbb{E}\big[X_1^{2k}\big] = \frac{\Gamma\left(k+\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)^{n-1}\Gamma\left(\frac{n}{2}\right)n^k}{\Gamma\left(k+\frac{n}{2}\right)\pi^{\frac{n}{2}}}.$$

Using the functional equation $\Gamma(t+1) = t\Gamma(t)$ and the fact that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, this simplifies to

$$\mathbb{E}\big[X_1^{2k}\big] = \frac{(2k-1)(2k-3)\dots(1)n^k}{(n+2k-2)(n+2k-4)\dots(n)}. \tag{2.3}$$

Equation (2.1) follows immediately.

□

We've learned a lot in the last century, even about this rather classical problem. In particular, we can give a much more precise statement that quantifies the central limit theorem of Borel's lemma. In order to do this, we will first need to explore some notions of distance between measures.

2.2 Metrics on probability measures

What is generally meant by quantifying a theorem like Borel's lemma is to give a *rate of convergence* of Y_n to Z in some metric; that is, to give a bound in terms of n on the distance between Y_n and Z, for some notion of distance. The following are some of the more widely used metrics on probability measures on $\mathbb{R}^n$. The definitions can be extended to measures on other spaces, but for now we'll stick with $\mathbb{R}^n$.

1. Let μ and ν be probability measures on $\mathbb{R}^n$. The **total variation distance** between μ and ν is defined by

$$d_{TV}(\mu, \nu) := 2 \sup_{A \subseteq \mathbb{R}^n} |\mu(A) - \nu(A)|,$$

where the supremum is over Borel measurable sets. Equivalently, one can define

$$d_{TV}(\mu, \nu) := \sup_{f:\mathbb{R}^n \to \mathbb{R}} \left| \int f d\mu - \int f d\nu \right|,$$

where the supremum is over functions f which are continuous, such that $\|f\|_\infty \leq 1$. The total variation distance is a very strong metric on probability measures; in particular, you cannot approximate a continuous distribution by a discrete distribution in total variation.

Exercise 2.6.

(a) Prove that these two definitions are equivalent.
Hint: The Hahn decomposition of $\mathbb{R}^n$ corresponding to the signed measure $\mu - \nu$ is useful here.

(b) Prove that the total variation distance between a discrete distribution and a continuous distribution is always 2.

2. The **bounded Lipschitz distance** is defined by

$$d_{BL}(\mu, \nu) := \sup_{\|g\|_{BL} \leq 1} \left| \int g \, d\mu - \int g \, d\nu \right|,$$

where the bounded-Lipschitz norm $\|g\|_{BL}$ of $g : \mathbb{R}^n \to \mathbb{R}$ is defined by

$$\|g\|_{BL} := \max \left\{ \|g\|_\infty, \sup_{x \neq y} \frac{|g(x) - g(y)|}{\|x - y\|} \right\}$$

and $\|\cdot\|$ denotes the standard Euclidean norm on $\mathbb{R}^n$. The bounded-Lipschitz distance is a metric for the weak topology on probability measures (see, e.g., [**?**, Theorem 11.3.3]).

3. The L_p **Wasserstein distance** for $p \geq 1$ is defined by

$$W_p(\mu, \nu) := \inf_{\pi} \left[\int \|x - y\|^p \, d\pi(x, y) \right]^{\frac{1}{p}},$$

where the infimum is over couplings π of μ and ν; that is, probability measures π on $\mathbb{R}^{2n}$ such that $\pi(A \times \mathbb{R}^n) = \mu(A)$ and $\pi(\mathbb{R}^n \times B) = \nu(B)$. The L_p Wasserstein distance is a metric for the topology of weak convergence plus convergence of moments of order p or less. (See [**?**, Section 6] for a proof of this fact, and a lengthy discussion of the many fine mathematicians after whom this distance could reasonably be named.)

When $p = 1$, there is the following alternative formulation:

$$W_1(\mu, \nu) := \sup_{|f|_L \leq 1} \left| \int f \, d\mu - \int f \, d\nu \right|,$$

where $|f|_L$ denotes the Lipschitz constant of f. That this is the same thing as W_1 defined above is the *Kantorovich-Rubenstein* theorem.

As a slight extension of the notation defined above, we will also write things like $d_{TV}(X, Y)$, where X and Y are random vectors in $\mathbb{R}^n$, to mean the total variation distance between the distributions of X and Y.

2.3 More refined properties of the entries of Haar-distributed matrices

We saw in Section 2.1 that if $U = \left[u_{ij}\right]_{i,j=1}^n$ is a Haar-distributed random orthogonal matrix, then the asymptotic distribution of $\sqrt{n}u_{11}$ is the standard Gaussian distribution. Moreover, this followed from the classical result (Borel's lemma) that the first coordinate of a uniform random point on $\sqrt{n}\mathbb{S}^{n-1}$ converges weakly to Gaussian, as $n \to \infty$. Borel's lemma has been strengthened considerably, as follows.

Theorem 2.7 (Diaconis-Freedman)**.** *Let X be a uniform random point on $\sqrt{n}\mathbb{S}^{n-1}$, for $n \geq 5$. Then if Z is a standard Gaussian random variable,*

$$d_{TV}(X_1, Z) \leq \frac{4}{n-4}.$$

This theorem is fine as far as it goes (it is in fact sharp in the dependence on n), but it's very limited as a means of understanding Haar measure on $\mathbb{O}(n)$, since it's only about the distribution of individual entries and not about their joint distributions. You will see (or already have seen) in Exercise 2.3 that the covariances between the squares of the entries are quite small; much smaller than the variances of individual entries. It's natural to conjecture

then that you could approximate some of the entries of a large random orthogonal matrix by a collection of independent Gaussian random variables. Diaconis and Freedman in fact showed rather more about the coordinates of a random point on the sphere:

Theorem 2.8 (Diaconis-Freedman)**.** *Let X be a uniform random point on $\sqrt{n}\mathbb{S}^{n-1}$, for $n \geq 5$, and let $1 \leq k \leq n-4$. Then if Z is a standard Gaussian random vector in $\mathbb{R}^k$,*

$$d_{TV}\big((X_1, \ldots, X_k), Z\big) \leq \frac{2(k+3)}{n-k-3}.$$

This means that one can approximate k entries from the same row or column of U by independent Gaussian random variables, as long as $k = o(n)$. Persi Diaconis then raised the question: How many entries of U can be simultaneously approximated by independent normal random variables? A non-sharp answer was given by Diaconis, Eaton and Lauritson in [**?**]; the question was definitively answered (in two ways) by Tiefeng Jiang, as follows.

Theorem 2.9 (Jiang's Theorem 1)**.** *Let $\{U_n\}$ be a sequence of random orthogonal matrices with $U_n \in \mathbb{O}(n)$ for each n, and suppose that $p_n, q_n = o(\sqrt{n})$. Let $\mathcal{L}(\sqrt{n}U(p_n, q_n))$ denote the joint distribution of the $p_n q_n$ entries of the top-left $p_n \times q_n$ block of $\sqrt{n}U_n$, and let $\Phi(p_n, q_n)$ denote the distribution of a collection of $p_n q_n$ i.i.d. standard normal random variables. Then*

$$\lim_{n\to\infty} d_{TV}(\mathcal{L}(\sqrt{n}U(p_n, q_n)), \Phi(p_n, q_n)) = 0.$$

That is, a $p_n \times q_n$ principle submatrix can be approximated in total variation by a Gaussian random matrix, as long as $p_n, q_n \ll \sqrt{n}$. The theorem is sharp in the sense that if $p_n \sim x\sqrt{n}$ and $q_n \sim y\sqrt{n}$ for $x, y > 0$, then $d_{TV}(\mathcal{L}(\sqrt{n}U(p_n, q_n)), \Phi(p_n, q_n))$ does not tend to zero.

As we said above, total variation distance is a very strong metric on the space of probability measures. Jiang also proved the following theorem, which says that if you accept a much weaker notion of approximation, then you can approximate many more entries of U by i.i.d. Gaussians. Recall that a sequence of random variables $\{X_n\}$ tends to zero in probability (denoted $X_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$) if for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[|X_n| > \epsilon\right] = 0.$$

Theorem 2.10 (Jiang's Theorem 2)**.** *For each n, let $Y_n = \left[y_{ij}\right]_{i,j=1}^n$ be an $n \times n$ matrix of independent standard Gaussian random variables and let $\Gamma_n = \left[\gamma_{ij}\right]_{i,j=1}^n$ be the matrix obtained from Y_n by performing the Gram-Schmidt process; i.e., Γ_n is a random orthogonal matrix. Let*

$$\epsilon_n(m) = \max_{1\leq i\leq n, 1\leq j\leq m} \left|\sqrt{n}\gamma_{ij} - y_{ij}\right|.$$

Then

$$\epsilon_n(m_n) \xrightarrow[n\to\infty]{\mathbb{P}} 0$$

if and only if $m_n = o\left(\frac{n}{\log(n)}\right)$.

That is, in an "in probability" sense, $o\left(\frac{n^2}{\log(n)}\right)$ entries of U (so nearly all of them!) can be simultaneously approximated by independent Gaussians.

2.4 A first look at eigenvalues

Suppose U is a random orthogonal or unitary matrix. Then U has eigenvalues (U is normal by definition), all of which lie on the unit circle $\mathbb{S}^1 \subseteq \mathbb{C}$. Since U is random, its set of eigenvalues is a *random point process*; that is, it is a collection of n random points on $\mathbb{S}^1$. The eigenvalue process of a random orthogonal or unitary matrix has many remarkable properties, the first of which is that there is an explicit formula (due to H. Weyl) for its density. The situation is simplest for random unitary matrices.

Lemma 2.11 (Weyl density formula)**.** *The unordered eigenvalues of an $n \times n$ random unitary matrix have eigenvalue density*

$$\frac{1}{n!(2\pi)^n} \prod_{1 \leq j < k \leq n} |e^{i\theta_j} - e^{i\theta_k}|^2,$$

with respect to $d\theta_1 \cdots d\theta_n$ on $(2\pi)^n$.

That is, for any central *function $g : \mathbb{U}(n) \to \mathbb{R}$ (g is central if $g(U) = g(VUV^*)$ for any $U, V \in \mathbb{U}(n)$, or alternatively, if g depends on U only through its eigenvalues),*

$$\int_{\mathbb{U}(n)} g dHaar = \frac{1}{n!(2\pi)^n} \int_{[0,2\pi)^n} \tilde{g}(\theta_1, \ldots, \theta_n) \prod_{1 \leq j < k \leq n} |e^{i\theta_j} - e^{i\theta_k}|^2 d\theta_1 \cdots d\theta_n,$$

where $\tilde{g} : [0, 2\pi)^n \to \mathbb{R}$ is the expression of $g(U)$ as a function of the eigenvalues of U. Note that any $\tilde{g}$ arising in this way is therefore invariant under permutations of coordinates on $[0, 2\pi)^n$: $\tilde{g}(\theta_1, \ldots, \theta_n) = \tilde{g}(\theta_{\sigma(1)}, \ldots, \theta_{\sigma(n)})$ for any $\sigma \in S_n$.

More concretely, let $\{e^{i\phi_j}\}_{j=1}^n$ be the eigenvalues of a Haar-distributed random orthogonal matrix, with $0 \leq \phi_1 < \phi_2 < \cdots < \phi_n < 2\pi$. Let $\sigma \in S_n$ be a random permutation, independent of U. Then for any measurable $A \subseteq [0, 2\pi)^n$,

$$\mathbb{P}\left[(e^{i\phi_{\sigma(1)}}, \ldots, e^{i\phi_{\sigma(n)}}) \in A\right] = \frac{1}{n!(2\pi)^n} \int_A \cdots \int \prod_{j<k} |e^{i\theta_j} - e^{i\theta_k}|^2 d\theta_1 \ldots d\theta_n.$$

Equivalently, if A is a measureable subset of $\{0 \leq \theta_1 < \theta_2 < \cdots < \theta_n < 2\pi\} \subseteq [0, 2\pi)^n$, then

$$\mathbb{P}\left[(e^{i\phi_1}, \ldots, e^{i\phi_n}) \in A\right] = \frac{1}{(2\pi)^n} \int_A \cdots \int \prod_{j<k} |e^{i\theta_j} - e^{i\theta_k}|^2 d\theta_1 \ldots d\theta_n.$$

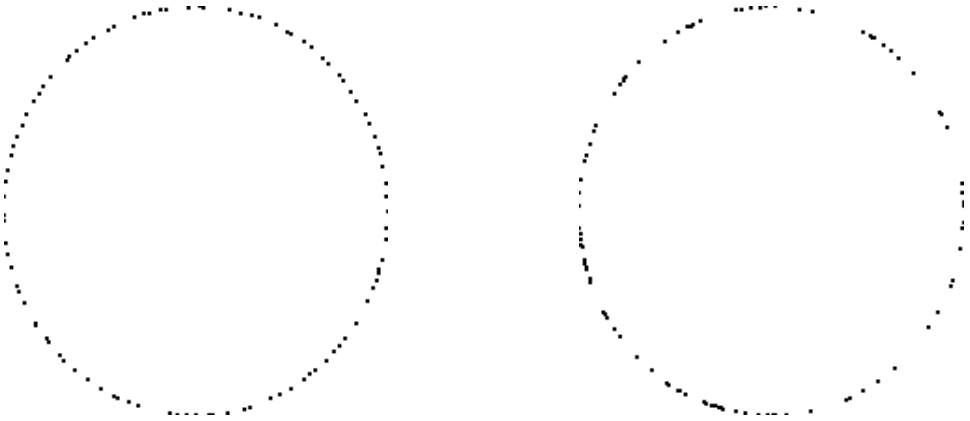


Figure 2.1: On the left are the eigenvalues of a 100×100 random unitary matrix; on the right are 100 i.i.d. uniform random points. *Figures courtesy of E. Rains.*

The factor $\prod_{1\leq j<k\leq n} |e^{i\theta_j} - e^{i\theta_k}|^2$ is the norm-squared of a Vandermonde determinant, which means one can also write it as

$$\prod_{1\leq j<k\leq n} |e^{i\theta_j} - e^{i\theta_k}|^2 = \left|\det\left[e^{i\theta_j(k-1)}\right]_{j,k=1}^{n}\right|^2 = \sum_{\sigma,\tau\in S_n} \operatorname{sgn}(\sigma\tau) \prod_{1\leq k\leq n} e^{i\theta_k(\sigma(k)-\tau(k))}. \tag{2.4}$$

This last expression can be quite useful in computations.

Either expression for the eigenvalue density is pretty hard to look at, but one thing to notice right away is that for any given pair (j, k), $|e^{i\theta_k} - e^{i\theta_j}|^2$ is zero if $\theta_j = \theta_k$ (and small if they are close), but $|e^{i\theta_k} - e^{i\theta_j}|^2$ is 4 if $\theta_j = \theta_k + \pi$ (and in that neighborhood if θ_j and θ_k are roughly antipodal). This produces the effect known as "eigenvalue repulsion": the eigenvalues *really* want to spread out. You can see this pretty dramatically in pictures, even for matrices which aren't really that large.

In the picture on the right in Figure 2.4, where 100 points were dropped uniformly and independently, there are several large clumps of points close together, and some biggish gaps. In the picture on the right, this is much less true: the eigenvalues are spread pretty evenly, and there are no big clumps.

Things are similar for the other matrix groups, just a little fussier. Each matrix in $\mathbb{SO}(2N+1)$ has 1 as an eigenvalue, each matrix in $\mathbb{SO}^-(2N+1)$ has -1 as an eigenvalue, and each matrix in $\mathbb{SO}^-(2N+2)$ has both -1 and 1 as eigenvalues; we refer to all of these as trivial eigenvalues. The remaining eigenvalues of matrices in $\mathbb{SO}(N)$ or $\mathbb{Sp}(2N)$ occur in complex conjugate pairs. For this reason, when discussing $\mathbb{SO}(N)$, $\mathbb{SO}^-(N)$, or $\mathbb{Sp}(2N)$, the eigenvalue angles corresponding to the eigenvalues in the open upper half-circle are the nontrivial ones and we generally restrict our attention there. For $\mathbb{U}(N)$, all the eigenvalue angles are considered nontrivial; there are no automatic symmetries in the eigenvalue process in this case.

In the case of the orthogonal and symplectic groups, one can give a similar formula for the density of the non-trivial eigenangles as in the unitary case, although it is not as easy to work with because it doesn't take the form of a norm squared. The densities are as follows.

Theorem 2.12. *Let U be a Haar-distributed random matrix in S, where S is one of $\mathbb{SO}(2n+1)$, $\mathbb{SO}(2n)$, $\mathbb{SO}^-(2n+1)$, $\mathbb{SO}^-(2n+2)$, $\mathbb{Sp}(2n)$. Then a function g of U which is invariant under conjugation of U by a fixed orthogonal (in all but the last case) or symplectic (in the last case) matrix is associated as above with a function $\tilde{g} : [0,\pi)^n \to \mathbb{R}$ (of the non-trivial eigenangles) which is invariant under permutations of coordinates, and*

$$\int_G g dHaar = \int_{[0,\pi)^n} \tilde{g} d\mu_G,$$

where the measures μ_G on $[0,\pi)^n$ have densities with respect to $d\theta_1 \cdots d\theta_n$ as follows.

G	μ_G
$\mathbb{SO}(2n)$	$\frac{2}{n!(2\pi)^n} \prod_{1\le j<k\le n} \left(2\cos(\theta_k) - 2\cos(\theta_j)\right)^2$
$\mathbb{SO}(2n+1), \mathbb{SO}^-(2n+1)$	$\frac{2^n}{n!\pi^n} \prod_{1\le j\le n} \sin^2\left(\frac{\theta_j}{2}\right) \prod_{1\le j<k\le n} \left(2\cos(\theta_k) - 2\cos(\theta_j)\right)^2$
$\mathbb{Sp}(2n), \mathbb{SO}^-(2N+2)$	$\frac{2^n}{n!\pi^n} \prod_{1\le j\le n} \sin^2(\theta_j) \prod_{1\le j<k\le n} \left(2\cos(\theta_k) - 2\cos(\theta_j)\right)^2$

One of the technical tools that is commonly used to study the ensemble of eigenvalues of a random matrix is the *empirical spectral measure*. Given a random matrix U with eigenvalues $\lambda_1, \ldots, \lambda_n$, the **empirical spectral measure** of U is the random measure

$$\mu_U := \frac{1}{n} \sum_{j=1}^{n} \delta_{\lambda_j}.$$

The empirical spectral measure is a handy way to encode the ensemble of eigenvalues. In particular, it lets us formalize the idea that the eigenvalues are evenly spread out on the circle, and indeed more so than i.i.d. points are, by comparing the empirical spectral measure to the uniform measure on the circle. In dealing with convergence of random measures, one notion that comes up a lot is that of convergence "weakly in probability", although it is seldom actually defined. This is an unfortunate practice, with which we will break. However, we will specialize to the situation that the random measures are defined on $\mathbb{S}^1$, since that is where our empirical spectral measures live.

Definition. A sequence of random probability measures μ_n on $\mathbb{S}^1$ converge weakly in probability to a measure μ on $\mathbb{S}^1$ (written $\mu_n \stackrel{\mathbb{P}}{\Longrightarrow} \mu$) if for each continuous $f : \mathbb{S}^1 \to \mathbb{R}$,

$$\int f d\mu_n \xrightarrow[n\to\infty]{\mathbb{P}} \int f d\mu.$$

There are many equivalent viewpoints:

Lemma 2.13. *For $j \in \mathbb{Z}$ and μ a probability measure on $[0, 2\pi)$, let $\hat{\mu}(j) = \int_0^{2\pi} e^{ij\theta} d\mu(\theta)$ denote the Fourier transform of μ at j. The following are equivalent:*

1. *$\mu_n \overset{\mathbb{P}}{\Longrightarrow} \mu$;*
2. *for each $j \in \mathbb{Z}$, $\hat{\mu}_n(j) \xrightarrow[n\to\infty]{\mathbb{P}} \hat{\mu}(j)$;*
3. *for every subsequence n' in $\mathbb{N}$ there is a further subsequence n'' such that with probability one, $\mu_{n''} \Longrightarrow \mu$ as $n \to \infty$.*

The first result showing that the eigenvalues of a Haar-distributed matrix evenly fill out the circle was proved by Diaconis and Shahshahani.

Theorem 2.14 (Diaconis–Shahshahani)**.** *Let $\{G_n\}$ be one of the sequences $\{\mathbb{O}(n)\}$, $\{\mathbb{U}(n)\}$, or $\{\mathbb{Sp}(2n)\}$ of groups, and let μ_n be the empirical spectral measure of U_n, where U_n is Haar-distributed in G_n. Then as $n \to \infty$, μ_n converges, weakly in probability, to the uniform measure ν on $\mathbb{S}^1$.*

In fact, it is possible to show that the measures μ_n converge quite quickly to the uniform measure on the circle, and indeed more quickly than the empirical measure of n i.i.d. uniform points on the circle.

Theorem 2.15 (E. Meckes-M. Meckes)**.** *Suppose that for each n, U_n is Haar distributed in G_n, where G_n is one of $\mathbb{O}(n)$, $\mathbb{SO}(n)$, $\mathbb{SO}^-(n)$, $\mathbb{U}(n)$, $\mathbb{SU}(n)$, or $\mathbb{Sp}(2n)$. Let ν denote the uniform measure on $\mathbb{S}^1$. There is an absolute constant C such that with probability 1, for all sufficiently large n, and all $1 \le p \le 2$,*

$$W_p(\mu_n, \nu) \le C \frac{\sqrt{\log(n)}}{n}.$$

By way of comparison, it is known (cf. [**?**]) that if μ_n is the empirical measure of n i.i.d. uniform points on $\mathbb{S}^1$, then $W_p(\mu_n, \nu)$ is typically of order $\frac{1}{\sqrt{n}}$.

The above result gives one explicit demonstration that there is important structure to the eigenvalue processes of these random matrices; in at least some ways, they are quite different from collections of independent random points. There are indeed many beautiful patterns to be found; a particularly striking example is the following.

Theorem 2.16 (Rains)**.** *Let $m \in \mathbb{N}$ be fixed and let $\widetilde{m} := \min\{m, N\}$. If $\sim$ denotes equality of eigenvalue distributions, then*

$$\mathbb{U}(N)^m \sim \bigoplus_{0 \le j < \widetilde{m}} \mathbb{U}\left(\left\lceil \frac{N-j}{\widetilde{m}} \right\rceil\right)$$

That is, if U is a uniform $N \times N$ unitary matrix, the eigenvalues of U^m are distributed as those of $\widetilde{m}$ independent uniform unitary matrices of sizes

$$\left\lfloor \frac{N}{\widetilde{m}} \right\rfloor := \max\left\{k \in \mathbb{N} \mid k \le \frac{N}{\widetilde{m}}\right\} \qquad \textit{and} \qquad \left\lceil \frac{N}{\widetilde{m}} \right\rceil := \min\left\{k \in \mathbb{N} \mid k \ge \frac{N}{\widetilde{m}}\right\},$$

such that the sum of the sizes of the matrices is N. In particular, if $m \geq N$, the eigenvalues of U^m are distributed exactly as N i.i.d. uniform points on $\mathbb{S}^1$.

We conclude this lecture by mentioning a special algebraic structure of the eigenvalue processes of the compact classical groups. These processes are what's known as *determinantal point processes*. The definition is a little complicated, but it turns out that processes that satisfy it have very special (very useful) properties, as we'll see.

Firstly, a **point process** $\mathcal{X}$ in a locally compact Polish space Λ is a random discrete subset of Λ. Abusing notation, we denote by $\mathcal{X}(D)$ the number of points of $\mathcal{X}$ in D. A point process may or may not have *k-point correlation functions*, defined as follows.

Definition. For a point process $\mathcal{X}$ in Λ, suppose there exist functions $\rho_k : \Lambda^k \to [0, \infty)$ such that, for pairwise disjoint subsets $D_1, \ldots, D_k \subseteq \Lambda$,

$$\mathbb{E}\left[\prod_{j=1}^{k} \mathcal{X}(D_i)\right] = \int_{D_1} \cdots \int_{D_k} \rho_k(x_1, \ldots, x_k) dx_1 \cdots dx_k.$$

Then the ρ_k are called the **k-point correlation functions** (or **joint intensities**) of $\mathcal{X}$.

A determinantal point process is a point process whose k-point correlation functions have a special form:

Definition. Let $K : \Lambda \times \Lambda \to [0, 1]$. A point process $\mathcal{X}$ is a **determinantal point process with kernel** K if for all $k \in \mathbb{N}$,

$$\rho_k(x_1, \ldots, x_k) = \det\left[K(x_i, x_j)\right]_{i,j=1}^{k}.$$

Proposition 2.17. *The nontrivial eigenvalue angles of uniformly distributed random matrices in any of $\mathbb{SO}(N)$, $\mathbb{SO}^-(N)$, $\mathbb{U}(N)$, $\mathbb{Sp}(2N)$ are a determinantal point process, with respect to uniform measure on Λ, with kernels as follows.*

	$K_N(x, y)$	Λ
$\mathbb{SO}(2N)$	$1 + \sum_{j=1}^{N-1} 2\cos(jx)\cos(jy)$	$[0, \pi)$
$\mathbb{SO}(2N+1), \mathbb{SO}^-(2N+1)$	$\sum_{j=0}^{N-1} 2\sin\left(\frac{(2j+1)x}{2}\right)\sin\left(\frac{(2j+1)y}{2}\right)$	$[0, \pi)$
$\mathbb{U}(N)$	$\sum_{j=0}^{N-1} e^{ij(x-y)}$	$[0, 2\pi)$
$\mathbb{Sp}(2N), \mathbb{SO}^-(2N+2)$	$\sum_{j=1}^{N} 2\sin(jx)\sin(jy)$	$[0, \pi)$

For some purposes, the following alternatives can be more convenient. In all but the unitary case, they are the same functions; for the unitary case, the kernels are different but define the same point processes.

First define

$$S_N(x) := \begin{cases} \sin\left(\frac{Nx}{2}\right) / \sin\left(\frac{x}{2}\right) & \text{if } x \neq 0, \\ N & \text{if } x = 0. \end{cases}$$

Proposition 2.18. *The nontrivial eigenvalue angles of uniformly distributed random matrices in any of* $\mathbb{SO}(N)$, $\mathbb{SO}^-(N)$, $\mathbb{U}(N)$, $\mathbb{Sp}(2N)$ *are a determinantal point process, with respect to uniform measure on* Λ*, with kernels as follows.*

	$L_N(x,y)$	Λ
$\mathbb{SO}(2N)$	$\frac{1}{2}\Big(S_{2N-1}(x-y) + S_{2N-1}(x+y)\Big)$	$[0,\pi)$
$\mathbb{SO}(2N+1), \mathbb{SO}^-(2N+1)$	$\frac{1}{2}\Big(S_{2N}(x-y) - S_{2N}(x+y)\Big)$	$[0,\pi)$
$\mathbb{U}(N)$	$S_N(x-y)$	$[0,2\pi)$
$\mathbb{Sp}(2N), \mathbb{SO}^-(2N+2)$	$\frac{1}{2}\Big(S_{2N+1}(x-y) - S_{2N+1}(x+y)\Big)$	$[0,\pi)$

One thing that is convenient about determinantal point processes is that there are easy-to-use (in principle, at least) formulas for computing things like means and variances of the number of points in a given set, such as those given in the following lemma.

Lemma 2.19. *Let* $K : I \times I \to \mathbb{R}$ *be a continuous kernel on an interval* I *such that the corresponding operator*

$$\mathcal{K}(f)(x) := \int_I K(x,y) f(y) d\mu(y)$$

on $L^2(\mu)$*, where* μ *is the uniform measure on* I*, is an orthogonal projection. For a subinterval* $D \subseteq I$*, denote by* $\mathcal{N}_D$ *the number of particles of the determinantal point process with kernel* K *which lie in* D*. Then*

$$\mathbb{E}\mathcal{N}_D = \int_D K(x,x)\, d\mu(x)$$

and

$$\operatorname{Var}\mathcal{N}_D = \int_D \int_{I\setminus D} K(x,y)^2\, d\mu(x)\, d\mu(y).$$

Lecture 3

Concentration of Haar measure

3.1 The concentration of measure phenomenon

The phenomenon of concentration of measure has poked its head out of the water in many places in the history of mathematics, but was first explicitly described and used by Vitali Milman in the 1970's in his probabilistic proof of Dvoretzky's theorem. Milman strongly advocated that the phenomenon was both fundamental and useful, and its study and application has become a large and influential field since then. The basic idea is that functions with small local fluctuations are often essentially constant, where "essentially" means that we consider a function on a probability space (i.e., a random variable) and it is close to a particular value with probability close to 1.

An example of a concentration phenomenon in classical probability theory is the following.

Theorem 3.1 (Bernstein's inequality)**.** *Let $\{X_j\}_{j=1}^n$ be independent random variables such that, for each i, $|X_j| \leq 1$ almost surely. Let $\sigma^2 = \mathrm{Var}\left(\sum_{j=1}^n X_j\right)$. Then for all $t > 0$,*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{j=1}^n X_j - \mathbb{E}\left(\frac{1}{n}\sum_{j=1}^n X_j\right)\right| > t\right] \leq C\exp\left(-\min\left\{\frac{n^2t^2}{2\sigma^2}, \frac{nt}{2}\right\}\right).$$

That is, the average of independent bounded random variables is essentially constant, in that it is very likely to be close to its mean. We can reasonably think of the average of n random variables as a statistic with small local fluctuations, since if we just change the value of one (or a few) of the random variables, the average can only change on the order $\frac{1}{n}$.

In a more geometric context, we have the following similar statement about Lipschitz functions of a random point on the sphere.

Theorem 3.2 (Lévy's lemma)**.** *Let $f : \mathbb{S}^{n-1} \to \mathbb{R}$ be Lipschitz with Lipschitz constant L, and let X be a uniform random vector in $\mathbb{S}^{n-1}$. Let M be the median of f; that is,*

$\mathbb{P}[f(X) \geq M] \geq \frac{1}{2}$ *and* $\mathbb{P}[f(X) \leq M] \geq \frac{1}{2}$. *Then*

$$\mathbb{P}\left[\left|f(X) - M\right| \geq Lt\right] \leq 2e^{-(n-2)t^2}.$$

Again, this says that if the local fluctions of a function on the sphere are controlled (the function is Lipschitz), then the function is essentially constant.

We often prefer to state concentration results about the mean rather than the median, as follows.

Corollary 3.3. *Let $f : \mathbb{S}^{n-1} \to \mathbb{R}$ be Lipschitz with Lipschitz constant L, and let X be a uniform random vector in $\mathbb{S}^{n-1}$. Then for M_f denoting the median of f with respect to uniform measure on $\mathbb{S}^{n-1}$, $|\mathbb{E}f(X) - M_f| \leq L\sqrt{\frac{\pi}{n-2}}$ and*

$$\mathbb{P}[|f(X) - \mathbb{E}f(X)| \geq Lt] \leq e^{\pi - \frac{nt^2}{4}}.$$

That is, a Lipschitz function on the sphere is essentially constant, and we can take that constant value to be either the median or the mean of the function.

Proof. First note that Lévy's lemma and Fubini's theorem imply that

$$\begin{aligned}\left|\mathbb{E}f(X) - M_f\right| &\leq \mathbb{E}\left|f(X) - M_f\right| \\ &= \int_0^\infty \mathbb{P}\Big[\left|f(X) - M_f\right| > t\Big]dt \leq \int_0^\infty 2e^{-\frac{(n-2)t^2}{L^2}}dt = L\sqrt{\frac{\pi}{n-2}}.\end{aligned}$$

If $t > 2\sqrt{\frac{\pi}{n-2}}$, then

$$\begin{aligned}\mathbb{P}\left[|f(X) - \mathbb{E}f(X)| > Lt\right] &\leq \mathbb{P}\left[|f(X) - M_f| > Lt - |M_f - \mathbb{E}f(X)|\right] \\ &\leq \mathbb{P}\left[|f(X) - M_f| > L\left(t - \sqrt{\frac{\pi}{n-2}}\right)\right] \\ &\leq 2e^{-\frac{(n-2)t^2}{4}}.\end{aligned}$$

On the other hand, if $t \leq 2\sqrt{\frac{\pi}{n-2}}$, then

$$e^{\pi - \frac{(n-2)t^2}{4}} \geq 1,$$

so the statement holds trivially. □

3.2 Log-Sobolev inequalities and concentration

Knowing that a metric probability space possesses a concentration of measure property along the lines of Lévy's lemma opens many doors; however, it is not *a priori* clear how to show that such a property holds or to determine what the optimal (or even good) constants are. In this section we discuss one approach to obtaining measure concentration, which is in particular one way to prove Lévy's lemma.

We begin with the following general definitions for a metric space (X, d) equipped with a Borel probability measure $\mathbb{P}$.

Definition.

1. The **entropy** of a measurable function $f : X \to [0, \infty)$ with respect to $\mathbb{P}$ is
$$\mathrm{Ent}(f) := \mathbb{E}\big[f \log(f)\big] - (\mathbb{E}f) \log(\mathbb{E}f).$$
2. For a locally Lipschitz function $g : X \to \mathbb{R}$,
$$|\nabla g|(x) := \limsup_{y \to x} \frac{|g(y) - g(x)|}{d(y, x)}.$$

Exercise 3.4. Show that $\mathrm{Ent}(f) \geq 0$ and that for $c > 0$, $\mathrm{Ent}(cf) = c\,\mathrm{Ent}(f)$.

Definition. We say that $(X, d, \mathbb{P})$ satisfies a **logarithmic Sobolev inequality** (or **log-Sobolev inequality** or **LSI**) with constant $C > 0$ if, for every locally Lipschitz $f : X \to \mathbb{R}$,
$$\mathrm{Ent}(f^2) \leq 2C\mathbb{E}\left(|\nabla f|^2\right). \tag{3.1}$$

The reason for our interest in log-Sobolev inequalities is that they imply measure concentration for Lipschitz functions, via the "Herbst argument". The argument was outlined by Herbst in a letter to Len Gross, who was studying something called "hypercontractivity". The argument made it into folklore, and then books (e.g., [**?**]) without most of the people involved ever having seen the letter (Gross kept it, though, and if you ask nicely he'll let you see it).

Specifically, the following result holds.

Theorem 3.5. *Suppose that $(X, d, \mathbb{P})$ satisfies a log-Sobolev inequality with constant $C > 0$. Then for every 1-Lipschitz function $F : X \to \mathbb{R}$, $\mathbb{E}|F| < \infty$, and for every $r \geq 0$,*
$$\mathbb{P}\Big[\big|F - \mathbb{E}_\mu F\big| \geq r\Big] \leq 2e^{-r^2/2C}.$$

Proof (the Herbst argument).

We begin with the standard observation that for any $\lambda > 0$,
$$\mathbb{P}\left[F \geq \mathbb{E}F + r\right] = \mathbb{P}\left[e^{\lambda F - \mathbb{E}\lambda F} \geq e^{\lambda r}\right] \leq e^{-\lambda r}\mathbb{E}e^{\lambda F - \mathbb{E}\lambda F}, \tag{3.2}$$
assuming that $\mathbb{E}e^{\lambda F} < \infty$. For now, assume that F is bounded as well as Lipschitz, so that the finiteness is assured. Also, observe that we can always replace F with $F - \mathbb{E}F$, so we may assume that $\mathbb{E}F = 0$.

That is, given a bounded, 1-Lipschitz function $F : X \to \mathbb{R}$ with $\mathbb{E}F = 0$, we need to estimate $\mathbb{E}e^{\lambda F}$ under the assumption of an LSI with constant C; a natural thing to do is to apply the LSI to the function f with
$$f^2 := e^{\lambda F}.$$
For notational convenience, let $H(\lambda) := \mathbb{E}e^{\lambda F}$. Then
$$\mathrm{Ent}(f^2) = \mathbb{E}\left[\lambda F e^{\lambda F}\right] - H(\lambda)\log H(\lambda),$$

whereas

$$|\nabla f(x)| \leq e^{\frac{\lambda F(x)}{2}} \left(\frac{\lambda}{2}\right) |\nabla F(x)|,$$

because the exponential function is smooth and F is 1-Lipschitz, and so

$$\mathbb{E}|\nabla f|^2 \leq \frac{\lambda^2}{4}\mathbb{E}\left[|\nabla F|^2 e^{\lambda F}\right] \leq \frac{\lambda^2}{4}\mathbb{E}\left[e^{\lambda F}\right] = \frac{\lambda^2}{4}H(\lambda)$$

(since $|\nabla F| \leq 1$). Applying the LSI with constant C to this f thus yields

$$\mathbb{E}\left[\lambda F e^{\lambda F}\right] - H(\lambda)\log H(\lambda) = \lambda H'(\lambda) - H(\lambda)\log H(\lambda) \leq \frac{C\lambda^2}{2}H(\lambda),$$

or rearranging,

$$\frac{H'(\lambda)}{\lambda H(\lambda)} - \frac{\log H(\lambda)}{\lambda^2} \leq \frac{C}{2}.$$

Indeed, if we define $K(\lambda) := \frac{\log H(\lambda)}{\lambda}$, then the right-hand side is just $K'(\lambda)$, and so we have the simple differential inequality

$$K'(\lambda) \leq \frac{C}{2}.$$

Now, $H(0) = 1$, so

$$\lim_{\lambda\to 0} K(\lambda) = \lim_{\lambda\to 0} \frac{H'(\lambda)}{H(\lambda)} = \lim_{\lambda\to 0} \frac{\mathbb{E}\left[Fe^{\lambda F}\right]}{\mathbb{E}\left[e^{\lambda F}\right]} = \mathbb{E}F = 0,$$

and thus

$$K(\lambda) = \int_0^\lambda K'(s)ds \leq \int_0^\lambda \frac{C}{2}ds = \frac{C\lambda}{2}.$$

In other words,

$$H(\lambda) = \mathbb{E}\left[e^{\lambda F}\right] \leq e^{\frac{C\lambda^2}{2}}.$$

It follows from (3.2) that for $F : X \to \mathbb{R}$ which is 1-Lipschitz and bounded,

$$\mathbb{P}\left[F \geq \mathbb{E}F + r\right] \leq e^{-\lambda r + \frac{C\lambda^2}{2}}.$$

Now choose $\lambda = \frac{r}{C}$ and the statement of the result follows under the assumption that F is bounded.

In the general case, let $\epsilon > 0$ and define the truncation F_ϵ by

$$F_\epsilon(x) := \begin{cases} -\frac{1}{\epsilon}, & F(x) \leq -\frac{1}{\epsilon}; \\ F(x), & -\frac{1}{\epsilon} \leq F(x) \leq \frac{1}{\epsilon}; \\ \frac{1}{\epsilon}, & F(x) \geq \frac{1}{\epsilon}. \end{cases}$$

Then F_ϵ is 1-Lipschitz and bounded so that by the argument above,

$$\mathbb{E}\left[e^{\lambda F_\epsilon}\right] \leq e^{\lambda \mathbb{E}F_\epsilon + \frac{C\lambda^2}{2}}.$$

The truncation F_ϵ approaches F pointwise as $\epsilon \to 0$, so by Fatou's lemma,

$$\mathbb{E}\left[e^{\lambda F}\right] \leq e^{\liminf_{\epsilon\to 0} \lambda \mathbb{E}F_\epsilon} e^{\frac{C\lambda^2}{2}}.$$

It remains to show that $\mathbb{E}F_\epsilon \xrightarrow{\epsilon\to 0} \mathbb{E}F$; we can then complete the proof in the unbounded case exactly as before.

Now, we've already proved the concentration inequality

$$\mathbb{P}\left[|F_\epsilon - \mathbb{E}F_\epsilon| > t\right] \leq 2e^{-\frac{t^2}{2C}} \tag{3.3}$$

and F_ϵ converges pointwise (hence also in probability) to F, which has some finite value at each point in X, so there is a constant K such that

$$\mathbb{P}\left[|F| \leq K\right] \geq \frac{3}{4},$$

and the convergence of F_ϵ in probability to F means that there is some K' such that for $\epsilon < \epsilon_o$,

$$\mathbb{P}\left[|F_\epsilon - F| > K'\right] < \frac{1}{4}.$$

It follows that for $\epsilon < \epsilon_o$,

$$\mathbb{E}|F_\epsilon| < K + K'.$$

It also follows from (3.3) and Fubini's theorem that

$$\mathbb{E}\left|F_\epsilon - \mathbb{E}F_\epsilon\right|^2 = \int_0^\infty t\mathbb{P}\left[|F_\epsilon - \mathbb{E}F_\epsilon| > t\right] dt \leq \int_0^\infty 2te^{-\frac{t^2}{2C}} dt = 2C,$$

so that in fact $\mathbb{E}F_\epsilon^2 \leq 2C + K + K'$. Using Fatou's lemma again gives that

$$\mathbb{E}F^2 \leq \liminf_{\epsilon\to 0} \mathbb{E}F_\epsilon^2 \leq 2C + K + K'.$$

We can then use convergence in probability again

$$|\mathbb{E}F - \mathbb{E}F_\epsilon| =\leq \delta + \mathbb{E}|F_\epsilon - F|\mathbb{1}_{|F_\epsilon - F|>\delta} \leq \delta + \sqrt{\mathbb{E}|F_\epsilon - F|^2 \mathbb{P}\left[|F_\epsilon - F| > \delta\right]} \xrightarrow{\epsilon\to 0} 0.$$

□

One of the reasons that the approach to concentration via log-Sobolev inequalities is so nice is that log-Sobolev inequalities *tensorize*; that is, if one has the same LSI on each of some finite collection of spaces, one can get *the same* LSI again on the product space, independent of the number of factors. Specifically, we have the following.

Theorem 3.6 (see [**?**])**.** *Suppose that each of the metric probability spaces (X_i, d_i, μ_i) $(1 \leq i \leq n)$ satisifes a log-Sobolev inequality: for each i there is a $C_i > 0$ such that for every locally Lipschitz function $f : X_i \to \mathbb{R}$,*

$$\mathrm{Ent}_{\mu_i}(f^2) \leq 2C_i \int |\nabla_{X_i} f|^2 d\mu_i.$$

Let $X = X_1 \times \cdots \times X_n$, and equip X with the L_2-sum metric

$$d\big((x_1, \ldots, x_n), (y_1, \ldots, y_n)\big) := \sqrt{\sum_{i=1}^n d_i^2(x_i, y_i)}$$

and the product probability measure $\mu := \mu_1 \otimes \cdots \otimes \mu_n$. Then (X, d, μ) satisfies a log-Sobolev inequality with constant $C := \max_{1 \leq i \leq n} C_i$.

Note that with respect to the L_2-sum metric, if $f : X \to \mathbb{R}$, then

$$|\nabla f|^2 = \sum_{i=1}^n |\nabla_{X_i} f|^2,$$

where

$$|\nabla_{X_i} f(x_1, \ldots, x_n)| = \limsup_{y_i \to x_i} \frac{|f(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_n)|}{d_i(y_i, x_i)}.$$

A crucial point to notice above is that the constant C doesn't get worse with the number of factors; that is, the lemma gives *dimension-free* tensorization.

The theorem follows immediately from the following property of entropy.

Proposition 3.7. *Let $X = X_1 \times \cdots \times X_n$ and $\mu = \mu_1 \otimes \cdots \otimes \mu_n$ as above, and suppose that $f : X \to [0, \infty)$. For $\{x_1, \ldots, x_n\} \setminus \{x_i\}$ fixed, write*

$$f_i(x_i) = f(x_1, \ldots, x_n),$$

thought of as a function of x_i. Then

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

Proof. The proof is a good chance to see a dual formulation of the definition of entropy. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we defined entropy for a function $f : \Omega \to \mathbb{R}$ by

$$\mathrm{Ent}_{\mathbb{P}}(f) := \int f \log(f) d\mathbb{P} - \left(\int f d\mathbb{P} \right) \log \left(\int f d\mathbb{P} \right).$$

It turns out to be equivalent to define

$$\mathrm{Ent}_{\mathbb{P}}(f) := \sup\left\{ \int f g d\mathbb{P} \middle| \int e^g d\mathbb{P} \leq 1 \right\},$$

which can be seen as follows.

First, for simplicity we may assume that $\int f d\mathbb{P} = 1$, since both expressions we've given for the entropy are homogeneous of degree 1. Then our earlier expression becomes

$$\mathrm{Ent}_{\mathbb{P}}(f) = \int f \log(f) d\mathbb{P}.$$

Now, if $g := \log(f)$, then $\int e^g = \int f = 1$, and so we have that

$$\int f \log(f) d\mathbb{P} = \int f g d\mathbb{P} \leq \sup\left\{ \int f g d\mathbb{P} \middle| \int e^g d\mathbb{P} \leq 1 \right\}.$$

On the other hand, Young's inequality says that for $u \geq 0$ and $v \in \mathbb{R}$,

$$uv \leq u\log(u) - u + e^v;$$

applying this to $u = f$ and $v = g$ and integrating shows that

$$\sup\left\{ \int f g d\mathbb{P} \middle| \int e^g d\mathbb{P} \leq 1 \right\} \leq \int f \log(f) d\mathbb{P}.$$

With this alternative definition of entropy, given g such that $\int e^g d\mu \leq 1$, for each i define

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

(note that g^i only actually depends on $x_i, \ldots, x_n$). Then

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

and by construction,

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

Applying these two estimates together with Fubini's theorem yields

$$\int f g d\mu \leq \sum_{i=1}^n \int f g^i d\mu = \sum_{i=1}^n \int \left(\int f_i (g^i)_i d\mu_i \right) d\mu \leq \sum_{i=1}^n \int \mathrm{Ent}_{\mu_i}(f_i) d\mu.$$

□

The optimal (i.e., with smallest constants) log-Sobolev inequalities on most of the compact classical matrix groups were proved using the *Bakry-Émery curvature criterion*. To state it, we need to delve a little bit into the world of Riemannian geometry and define some quantities on Riemannian manifolds, most notably, the Ricci curvature.

Recall that a Riemannian manifold (M, g) is a smooth manifold (every point has an open neighborhood which is diffeomorphic to an open subset of Euclidean space) together with a **Riemannian metric** g. The metric g is a family of inner products: at each point $p \in M$, $g_p : T_pM \times T_pM \to \mathbb{R}$ defines an inner product on the tangent space T_pM to M at p. Our manifolds are all embedded in Euclidean space already, and the metrics are just those inherited from the ambient Euclidean space (this came up briefly in our discussion of the Riemannian construction of Haar measure in Lecture 1).

A **vector field** X on M is a smooth (infinitely differentiable) map $X : M \to TM$ such that for each $p \in M$, $X(p) \in T_pM$. To have a smooth Riemannian manifold, we require the metric g to be smooth, in the sense that for any two smooth vector fields X and Y on M, the map

$$p \longmapsto g_p(X(p), Y(p))$$

is a smooth real-valued function on M.

In Riemannian geometry, we think of vector fields as differential operators (think of them as directional derivatives): given a smooth function $f : M \to \mathbb{R}$ and a vector field X on M, we define the function $X(f)$ by the requirement that for any curve $\gamma : [0, T] \to M$ with $\gamma(0) = p$ and $\gamma'(0) = X(p)$ (here, $\gamma'(0)$ denotes the tangent vector to the curve γ at $\gamma(0) = p$),

$$X(f)(p) = \frac{d}{dt} f(\gamma(t))\bigg|_{t=0}.$$

Given two vector fields X and Y on M, there is a unique vector field $[X, Y]$, called the **Lie Bracket** of X and Y, such that

$$[X, Y](f) = X(Y(f)) - Y(X(f)).$$

It is sometimes convenient to work in coordinates. A **local frame** $\{L_i\}$ is a collection of vector fields defined on an open set $U \subseteq M$ such that at each point $p \in U$, the vectors $\{L_i(p)\} \subseteq T_pM$ form a basis of T_pM. The vector fields $\{L_i\}$ are called a **local orthonormal frame** if at each point in U, the $\{L_i\}$ are orthonormal with repect to g. Some manifolds only have local frames, not global ones; that is, you can't define a smooth family of vector fields over the whole manifold which forms a basis of the tangent space at each point. This is true, for example of $\mathbb{S}^2 \subseteq \mathbb{R}^3$.

We need a few more notions in order to get to curvature. Firstly, a **connection** ∇ on M is a way of differentiating one vector field in the direction of another: a connection ∇ is a bilinear form on vector fields that assigns to vector fields X and Y a new vector field $\nabla_X Y$, such that for any smooth function $f : M \to \mathbb{R}$,

$$\nabla_{fX} Y = f \nabla_X Y \qquad \textit{and} \qquad \nabla_X (fY) = f \nabla_X (Y) + X(f) Y.$$

A connection is called **torsion-free** if $\nabla_X Y - \nabla_Y X = [X, Y]$. There is a special connection on a Riemannian manifold, called the **Levi-Civita connection**, which is the unique torsion-free connection with the property that

$$X(g(Y, Z)) = g(\nabla_X Y, Z) + g(Y, \nabla_X Z).$$

This property may look not obviously interesting, but geometrically, it is a compatibility condition of the connection ∇ with g. There is a notion of transporting a vector field in a "parallel way" along a curve, which is defined by the connection. The condition above means (this is not obvious) that the inner product defined by g of two vector fields at a point is unchanged if you parallel-transport the vector fields (using ∇ to define "parallel") along any curve.

Finally, we can define the **Riemannian curvature tensor** $R(X, Y)$: to each pair of vector fields X and Y on M, we associate an operator $R(X, Y)$ on vector fields defined by

$$R(X, Y)(Z) := \nabla_X(\nabla_Y Z) - \nabla_Y(\nabla_X Z) - \nabla_{[X,Y]} Z.$$

The **Ricci curvature tensor** is the function $\mathrm{Ric}(X, Y)$ on M which, at each point $p \in M$, is the trace of the linear map on $T_p M$ defined by $Z \mapsto R(Z, Y)(X)$. In orthonormal local coordinates $\{L_i\}$,

$$\mathrm{Ric}(X, Y) = \sum_i g(R(X, L_i)L_i, Y).$$

(Note that seeing that this coordinate expression is right involves using some of the symmetries of R.) The Bakry-Émery criterion can be made more general, but for our purposes it suffices to formulate it as follows.

Theorem 3.8 (Bakry–Émery)**.** *Let (M, g) be a compact, connected, m-dimensional Riemannian manifold with normalized volume measure μ. Suppose that there is a constant $c > 0$ such that for each $p \in M$ and each $v \in T_p M$,*

$$\mathrm{Ric}_p(v, v) \geq \frac{1}{c} g_p(v, v).$$

Then μ satisfies a log-Sobolev inequality with constant c.

3.3 Concentration for the compact classical groups

Theorem 3.9. *The matrix groups and cosets $\mathbb{SO}(n)$, $\mathbb{SO}^-(n)$, $\mathbb{SU}(n)$, $\mathbb{U}(n)$, and $\mathbb{Sp}(2n)$ with Haar probability measure and the Hilbert–Schmidt metric, satisfy logarithmic Sobolev inequalities with constants*

G	C_G
$\mathbb{SO}(n)$, $\mathbb{SO}^-(n)$	$\frac{4}{n-2}$
$\mathbb{SU}(n)$	$\frac{2}{n}$
$\mathbb{U}(n)$	$\frac{6}{n}$
$\mathbb{Sp}(2n)$	$\frac{1}{2n+1}$

We saw in Lecture 1 (Lemma 1.4) that the the geodesic distance on $\mathbb{U}(n)$ is bounded above by $\pi/2$ times the Hilbert–Schmidt distance. Thus Theorem 3.9 implies, for example that $\mathbb{U}(n)$ equipped with the geodesic distance also satisfies a log-Sobolev inequality, with constant $3\pi^2/2n$.

Exercise 3.10. Prove that Theorem 3.9 does indeed imply a LSI for the geodesic distance with constant $3\pi^2/n$.

The proof of Theorem 3.9 for each G except $\mathbb{U}(n)$ follows immediately from the Bakry–Émery Theorem, by the following curvature computations.

Proposition 3.11. *If G_n is one of $\mathbb{SO}(n)$, $\mathbb{SO}^-(n)$ $\mathbb{SU}(n)$, or $\mathbb{Sp}(2n)$, then for each $U \in G_n$ and each $X \in T_U G_n$,*

$$\mathrm{Ric}_U(X, X) = c_{G_n} g_U(X, X),$$

where g_U is the Hilbert-Schmidt metric and c_{G_n} is given by

G	c_G
$\mathbb{SO}(n)$, $\mathbb{SO}^-(n)$	$\frac{n-2}{4}$
$\mathbb{SU}(n)$	$\frac{n}{2}$
$\mathbb{Sp}(2n)$	$2n+1$

This gives us the concentration phenomenon we're after on all of the G_n except $\mathbb{O}(n)$ and $\mathbb{U}(n)$. Now, on $\mathbb{O}(n)$ we can't actually expect more and indeed more is not true, because $\mathbb{O}(n)$ is disconnected. We have the best we can hope for already, namely concentration on each of the pieces. In the case of $\mathbb{U}(n)$, though, we in fact do have the same kind of concentration that we have on $\mathbb{SU}(n)$. There is no non-zero lower bound on the Ricci curvature on $\mathbb{U}(n)$, but the log-Sobolev inequality there follows from the one on $\mathbb{SU}(n)$. The crucial observation is the following coupling of the Haar measures on $\mathbb{SU}(n)$ and $\mathbb{U}(n)$.

Lemma 3.12. *Let θ be uniformly distributed in $\left[0, \frac{2\pi}{n}\right]$ and let $V \in \mathbb{SU}(n)$ be uniformly distributed, with θ and V independent. Then $e^{i\theta}V$ is uniformly distributed in $\mathbb{U}(n)$.*

Proof. Let X be uniformly distributed in $[0,1)$, K uniformly distributed in $\{0,\dots,n-1\}$, and V uniformly distributed in $\mathbb{SU}(n)$ with (X,K,V) independent. Consider

$$U = e^{2\pi i X/n} e^{2\pi i K/n} V.$$

On one hand, it is easy to see that $(X+K)$ is uniformly distributed in $[0,n]$, so that $e^{2\pi i(X+K)/n}$ is uniformly distributed on $\mathbb{S}^1$. Thus $U \stackrel{d}{=} \omega V$ for ω uniform in $\mathbb{S}^1$ and independent of V. One can then show that the distribution of ωV is translation-invariant on $\mathbb{U}(n)$, and thus yields Haar measure.

On the other hand, if I_n is the $n \times n$ identity matrix, then $e^{2\pi i K/n} I_n \in \mathbb{SU}(n)$. By the translation invariance of Haar measure on $\mathbb{SU}(n)$ this implies that $e^{2\pi i K/n} V \stackrel{d}{=} V$, and so $e^{2\pi i X/n} V \stackrel{d}{=} U$. □

Exercise 3.13. Prove carefully that if ω is uniform in $\mathbb{S}^1$ and U is Haar-distributed in $\mathbb{SU}(n)$ with ω, U independent, then ωU is Haar-distributed in $\mathbb{U}(n)$.

Using this coupling lets us prove the log-Sobolev inequality on $\mathbb{U}(n)$ via the tensorization property of LSI.

Proof of Theorem 3.9. First, for the interval $[0,2\pi]$ equipped with its standard metric and uniform measure, the optimal constant in (3.1) for functions f with $f(0) = f(2\pi)$ is known to be 1, see e.g. [**?**]. This fact completes the proof — with a better constant than stated above — in the case $n = 1$; from now on, assume that $n \geq 2$.

Suppose that $f : [0,\pi] \to \mathbb{R}$ is locally Lipschitz, and define a function $\tilde{f} : [0,2\pi] \to \mathbb{R}$ by reflection:

$$\tilde{f}(x) := \begin{cases} f(x), & 0 \leq x \leq \pi; \\ f(2\pi - x), & \pi \leq x \leq 2\pi. \end{cases}$$

Then $\tilde{f}$ is locally Lipschitz and $\tilde{f}(2\pi) = \tilde{f}(0)$, so $\tilde{f}$ satisfies a LSI for uniform measure on $[0,2\pi]$ with constant 1. If $\mu_{[a,b]}$ denotes uniform (probability) measure on $[a,b]$, then

$$\mathrm{Ent}_{\mu_{[0,2\pi]}}(\tilde{f}^2) = \mathrm{Ent}_{\mu_{[0,\pi]}}(f^2),$$

and

$$\frac{1}{2\pi}\int_0^{2\pi} |\nabla \tilde{f}(x)|^2 dx = \frac{1}{\pi}\int_0^{\pi} |\nabla f(x)|^2 dx,$$

so f itself satisfies a LSI for uniform measure on $[0,\pi]$ with constant 1 as well. In fact, the constant 1 here is optimal (see Exercise 3.14).

It then follows by a scaling argument that the optimal logarithmic Sobolev constant on $\left[0, \frac{\pi\sqrt{2}}{\sqrt{n}}\right)$ is $2/n$ (for $g : \left[0, \frac{\pi\sqrt{2}}{\sqrt{n}}\right) \to \mathbb{R}$, apply the LSI to $g\left(\sqrt{\frac{2}{n}}x\right)$ and rearrange it to get the LSI on $\left[0, \frac{\pi\sqrt{2}}{\sqrt{n}}\right)$.)

By Theorem 3.9 $\mathbb{SU}(n)$ satisfies a log-Sobolev inequality with constant $2/n$ when equipped with its geodesic distance, and hence also when equipped with the Hilbert–Schmidt metric. By the tensorization property of log-Sobolev inequalities in Euclidean spaces (Lemma 3.6), the product space $\left[0, \frac{\pi\sqrt{2}}{\sqrt{n}}\right) \times \mathbb{SU}(n)$, equipped with the L_2-sum metric, satisfies a log-Sobolev inequality with constant $2/n$ as well.

Define the map $F : \left[0, \frac{\pi\sqrt{2}}{\sqrt{n}}\right) \times \mathbb{SU}(n) \to \mathbb{U}(n)$ by $F(t, V) = e^{\sqrt{2}it/\sqrt{n}}V$. By Lemma 3.12, the push-forward via F of the product of uniform measure on $\left[0, \frac{\pi\sqrt{2}}{\sqrt{n}}\right)$ with uniform measure on $\mathbb{SU}(n)$ is uniform measure on $\mathbb{U}(n)$. Moreover, this map is $\sqrt{3}$-Lipschitz:

$$\begin{aligned}
\left\|e^{\sqrt{2}it_1/\sqrt{n}}V_1 - e^{\sqrt{2}it_2/\sqrt{n}}V_2\right\|_{HS} &\leq \left\|e^{\sqrt{2}it_1/\sqrt{n}}V_1 - e^{\sqrt{2}it_1/\sqrt{n}}V_2\right\|_{HS} \\
&\quad + \left\|e^{\sqrt{2}it_1/\sqrt{n}}V_2 - e^{\sqrt{2}it_2/\sqrt{n}}V_2\right\|_{HS} \\
&= \|V_1 - V_2\|_{HS} + \left\|e^{\sqrt{2}it_1/\sqrt{n}}I_n - e^{\sqrt{2}it_2/\sqrt{n}}I_n\right\|_{HS} \\
&\leq \|V_1 - V_2\|_{HS} + \sqrt{2}\,|t_1 - t_2| \\
&\leq \sqrt{3}\sqrt{\|V_1 - V_2\|_{HS}^2 + |t_1 - t_2|^2}.
\end{aligned}$$

Since the map F is $\sqrt{3}$-Lipschitz, its image $\mathbb{U}(n)$ with the (uniform) image measure satisfies a logarithmic Sobolev inequality with constant $(\sqrt{3})^2\frac{2}{n} = \frac{6}{n}$. □

Exercise 3.14. Prove that the optimal log-Sobolev constant for uniform measure on $[0, \pi]$ is 1. Here are two possible approaches:

1. Suppose uniform measure on $[0, \pi]$ satisfied a LSI with constant $C < 1$. Let $f : [0, 2\pi]$ be locally Lipschitz with $f(0) = f(2\pi)$. Decompose all of the integrals in the expression for the entropy of f^2 w.r.t. uniform measure on $[0, 2\pi]$ into the part on $[0, \pi]$ and the part on $[\pi, 2\pi]$. Use the concavity of the logarithm and your assumed LSI to get an estimate for the entropy of f^2 on $[0, 2\pi]$. Now obtain a contraditction by observing that whether or not the crucial inequality holds is invariant under the transformation $f(x) \longmapsto f(2\pi - x)$, and so you may use whichever version of f is more convenient.
2. Consider the function $f(x) = 1 + \epsilon\cos(x)$ on $[0, \pi]$. Suppose you had an LSI on $[0, \pi]$ with constant $C < 1$. Apply it to this f and expand $\mathrm{Ent}_{\mu_{[0,\pi]}}(f^2)$ in powers of ϵ to get a contradiction.

From our log-Sobolev inequalities, we finally get the concentration we want on the compact classical groups, as follows.

Corollary 3.15. *Given $n_1, \ldots, n_k \in \mathbb{N}$, let $X = G_{n_1} \times \cdots \times G_{n_k}$, where for each of the n_i, G_{n_i} is one of $\mathbb{SO}(n_i)$, $\mathbb{SO}^-(n_i)$, $\mathbb{SU}(n_i)$, $\mathbb{U}(n_i)$, or $\mathbb{Sp}(2n_i)$. Let X be equipped with the L_2-sum of Hilbert–Schmidt metrics on the G_{n_i}. Suppose that $F : X \to \mathbb{R}$ is L-Lipschitz,*

and that $\{U_j \in G_{n_j} : 1 \leq j \leq k\}$ *are independent, Haar-distributed random matrices. Then for each* $t > 0$*,*

$$\mathbb{P}\big[F(U_1, \ldots, U_k) \geq \mathbb{E}F(U_1, \ldots, U_k) + t\big] \leq e^{-(n-2)t^2/12L^2},$$

where $n = \min\{n_1, \ldots, n_k\}$.

Proof. By Theorem 3.9 and Lemma 3.6, X satisfies a log-Sobolev inequality with constant $6/(n-2)$. The stated concentration inequality then follows from the Herbst argument. □

Lecture 4

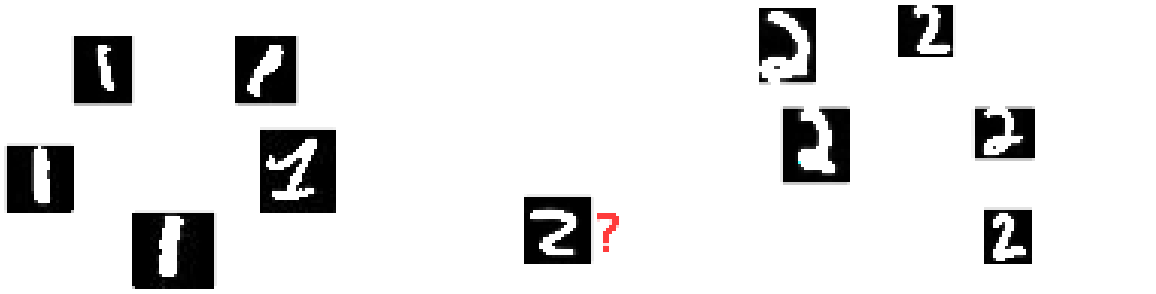
Applications of Concentration

4.1 The Johnson-Lindenstrauss Lemma

A huge area of application in computing is that of *dimension-reduction*. In this day and age, we are often in the situation of having (sometimes large, sometimes not so large) data sets that live in very high-dimension. For example, a digital image can be encoded as a matrix, with each entry corresponding to one pixel, and the entry specifying the color of that pixel. So if you had a small black and white image whose resolution was, say 100×150 pixels, you would encode it as a vector in $\{0,1\}^{15,000}$. An issue that causes problems is that many algorithms for analyzing such high-dimensional data have their run-time increase very quickly as the dimension of the data increases, to the point that analyzing the data in the most obvious way becomes computationally infeasible. The idea of dimension reduction is that in many situations, the desired algorithm can be at least approximately carried out in a much lower-dimensional setting than the one the data come to you in, and that can make computationally infeasible problems feasible.

A motivating problem

Suppose you have a data set consisting of black and white images of hand-written examples of the numbers 1 and 2. So you have a reference collection $\mathcal{X}$ of n points in $\mathbb{R}^d$, where d is the number of pixels in each image. You want to design a computer program so that one can input an image of a hand-written number, and the computer can tell whether it's a 1 or a 2. So the computer will have a query point $q \in \mathbb{R}^d$, and the natural thing to do is to program it to find the closest point in the reference set $\mathcal{X}$ to q; the computer then reports that the input image was of the same number as that closest point in $\mathcal{X}$.



P. Indyk

The naïve approach would be for the computer to calculate the distance from q to each of the points of $\mathcal{X}$ in turn, keeping track of which point in $\mathcal{X}$ has so far been the closest. Such an algorithm runs in $O(nd)$ steps. Remember that d can be extremely large if our images are fairly high resolution, so nd steps might be computationally infeasible. Many mathematicians' first remark at this point is that the problem only has to be solved within the span of the points of $\mathcal{X}$ and q, so that one can *a priori* replace d by n. Actually doing this, though, means you have to find an orthonormal basis for the subspace you plan to work in, so in general you can't save time this way.

The idea of dimension reduction is to find a way to carry out the nearest point algorithm within some much lower-dimensional space, in such a way that you are guarranteed (or to be more realistic, very likely) to still find the closest point, and without having to do much work to figure out which lower-dimensional space to work in. This sounds impossible, but the geometry of high-dimensional spaces often turns out to be surprising. An important result about high-dimensional geometry that has inspired many randomized algorithms incorporating dimension-reduction is the following.

Lemma 4.1 (The Johnson–Lindenstrauss Lemma)**.** *There are absolute constants c, C such that the following holds.*

Let $\{x_j\}_{j=1}^n \subseteq \mathbb{R}^d$, and let P be a random $k \times d$ matrix, consisting of the first k rows of a Haar-distributed random matrix in $\mathbb{O}(d)$. Fix $\epsilon > 0$ and let $k = \frac{a \log(n)}{\epsilon^2}$. With probability $1 - Cn^{2-ac}$

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \left(\frac{d}{k}\right) \|Px_i - Px_j\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2 \tag{4.1}$$

for all $i, j \in \{1, \ldots, n\}$.

What the lemma says is that one can take a set of n points in $\mathbb{R}^d$ and project them onto a random subspace of dimension on the order of $\log(n)$ so that, after appropriate rescaling, the pairwise distances between the points hardly changes. The practical conclusion of this is that if your problem is about the metric structure of the data (finding the closest point as above, finding the most separated pair of points, finding the minimum length spanning tree

of a graph,etc.), there is no need to work in the high-dimensional space that the data naturally live in, and that moreover there is no need to work hard to pick a lower-dimensional subspace onto which to project: a random one should do.

Exercise 4.2. Verify that for $x \in \mathbb{R}^d$ and P as above, $\frac{d}{k}\mathbb{E}\|Px\|^2 = \|x\|^2$.

Getting an almost-solution, with high probability

The discussion above suggests that we try to solve the problem of finding the closest point to q in $\mathcal{X}$ by choosing a random $k \times d$ matrix P to be the first k rows of a Haar-distributed $U \in \mathbb{O}(d)$, then finding the closest point in $\{Px :\in \mathcal{X}\}$ to Pq. There are two obvious issues here. One is that we might have the bad luck to choose a bad matrix P that doesn't satisfy (4.1). But that is *very* unlikely, and so we typically just accept the risk and figure it won't actually happen.

There is a second issue, though, which is that it's possible that we choose P that does satisfy (4.1), but that the closest point in $\{Px :\in \mathcal{X}\}$ to Pq is Py, whereas the closest point in $\mathcal{X}$ to q is z, with $y \neq z$. In that case, although our approach will yield the wrong value for the closest point (y instead of z), we have by choice of y and (4.1) that

$$\|q-y\| \leq \sqrt{\frac{d}{k(1-\epsilon)}}\|Pq-Py\| \leq \sqrt{\frac{d}{k(1-\epsilon)}}\|Pq-Pz\| \leq \sqrt{\frac{1+\epsilon}{1-\epsilon}}\|q-z\|.$$

So even though z is the true closest point to q, y is almost as close. In our example of recognizing whether a hand-written number is a 1 or a 2, it seems likely that even if we don't find the exact closest point in the reference set, we'll still manage to correctly identify the number, which is all we actually care about.

For being willing to accept an answer which may be not quite right, and accept the (tiny) risk that we'll choose a bad matrix, we get a lot in return. The naïve algorithm we mentioned at the beginning now runs in $O(n\log(n))$ steps.

Proof of Johnson-Lindenstrauss

Given $\{x_i\}_{i=1}^n \subseteq \mathbb{R}^d$, $\epsilon > 0$, and U a Haar-distributed random matrix in $\mathbb{O}(d)$, let P be the $k \times d$ matrix consisting of the first k rows of U. We want to show that for each pair (i,j),

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \left(\frac{d}{k}\right)\left\|Px_i - Px_j\right\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2$$

with high probability, or equivalently,

$$\sqrt{1-\epsilon} \leq \sqrt{\frac{d}{k}}\left\|Px_{i,j}\right\| \leq \sqrt{1+\epsilon}$$

for $x_{i,j} := \frac{x_i - x_j}{\|x_i - x_j\|}$.

For notational convenience, fix i and j for the moment and let $x = x_{i,j}$. For such an $x \in \mathbb{S}^{d-1}$ fixed, consider the function $F_x : \mathbb{O}(d) \to \mathbb{R}$ defined by

$$F_x(U) = \sqrt{\frac{d}{k}}\Big\|Px\Big\|.$$

Let $U, U' \in \mathbb{O}(d)$, and let P, P' denote the matrices of the first k rows of U and U'. Then

$$\Big|F_x(U) - F_x(U')\Big| = \sqrt{\frac{d}{k}}\Big|\|Px\| - \|P'x\|\Big| \leq \sqrt{\frac{d}{k}}\big\|(P - P')x\big\|.$$

Exercise 4.3. Prove that $\big\|(P - P')x\big\| \leq d_{H,S}(U, U')$.

Hint: First show $\big\|(P - P')x\big\| \leq \big\|(U - U')x\big\|$.

That is, the function F_x is $\sqrt{\frac{d}{k}}$-Lipschitz on $\mathbb{O}(d)$ with respect to $d_{HS}(\cdot, \cdot)$ In particular, F_x is also $\sqrt{\frac{d}{k}}$-Lipschitz when restricted to either $\mathbb{SO}(d)$ or $\mathbb{SO}^-(d)$.

The idea is to apply concentration of measure to the function F_x. We want a concentration inequality on the full orthogonal group, which we will get from similar inequalities on the two components.

Haar measure on $\mathbb{O}(d)$ can be described in terms of Haar measure on $\mathbb{SO}(d)$ and $\mathbb{SO}^-(d)$ as follows. Let U_1 be Haar-distributed in $\mathbb{SO}(d)$ and let U_2 be the matrix obtained from U by swapping the last two rows. Then we saw in Lecture 1 that U_2 is Haar-distributed in $\mathbb{SO}^-(d)$, and so if U is equal to U_1 with probability $\frac{1}{2}$ and equal to U_2 with probability $\frac{1}{2}$, then U is Haar distributed on $\mathbb{O}(d)$. Note that as long as $k \leq d - 2$, the matrix P_1 of the first k rows of U_1 is the same as the matrix P_2 of the first k rows of U_2. It follows that

$$\mathbb{E}F_x(U) = \mathbb{E}F_x(U_1) = \mathbb{E}F_x(U_2).$$

Because F_x is $\sqrt{\frac{d}{k}}$-Lipschitz when restricted to either $\mathbb{SO}(d)$ or $\mathbb{SO}^-(d)$, concentration of measure implies that, for $j = 1, 2$,

$$\mathbb{P}\left[|F_x(U_j) - \mathbb{E}F_x(U_j)| \geq \epsilon\right] \leq Ce^{-ck\epsilon^2}.$$

It then follows by conditioning on whether $U = U_1$ or $U = U_2$ that

$$\mathbb{P}\left[|F_x(U) - \mathbb{E}F_x(U)| \geq \epsilon\right] \leq Ce^{-ck\epsilon^2}. \tag{4.2}$$

To complete the proof, we need to show that $\mathbb{E}F_x(U) \approx 1$.

By the invariance of Haar measure under translation,

$$Px \overset{d}{=} Pe_1 = (U_{11}, \ldots, U_{k1}),$$

where e_1 is the first standard basis vector in $\mathbb{R}^d$. It follows that

$$F_x(U) \stackrel{d}{=} \sqrt{\left(\frac{d}{k}\right)(U_{11}^2 + \cdots + U_{k1}^2)}.$$

We saw in the first lecture that $\mathbb{E}U_{i1}^2 = \frac{1}{d}$ for each i, so $\mathbb{E}\big[F_x(U)\big]^2 = 1$; written slightly differently,

$$1 = \mathrm{Var}(F_x(U)) + \big(\mathbb{E}F_x(U)\big)^2.$$

By Fubini's theorem and the concentration inequality (4.2),

$$\mathrm{Var}(F_x(U)) = \int_0^\infty \mathbb{P}\big[|F_x(U) - \mathbb{E}F_x(U)|^2 \geq t\big]dt \leq \int_0^\infty Ce^{-ckt}dt = \frac{C}{ck},$$

so that

$$\sqrt{1 - \frac{C}{ck}} \leq \mathbb{E}F_x(U) \leq 1.$$

Recall that $k = \frac{a\log(n)}{\epsilon^2}$. As long as $\epsilon < \frac{ca\log(n)}{C+ca\log(n)}$, this means that $1 - \frac{\epsilon}{2} \leq \mathbb{E}F_x(U) \leq 1$, and so

$$\mathbb{P}\left[\big|F_x(U) - 1\big| > \epsilon\right] \leq Ce^{-\frac{ck\epsilon^2}{4}}; \tag{4.3}$$

that is, with probability at least $1 - Ce^{-\frac{ck\epsilon^2}{4}}$,

$$1 - \epsilon \leq \sqrt{\frac{d}{k}}\|Px\| \leq 1 + \epsilon.$$

Returning to the original formulation, for each pair (i, j), there is a set of probability at least $1 - Ce^{-ck\epsilon^2}$ such that

$$(1-\epsilon)^2\|x_i - x_j\|^2 \leq \left(\frac{d}{k}\right)\big\|Ux_i - Ux_j\big\|^2 \leq (1+\epsilon)^2\|x_i - x_j\|^2.$$

There are fewer than n^2 pairs (i, j), so a simple union bound gives that the above statement holds *for all pairs* (i, j) with probability at least $1 - \frac{C}{n^{ac-2}}$. □

Exercise 4.4. In the course of the proof, we assumed that $k \leq d - 2$ for convenience. While we certainly expect this to be true in applications, show that this wasn't necessary by checking the cases $k = d - 1$ and $k = d$.

4.2 Uniform approximation of the empirical spectral measure of powers of random unitary matrices

In this section we give the proof of Theorem 2.15 on the rate of convergence of the empirical spectral measure of a Haar random matrix to the uniform distribution. The result quoted there is a corollary of the following.

Theorem 4.5. *Let $\mu_{n,m}$ be the spectral measure of U^m, where $1 \leq m \leq n$ and $U \in \mathbb{U}(n)$ is distributed according to Haar measure, and let ν denote the uniform measure on $\mathbb{S}^1$. Then for each $p \geq 1$,*

$$\mathbb{E}W_p(\mu_{n,m}, \nu) \leq Cp\frac{\sqrt{m\left[\log\left(\frac{n}{m}\right)+1\right]}}{n},$$

where $C > 0$ is an absolute constant.

Moreover, for each $t > 0$,

$$\mathbb{P}\left[W_p(\mu_{n,m}, \nu) \geq C\frac{\sqrt{m\left[\log\left(\frac{n}{m}\right)+1\right]}}{n} + t\right] \leq \exp\left[-\frac{n^2t^2}{24m}\right]$$

for $1 \leq p \leq 2$ and

$$\mathbb{P}\left[W_p(\mu_{n,m}, \nu) \geq Cp\frac{\sqrt{m\left[\log\left(\frac{n}{m}\right)+1\right]}}{n} + t\right] \leq \exp\left[-\frac{n^{1+2/p}t^2}{24m}\right]$$

for $p > 2$, where $C > 0$ is an absolute constant.

This kind of change in behavior at $p = 2$ is typical for the Wasserstein distances.

By a simple application of the Borel-Cantelli lemma, one gets an almost sure rate of convergence, as follows.

Corollary 4.6. *Suppose that for each n, $U_n \in \mathbb{U}(n)$ is Haar-distributed and $1 \leq m_n \leq n$. Let ν denote the uniform measure on $\mathbb{S}^1$. There is an absolute constant C such that given $p \geq 1$, with probability 1, for all sufficiently large n,*

$$W_p(\mu_{n,m_n}, \nu) \leq C\frac{\sqrt{m_n \log(n)}}{n}$$

if $1 \leq p \leq 2$ and

$$W_p(\mu_{n,m_n}, \nu) \leq Cp\frac{\sqrt{m_n \log(n)}}{n^{\frac{1}{2}+\frac{1}{p}}}$$

if $p > 2$.

Exercise 4.7. Prove Corollary 4.6

The first step in proving Theorem 4.5 is to prove a concentration result for the number $\mathcal{N}_\theta^{(m)}$ of eigenangles of U^m in $[0, \theta)$. We get such a result as a concenquence of the following remarkable property of determinantal point processes.

Proposition 4.8. *Let $K : \Lambda \times \Lambda \to \mathbb{C}$ be a kernel on a locally compact Polish space Λ such that the corresponding integral operator $\mathcal{K} : L^2(\mu) \to L^2(\mu)$ defined by*

$$\mathcal{K}(f)(x) = \int K(x, y) f(y)\, d\mu(y)$$

is self-adjoint, nonnegative, and locally trace-class with eigenvalues in $[0, 1]$. For $D \subseteq \Lambda$ measurable, let $K_D(x, y) = \mathbb{1}_D(x) K(x, y) \mathbb{1}_D(y)$ be the restriction of K to D. Suppose that D is such that K_D is trace-class; denote by $\{\lambda_k\}_{k \in \mathcal{A}}$ the eigenvalues of the corresponding operator $\mathcal{K}_D$ on $L^2(D)$ ($\mathcal{A}$ may be finite or countable) and denote by $\mathcal{N}_D$ the number of particles of the determinantal point process with kernel K which lie in D. Then

$$\mathcal{N}_D \stackrel{d}{=} \sum_{k \in \mathcal{A}} \xi_k,$$

where "$\stackrel{d}{=}$" denotes equality in distribution and the ξ_k are independent Bernoulli random variables with $\mathbb{P}[\xi_k = 1] = \lambda_k$ and $\mathbb{P}[\xi_k = 0] = 1 - \lambda_k$.

This result is quite valuable for us, because it tells us (once we've checked the conditions; see Exercise 4.9) that $\mathcal{N}_\theta^{(1)}$ is distributed exactly as a sum of n independent Bernoulli random variables. Moreover, thanks to Rains' Theorem (Theorem 2.16), $\mathcal{N}_\theta^{(m)}$ is equal in distribution to the total number of eigenvalue angles in $[0, \theta)$ of each of $U_0 \ldots, U_{m-1}$, where $U_0, \ldots, U_{m-1}$ are independent and U_j is Haar-distributed in $\mathbb{U}\left(\left\lceil \frac{n-j}{m} \right\rceil\right)$; that is,

$$\mathcal{N}_\theta^{(m)} \stackrel{d}{=} \sum_{j=0}^{m-1} \mathcal{N}_{j,\theta},$$

where the $\mathcal{N}_{j,\theta}$ are the independent counting functions corresponding to $U_0, \ldots, U_{m-1}$. It is therefore also true that $\mathcal{N}_\theta^{(m)}$ is distributed exactly as a sum of n independent Bernoulli random variables.

Exercise 4.9. Show that the operator $\mathcal{K} : L^2([0, 2\pi)) \to L^2([0, 2\pi))$ defined by

$$\mathcal{K}(f)(x) = \frac{1}{2\pi} \int_0^{2\pi} K_N(x - y) f(y) dy,$$

for $K_N(z) = \sum_{j=0}^{N-1} e^{ijz}$ is self-adjoint, nonnegative, and trace-class.

Classical probability has quite a few things to say about sums of independent Bernoulli random variables; in particular, an application of Bernstein's inequality (Theorem 3.1) will give us the concentration we need for the eigenvalue counting function. Specifically, for each $t > 0$ we have

$$\mathbb{P}\left[\left|\mathcal{N}_\theta^{(m)} - \mathbb{E}\mathcal{N}_\theta^{(m)}\right| > t\right] \leq 2\exp\left(-\min\left\{\frac{t^2}{4\sigma^2}, \frac{t}{2}\right\}\right), \tag{4.4}$$

where $\sigma^2 = \operatorname{Var}\mathcal{N}_\theta^{(m)}$.

Of course, in order to apply (4.4), it is necessary to estimate $\mathbb{E}\mathcal{N}_\theta^{(m)}$ and σ^2. Rains' Theorem again means we only need to do any real work in the case $m = 1$. To do the computations, we use the kernel of the eigenvalue process given in Theorem 2.18 together with the formulae from Lemma 2.19.

Proposition 4.10. *Let U be uniform in $\mathbb{U}(N)$. Then for all $\theta \in [0, 2\pi)$,*

$$\mathbb{E}\mathcal{N}_\theta^{(1)} = \frac{n\theta}{2\pi},$$

and

$$\operatorname{Var}\mathcal{N}_\theta^{(1)} \leq \log n + 1.$$

Proof. Computing $\mathbb{E}\mathcal{N}_\theta^{(1)}$ is trivial, either by symmetry or by Lemma 2.19.

To compute $\operatorname{Var}\mathcal{N}_\theta^{(1)}$, note first that if $\theta \in (\pi, 2\pi)$, then $\mathcal{N}_\theta \stackrel{d}{=} N - \mathcal{N}_{2\pi-\theta}$, and so it suffices to assume that $\theta \leq \pi$. By Proposition 2.18 and Lemma 2.19,

$$\begin{aligned}
\operatorname{Var}\mathcal{N}_\theta &= \frac{1}{4\pi^2}\int_0^\theta\int_\theta^{2\pi} S_n(x-y)^2\,dx\,dy = \frac{1}{4\pi^2}\int_0^\theta\int_{\theta-y}^{2\pi-y}\frac{\sin^2\left(\frac{nz}{2}\right)}{\sin^2\left(\frac{z}{2}\right)}\,dz\,dy \\
&= \frac{1}{4\pi^2}\left[\int_0^\theta \frac{z\sin^2\left(\frac{nz}{2}\right)}{\sin^2\left(\frac{z}{2}\right)}\,dz + \int_\theta^{2\pi-\theta}\frac{\theta\sin^2\left(\frac{nz}{2}\right)}{\sin^2\left(\frac{z}{2}\right)}\,dz + \int_{2\pi-\theta}^{2\pi}\frac{(2\pi-z)\sin^2\left(\frac{nz}{2}\right)}{\sin^2\left(\frac{z}{2}\right)}\,dz\right] \\
&= \frac{1}{2\pi^2}\left[\int_0^\theta \frac{z\sin^2\left(\frac{nz}{2}\right)}{\sin^2\left(\frac{z}{2}\right)}\,dz + \int_\theta^{\pi}\frac{\theta\sin^2\left(\frac{nz}{2}\right)}{\sin^2\left(\frac{z}{2}\right)}\,dz\right].
\end{aligned}$$

For the first integral, since $\sin\left(\frac{z}{2}\right) \geq \frac{z}{\pi}$ for all $z \in [0, \theta]$, if $\theta > \frac{1}{n}$, then

$$\int_0^\theta \frac{z\sin^2\left(\frac{nz}{2}\right)}{\sin^2\left(\frac{z}{2}\right)}\,dz \leq \int_0^{\frac{1}{n}}\frac{(\pi n)^2 z}{4}\,dz + \int_{\frac{1}{n}}^\theta \frac{\pi^2}{z}\,dz = \pi^2\left(\frac{1}{8} + \log(n) + \log(\theta)\right).$$

If $\theta \leq \frac{1}{n}$, there is no need to break up the integral and one simply has the bound $\frac{(\pi n\theta)^2}{8} \leq \frac{\pi^2}{8}$. Similarly, if $\theta < \frac{1}{n}$, then

$$\begin{aligned}
\int_\theta^\pi \frac{\theta\sin^2\left(\frac{nz}{2}\right)}{\sin^2\left(\frac{z}{2}\right)}\,dz &\leq \int_\theta^{\frac{1}{n}}\frac{\theta(\pi n)^2}{4}\,dz + \int_{\frac{1}{n}}^\pi \frac{\pi^2\theta}{z^2}\,dz \\
&= \frac{\pi^2\theta n}{4}(1 - n\theta) + \pi^2 n\theta - \pi\theta \leq \frac{5\pi^2}{4};
\end{aligned}$$

if $\theta \geq \frac{1}{n}$, there is no need to break up the integral and one simply has a bound of π^2.

All together,

$$\operatorname{Var} \mathcal{N}_\theta \leq \log(n) + \frac{11}{16}. \qquad \square$$

Corollary 4.11. *Let U be uniform in $\mathbb{U}(n)$ and $1 \leq m \leq n$. For $\theta \in [0, 2\pi)$, let $\mathcal{N}_\theta^{(m)}$ be the number of eigenvalue angles of U^m in $[0, \theta)$. Then*

$$\mathbb{E}\mathcal{N}_\theta^{(m)} = \frac{n\theta}{2\pi} \qquad \textit{and} \qquad \operatorname{Var} \mathcal{N}_\theta^{(m)} \leq m\left(\log\left(\frac{n}{m}\right) + 1\right).$$

Proof. This follows immediately from Theorem 2.16; note that the n/m in the variance bound, as opposed to the more obvious $\lceil n/m \rceil$, follows from the concavity of the logarithm. $\square$

Putting these estimates together with Equation (4.4) gives that for all $t > 0$,

$$\mathbb{P}\left[\left|\mathcal{N}_\theta^{(m)} - \frac{n\theta}{2\pi}\right| > t\right] \leq 2\exp\left(-\min\left\{\frac{t^2}{4m\left(\log\left(\frac{n}{m}\right)+1\right)}, \frac{t}{2}\right\}\right). \tag{4.5}$$

It's fairly straighforward to use this inequality to obtain concentration for the individual eigenangles around their predicted values, as follows.

Lemma 4.12. *Let $1 \leq m \leq n$ and let $U \in \mathbb{U}(n)$ be uniformly distributed. Denote by $e^{i\theta_j}$, $1 \leq j \leq n$, the eigenvalues of U^m, ordered so that $0 \leq \theta_1 \leq \cdots \leq \theta_n < 2\pi$. Then for each j and $u > 0$,*

$$\mathbb{P}\left[\left|\theta_j - \frac{2\pi j}{n}\right| > \frac{4\pi}{n}u\right] \leq 4\exp\left[-\min\left\{\frac{u^2}{m\left(\log\left(\frac{n}{m}\right)+1\right)}, u\right\}\right]. \tag{4.6}$$

Proof. For each $1 \leq j \leq n$ and $u > 0$, if $j + 2u < n$ then

$$\begin{aligned}\mathbb{P}\left[\theta_j > \frac{2\pi j}{n} + \frac{4\pi}{n}u\right] &= \mathbb{P}\left[\mathcal{N}^{(m)}_{\frac{2\pi(j+2u)}{n}} < j\right] = \mathbb{P}\left[j + 2u - \mathcal{N}^{(m)}_{\frac{2\pi(j+2u)}{n}} > 2u\right] \\ &\leq \mathbb{P}\left[\left|\mathcal{N}^{(m)}_{\frac{2\pi(j+2u)}{n}} - \mathbb{E}\mathcal{N}^{(m)}_{\frac{2\pi(j+2u)}{n}}\right| > 2u\right].\end{aligned}$$

If $j + 2u \geq n$ then

$$\mathbb{P}\left[\theta_j > \frac{2\pi j}{n} + \frac{4\pi}{n}u\right] = \mathbb{P}\left[\theta_j > 2\pi\right] = 0,$$

and the above inequality holds trivially. The probability that $\theta_j < \frac{2\pi j}{n} - \frac{4\pi}{n}u$ is bounded in the same way. Inequality (4.6) now follows from (4.5). $\square$

We are now in a position to bound the expected distance between the empirical spectral measure of U^m and uniform measure. Let θ_j be as in Lemma 4.12. Then by Fubini's theorem,

$$\begin{aligned}
\mathbb{E}\left|\theta_j - \frac{2\pi j}{n}\right|^p &= \int_0^\infty pt^{p-1}\mathbb{P}\left[\left|\theta_j - \frac{2\pi j}{n}\right| > t\right]\, dt \\
&= \frac{(4\pi)^p p}{n^p}\int_0^\infty u^{p-1}\mathbb{P}\left[\left|\theta_j - \frac{2\pi j}{n}\right| > \frac{4\pi}{n}u\right]\, du \\
&\leq \frac{4(4\pi)^p p}{n^p}\left[\int_0^\infty u^{p-1}e^{-u^2/m[\log(n/m)+1]}\, du + \int_0^\infty u^{p-1}e^{-u}\, du\right] \\
&= \frac{4(4\pi)^p}{n^p}\left[\left(m\left[\log\left(\frac{n}{m}\right)+1\right]\right)^{p/2}\Gamma\left(\frac{p}{2}+1\right) + \Gamma(p+1)\right] \\
&\leq 8\Gamma(p+1)\left(\frac{4\pi}{n}\sqrt{m\left[\log\left(\frac{n}{m}\right)+1\right]}\right)^p.
\end{aligned}$$

Observe that in particular,

$$\operatorname{Var}\theta_j \leq C\frac{m\left[\log\left(\frac{n}{m}\right)+1\right]}{n^2}.$$

Let ν_n be the measure which puts mass $\frac{1}{n}$ at each of the points $e^{2\pi ij/n}$, $1 \leq j \leq n$. Then

$$\begin{aligned}
\mathbb{E}W_p(\mu_{n,m},\nu_n)^p &\leq \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n \left|e^{i\theta_j} - e^{2\pi ij/n}\right|^p\right] \leq \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n \left|\theta_j - \frac{2\pi j}{n}\right|^p\right] \\
&\leq 8\Gamma(p+1)\left(\frac{4\pi}{n}\sqrt{m\left[\log\left(\frac{n}{m}\right)+1\right]}\right)^p.
\end{aligned}$$

It is easy to check that $W_p(\nu_n,\nu) \leq \frac{\pi}{n}$, and thus

$$\mathbb{E}W_p(\mu_{n,m},\nu) \leq \mathbb{E}W_p(\mu_{n,m},\nu_n) + \frac{\pi}{n} \leq (\mathbb{E}W_p(\mu_{n,m},\nu_n)^p)^{\frac{1}{p}} + \frac{\pi}{n}. \tag{4.7}$$

Applying Stirling's formula to bound $\Gamma(p+1)^{\frac{1}{p}}$ completes the proof.

To prove the rest of the main theorem, namely the concentration of $W_p(\mu_{n,m},\nu)$ at its mean, the idea is essentially to show that $W_p(\mu_{n,m},\nu)$ is a Lipschitz function of U and apply concentration of measure.

The following lemma gives the necessary Lipschitz estimates for the functions to which the concentration property will be applied.

Lemma 4.13. *Let $p \geq 1$. The map $A \mapsto \mu_A$ taking an $n \times n$ normal matrix to its spectral measure is Lipschitz with constant $n^{-1/\max\{p,2\}}$ with respect to W_p. Thus if ρ is any fixed probability measure on $\mathbb{C}$, the map $A \mapsto W_p(\mu_A,\rho)$ is Lipschitz with constant $n^{-1/\max\{p,2\}}$.*

Proof. If A and B are $n \times n$ normal matrices, then the Hoffman–Wielandt inequality [**?**, Theorem VI.4.1] states that

$$\min_{\sigma \in \Sigma_n} \sum_{j=1}^{n} \left|\lambda_j(A) - \lambda_{\sigma(j)}(B)\right|^2 \leq \|A - B\|_{HS}^2 , \tag{4.8}$$

where $\lambda_1(A), \ldots, \lambda_n(A)$ and $\lambda_1(B), \ldots, \lambda_n(B)$ are the eigenvalues (with multiplicity, in any order) of A and B respectively, and Σ_n is the group of permutations on n letters. Defining couplings of μ_A and μ_B given by

$$\pi_\sigma = \frac{1}{n} \sum_{j=1}^{n} \delta_{(\lambda_j(A), \lambda_{\sigma(j)}(B))}$$

for $\sigma \in \Sigma_n$, it follows from (4.8) that

$$\begin{aligned} W_p(\mu_A, \mu_B) &\leq \min_{\sigma \in \Sigma_n} \left(\frac{1}{n} \sum_{j=1}^{n} \left|\lambda_j(A) - \lambda_{\sigma(j)}(B)\right|^p \right)^{1/p} \\ &\leq n^{-1/\max\{p,2\}} \min_{\sigma \in \Sigma_n} \left(\sum_{j=1}^{n} \left|\lambda_j(A) - \lambda_{\sigma(j)}(B)\right|^2 \right)^{1/2} \\ &\leq n^{-1/\max\{p,2\}} \|A - B\|_{HS} . \end{aligned}$$

□

Now, by Rains' Theorem, $\mu_{n,m}$ is equal in distribution to the spectral measure of a block-diagonal $n \times n$ random matrix $U_1 \oplus \cdots \oplus U_m$, where the U_j are independent and uniform in $\mathbb{U}\left(\left\lfloor \frac{n}{m} \right\rfloor\right)$ and $\mathbb{U}\left(\left\lceil \frac{n}{m} \right\rceil\right)$. Identify $\mu_{n,m}$ with this measure and define the function $F(U_1, \ldots, U_m) = W_p(\mu_{U_1 \oplus \cdots \oplus U_m}, \nu)$; the preceding discussion means that if $U_1, \ldots, U_m$ are independent and uniform in $\mathbb{U}\left(\left\lfloor \frac{n}{m} \right\rfloor\right)$ and $\mathbb{U}\left(\left\lceil \frac{n}{m} \right\rceil\right)$ as necessary, then $F(U_1, \ldots, U_m) \stackrel{d}{=} W_p(\mu_{n,m}, \nu)$.

Applying the concentration inequality in Corollary 3.15 to the function F gives that

$$\mathbb{P}\big[F(U_1, \ldots, U_m) \geq \mathbb{E}F(U_1, \ldots, U_m) + t\big] \leq e^{-nt^2/24mL^2},$$

where L is the Lipschitz constant of F, and we have used the trivial estimate $\left\lfloor \frac{n}{m} \right\rfloor \geq \frac{n}{2m}$. Inserting the estimate of $\mathbb{E}F(U_1, \ldots, U_m)$ from Equation (4.7) and the Lipschitz estimates of Lemma 4.13 completes the proof.