Lecturer: Joel A. Tropp Scribe: Kiran Vodrahalli

Contents

1 Introduction

Most of the technical work in this talk is due to Samet Oymak, not met (Google).

What is randomized dimension reduction? Let's start out with an example in numerical linear algebra. What may not have been obvious from the talk is that that lunch talk was mainstream. The basic idea is to do random dimension reduction on the input matrix you're trying to approximate, and then do further calculations. If the spectrum decays a lot, then you can get benefits from random Hadamard type methods. For most examples in machine learning, there isn't really a benefit (you need to use powering). Because you need to keep multiplying by A, you lose the benefit of Hadamard because only the first multiplication is fast. You can obtain improvements when A is sparse.

I'm going to give you another example in optimization which is kind of silly, but nevertheless it's something that people talk about. Randomized dimension reduction in optimization (people in machine learning don't like it that much: the $\mathcal{O}(1/\epsilon^2)$). Johnson-Lindenstrauss doesn't work that well. The stuff I am talking about is not Johnson-Lindenstrauss, it works better. You don't need to preserve distances.

Suppose you want to solve overdetermined least squares (this doesn't happen that much in a world where you need fancy algorithms): Tall $m \times n$ matrix A, and you want to minimize $\|Ax - b\|_{\ell_2}$ for $x \in C$. Typically, we're instead fitting models with a lot of variables. For this to make sense it needs to be massively overdetermined. An idea is to squash the problem, draw an $r \times m$ matrix Φ with $r << m$, and project the residual: Solve

$$\min \|\Phi(Ax - b)\|_{\ell_2}, x \in C$$

Then solve the compressed problem with your favorite algorithm.

Once you start adding small constraint sets, there may be some benefit.

1.1 Random Dimension Reduction

So randomized dimension reduction plays a role in contemporary computer science. The linear algebra in optimization problems are the most appealing applications of randomized dimension

reduction, since you can pick the optimization problem, etc. Here it's more of an engineering design choice. It's also in coding theory, statistical estimation, and signal processing, but here it's a bit more silly. Compressed sensing is nothing more than randomized matrix reduction for signal acquisition. People are doing this in large scale linear algebra. Nevertheless, compressed sensing and its friends are examples of randomized dimension reduction. If you believe you can design random experiments, then you can think of random experimental design as dimension reduction on the parameters in statistics. This also comes up in coding theory, where you can think of linear codes as dual to randomized dimension reduction (see APC 529 notes): By transmitting the linear image of the short vector (a long vector), you have redundancy (because the code matrix is a flat matrix).

2 Technical Setup for Dimension Reduction

You often think about the Johnson-Lindenstrauss setting where you care about distances between pairs of points. Here, we don't care about that at all. We start out with a set T that **does not contain the origin**. Then we take a matrix $\Phi : \mathbb{R}^n \to \mathbb{R}^d$, where $d << n$. Our goal when we apply Φ to T is that we want the image of T does NOT contain the origin. We get success if $0 \notin \Phi(T)$, failure otherwise. This is a weak notion, but it turns out to be relevant in all applications we've described. Because this notion doesn't care about the scale of points, we might as well forget about their scale for the moment. Thus, define $\Omega = \{t/\|t\| : t \in T\}$ (all norms in this talk are ℓ_2 norms). So our condition becomes

Definition 2.1. Success condition.

$$0 \notin \Phi(\Omega)$$

This is good enough for a lot of applications of random embeddings. It turns out that preserving distances is the wrong idea. So remember, Φ will be random and linear.

We're going to identify the range of Φ with a subspace of the ambient space. P will be the orthogonal projector with $\text{null}(P) = \text{null}(\Phi)$. Success is equivalent to the nullspace of P not intersecting Ω. So we want to pick a random embedding Φ, or a random orthogonal projector P so that we don't intersect the nullspace.

If you're going to pick a random distribution for Φ, what would you pick? You might pick Gaussian. You might also pick a uniformly random subspace of $\mathbb{R}^n$ with codimension d for the nullspace, which is actually the same thing. Note that we're parametrizing our random embedding Φ by its nullspace. It turns out that there's an exact analysis of uniformly random embedding applied to a spherically convex set. Suppose you map everything into 0 dimensions, then for sure you fail. If you pick the dimension of the space, then you certainly succeed. With a bit of work, you can see the probability increases monotonically. Attach a random normal at each step. It's not obvious that the probability doesn't go up as a straight line. What's interesting is that there's actually a phase transition in the property of successful embedding. Once the embedding dimension is big enough, the probability rises to 1 and then stays there. If you pick bigger examples, this starts looking like a step function. The statistical dimension is the location of the phase transition.

It turns out that the phenomena of having a phase transition in the uniformly random case is totally generic.

The benefits of uniformly random embeddings are: They work, and they admit precise analysis. However, they may not be implementable, they require expensive construction, may use a lot of storage, result in expensive arithmetic.

What if we use discrete, sparse, or structured random embeddings instead? Now how do we understand these things? For instance, sparse ± 1 matrices (sparse Rademacher matrix). It turns

out you see the same probability curve empirically. The success probability seems to not depend on the exact distribution. There is again a phase transition at the statistical dimension.

Remark 2.2. The universality property for eigenvectors, the span of the rows of these random matrices is NOT uniform, they concentrate (look at the Rademacher case). However, as $n \to \infty$, it is believed that it is close to uniform, but this is a big open problem.

Definition 2.3. Independent Random Embeddings.
Fix a parameter $B \geq 1$. Let $\Phi : \mathbb{R}^n \to \mathbb{R}^d$ be a random matrix with the following properties:

1. independence: statistically independent.
2. standardization: each entry has mean zero and variance one.
3. each entry X has same distribution as $-X$
4. bounded moments: Each entry satisfies $\mathbf{E}[|X|^5] \leq B$.

Examples are Gaussian, Rademacher, sparse Rademacher, Student t with 5 degrees of freedom. Symmetry is the least important property here (used only at one point in the proof).

Definition 2.4. Statistical dimension.
Let $T \subset \mathbb{R}^n$. Then the statistical dimension of T is

$$\delta(T) := \mathbf{E}[\left(\sup_{t \in T} \langle g, \frac{t}{\|t\|} \right)_+^2]$$

where $g \sim \mathcal{N}(0, I_n)$. You can think of this as the mean square width of the Gaussian times a constant, because you can normalize a Gaussian. $[]_+$ just means 0 if it's negative.

This is kind of like a central limit theorem for random matrices.
The basic properties of statistical dimension are as follows:

- If $E \subset T \subset \mathbb{R}^n$, then $0 \leq \delta(E) \leq \delta(T) \leq n$, so it's a measure of size.
- If L is a subspace, then $\delta(L) = \dim(L)$. But it's continuous, so it's a continuous extension of the dimension.
- If K is a closed convex cone, then $\delta(K^\circ) = n - \delta(K)$. This is where you need $[]_+$. The nonnegative orthant is the self-dual cone, which means its statistical dimension is half the ambient dimension.
- You can actually calculate $\delta(T)$ can be calculated very accurately for many choices of T.

This is somewhat similar to the Rademacher complexity. Here however, constants are essential. You can't relax it very much if you want to get the right answer.

Now we come to a theorem:

Theorem 2.5. *Oymak* 2015.
Let T be a compact subset of $\mathbb{R}^n$, $\Phi : \mathbb{R}^n \to \mathbb{R}^d$ be an independent random embedding with parameter B. Then

$$d \geq \delta(T) + o(n)$$

implies $0 \notin \Phi(T)$ with high probability (success). Furthermore, if the positive hull $\cup_{\alpha \geq 0} \alpha T$ is convex, then

$$d \leq \delta(T) - o(n)$$

implies $0 \in \Phi(T)$ with high probability (failure). You have a matching pair of bounds, you get successful embedding when the dimension is above the statistical dimension.

We care about this because we want to use things like sparse Rademacher from complexity standpoint; you don't want to give up constants and so on, they matter in the engineering context. So this basically says you can change out a Gaussian for sparse Rademacher.

2.1 The Compressed Sensing Problem

Let $x^\flat \in \mathbb{R}^n$ be an unknown sparse vector with s nonzero entries. Let $\Phi \in \mathbb{R}^{m\times n}$ be a random measurement matrix. Observe m random measurements $z = \Phi x^\flat$, we know Φ and z. Produce an estimate $\hat{x}$ by solving the convex program:

$$\min \|x\|_{\ell_1}$$

subject to $\Phi x = z$.

The hope is that $\hat{x} = x^\flat$.

We can change variables to $\min \|x^\flat + u\|_{\ell_1}$ subject to $\Phi u = 0$. Now we want to know when $x^\flat$ is the unique solution. That means there's no feasible perturbation which decreases the ℓ_1 norm. The success condition is that the nullspace of Φ should not intersect the ℓ_1 ball, we don't want to be able to decrease the ℓ_1 norm any further. The failure condition is that null(Φ) does intersect Ω. So understanding when the nullspace of the random matrix intersects our set is essential to compressed sensing with ℓ_1 minimization. We'vee translated this problem to our conditions. So we can now compute that statistical dimension (only depends on sparsity and ambient dimension) divided by the ambient dimension. We have an exact formula. ρ is the x-axis is the number of nonzeros / ambient dimension, and the y-axis is the normalized statistical dimension. It looks like a semicircle kind of, which is monotone increasing.

Then the universality law for random embedding predicts phase transition! This is viewable empirically.

What's new in this picture is that you can do this with sparse Rademacher matrices. This is the first explanation of why these pictures are identical, it doesn't matter what random matrices you use. The phase law is universal for everything in our family. You actually see the same picture for other things we can't explain.

The idea in the theorem is you want to take your random matrix P and you want to replace it with a Gaussian using Lindeberg. Then you want to use the Gaussian analysis to kill the problem, it doesn't work. You slice T into several pieces, then you use Lindeberg, and then you use Gaussian case and some asymptotics.

Remark 2.6. If you take a big n-dimensional distribution on the unit sphere. If you take a random matrix $n \times n$, look at top eigenvector, look at k of its entries, that this will converge to a standard Gaussian. This is open.

Certifying a solution for convex optimization is exactly the excluding the zero constraint.

3 Stability of Embeddings

There's also a question about how stable embeddings are. When you perturb the set T, do you move into the failure set easily? You might look at

Definition 3.1.

$$\sigma_{min}(\Phi; T) := \inf_{t\in T} \|\Phi t\|$$

Here we are not on the unit sphere.

Definition 3.2. d-excess width of T.

$$\mathcal{E}_d(T) := \mathbf{E}[\inf_{t\in T}(\sqrt{d}\|t\| - \langle g, t\rangle)]$$

where $g \sim \mathcal{N}(0, I_n)$.

is the important quantity to look at. If $m \leq d$, $\mathcal{E}_m(T) \leq \mathcal{E}_d(T)$.

It turns out that the stability of the random embeddings is also universal. The stability also doesn't depend on the distribution. The distance you stay away from 0 is empirically universal.

Theorem 3.3. *Stability (Oymak)* 2015.
Suppose that T is closed subset of the unit ball in $\mathbb{R}^n$, $\Phi : \mathbb{R}^n \to \mathbb{R}^d$ is an independent random embedding with parameter B, the embedding and ambient dimension satisfy $d \leq n^{6/5}$. Then

$$\sigma_{min}(\Phi; T) \geq (\mathcal{E}_d(T))_+ - o(\sqrt{d})$$

with high probability. Furthermore, if T is convex, then

$$\sigma_{min}(\Phi; T) \leq (\mathcal{E}_d(T))_+ + o(\sqrt{d})$$

This implies the success condition of the first theorem. This can be used to predict the mean square error of the Lasso for example, quality of subspace embedding, and so on. It has a lot of applications.

This has other applications:

- Signal processing
- statistical estimation
- coding theory (idealized)
- numerical analysis
- stochastic geometry
- random matrix theory
- neuroscience? The brain may perform dimension reduction... Does behavior depend on how you wire things together? One could argue that any activity performs dimension reduction behavior. It's not clear that 0 is not in the set. People look at the neural stuff is trajectories are preserved (dynamical system type of thing).

For Johnson-Lindenstrauss, you probably need subgaussinity to get $\log n$ factors. $1/\epsilon^2$ comes from Central limit theorem. But you really don't need as much preservation of distances in a lot of applications. That's why randomized linear algebra works. This is just because you don't care about distortions, you didn't collapse any point in the subspace to zero.

4 Parting Thought/Shot

Constants matter! To learn more, look at Oymak and Tropp, Universality laws for randomized dimension reduction with applications.