CPSC 536N: Randomized Algorithms 2014-15 Term 2

Lecture 7

Prof. Nick Harvey *University of British Columbia*

Dimensionality reduction is the process of mapping a high dimensional dataset to a lower dimensional space, while preserving much of the important structure. In statistics and machine learning, this often refers to the process of finding a few directions in which a high dimensional random vector has maximum variance. Principal component analysis is a standard technique for that purpose.

In this lecture, we consider a different sort of dimensionality reduction where the goal is to preserve *pairwise distances* between the data points. We present a technique, known as the ***random projection method*** or ***Johnson-Lindenstrauss method***, for solving this problem.

In the past few lectures, our main tool has been the Chernoff bound. In this lecture we will not directly use the Chernoff bound, but the main proof uses very similar ideas.

1 Dimensionality Reduction

Suppose we have n points $x_1, \ldots, x_n \in \mathbb{R}^d$. We would like to find n points $y_1, \ldots, y_n \in \mathbb{R}^t$, where $t \ll d$, such that the lengths and pairwise distances of the y vectors are approximately the same as for the x vectors. We will show that this can be accomplished while taking t to be surprisingly small. We will measure lengths using the Euclidean norm. Our notation for the length of v is $\|v\| = \sqrt{\sum_i v_i^2}$.

Theorem 1 (Johnson-Lindenstrauss 1984) *Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be arbitrary. Pick any $\epsilon = (0, 1)$. Then for some $t = O(\log(n)/\epsilon^2)$ there exist points $y_1, \ldots, y_n \in \mathbb{R}^t$ such that*

$$\begin{array}{rcccll} (1-\epsilon)\|x_j\| & \leq & \|y_j\| & \leq & (1+\epsilon)\|x_j\| & \forall j \\ (1-\epsilon)\|x_j - x_{j'}\| & \leq & \|y_j - y_{j'}\| & \leq & (1+\epsilon)\|x_j - x_{j'}\| & \forall j, j'. \end{array} \tag{1}$$

References: Dubhashi-Panconesi Section 2.5.

Remarks. The theorem is actually much stronger than we stated above. It shows that there is a *random linear map* such that *for any* $x_1, \ldots, x_n$ the condition (1) holds with probability at least $1/2n$. This linear map is *oblivious*: it does not depend on $x_1, \ldots, x_n$ at all! In fact, the linear map is just a matrix whose entries are independent Gaussians.

Whereas principal component analysis is only useful when the original data points $\{x_1, \ldots, x_n\}$ are inherently low dimensional, this theorem requires *absolutely no assumption* on the original data. Also, note that the final data points $\{y_1, \ldots, y_n\}$ have no dependence on d. The original data could live in an arbitrarily high dimension. (Although one can always assume $d \leq n$ since $x_1, \ldots, x_n$ lie in their linear span, which is a Euclidean space of dimension at most n.)

To prove the theorem, let us define the random linear map that is used to construct the points y_j. Define the function $f : \mathbb{R}^d \to \mathbb{R}^t$ by

$$f(v) \;=\; Rv,$$

where R is a $t \times d$ matrix whose entries are independently drawn from $N(0, 1)$, the normal distribution with mean 0 and variance 1. The point y_j in Theorem 1 is obtained linearly from x_j by setting $y_j \leftarrow f(x_j)/\sqrt{t} = Rx_j/\sqrt{t}$.

Lemma 2 (Distributional JL) *Let $\delta \in (0,1]$ be arbitrary. There exists a $t = O(\log(1/\delta)/\epsilon^2)$ and a random linear map $f : \mathbb{R}^d \to \mathbb{R}^t$ such that, for any vector $v \in \mathbb{R}^d$ with $\|v\| = 1$,*

$$\Pr\left[\ 1-\epsilon \ \le\ \frac{\|f(v)\|}{\sqrt{t}}\ \le\ 1+\epsilon\ \right]\ \ge\ 1-2\delta.$$

Given this lemma, our main theorem follows easily.

PROOF:[of Theorem 1] Set $\delta = 1/n^3$. Consider the set of vectors

$$W = \{\ x_i\ :\ i = 1,\ldots,n\ \}\ \cup\ \{\ x_i - x_j\ :\ i \neq j\ \}.$$

There are at most n^2 vectors in W.

For any $w \in W$, we may apply the DJL lemma to $v = w/\|w\|$. Consider the event

$$\mathcal{E}_w\ =\ \left\{\ \frac{\|f(v)\|}{\sqrt{t}} \notin [1-\epsilon, 1+\epsilon]\ \right\}$$

Since f is linear, we have $\|f(w)\| = \|\,\|w\| \cdot f(v)\,\| = \|w\| \cdot \|f(v)\|$, so

$$\mathcal{E}_w\ =\ \left\{\ \frac{\|f(w)\|}{\sqrt{t}} \notin [1-\epsilon, 1+\epsilon] \cdot \|w\|\ \right\}.$$

The DJL lemma shows that $\Pr[\,\mathcal{E}_w\,] \le 2\delta$. By a union bound,

$$\Pr[\,\text{condition (1) fails to hold}\,]\ =\ \Pr\left[\ \bigcup_{w\in W} \mathcal{E}_w\ \right]\ \le\ \sum_{w\in W} \Pr[\,\mathcal{E}_w\,]\ \le\ |W| \cdot (2\delta)\ \le\ 2/n.$$

□

1.1 Discussion

First of all, you have probably noticed that we've now jumped from the world of discrete probability to continuous probability. This is to make our lives easier. The same theorem would be true if we picked the coordinates of r_i to be uniform in $\{+1,-1\}$ rather than Gaussian. But the analysis of the $\{+1,-1\}$ case is trickier, and most proofs analyze that case by showing that its failure probability is not much worse than in the Gaussian case. So the Gaussian case is really the central problem.

Second of all, you might be wondering where the **random projection method** name comes from. Earlier versions of the Johnson-Lindenstrauss theorem used a slightly different linear map. Specifically, they chose a map $L(v) = Rv$ where $R^T R$ is a *projection* onto a *uniformly random subspace* of dimension t. (Recall that an orthogonal projection matrix is any symmetric, positive semidefinite matrix whose eigenvalues are either 0 or 1.) One advantage of that setup is its symmetry: one can argue that the failure probability in Lemma 2 would be the same if one instead chose a *fixed* subspace of dimension t and a *random* unit vector v. The latter problem can be analyzed by choosing the subspace to be the most convenient one of all: the span of the first t vectors in the standard basis.

So how is our map $f/\sqrt{t}$ different? It is almost a projection, but not quite. If we chose R to be a matrix of independent Gaussians, it turns out that the range of $R^T R$ is indeed a uniformly random subspace, but its eigenvalues are not necessarily in $\{0,1\}$. If we had insisted that the random vectors r_i that we choose were *orthonormal*, then we would have obtained a projection matrix. We could explicitly orthonormalize them by the Gram-Schmidt method, but fortunately that turns out to be unnecessary: the Johnson-Lindenstrauss theorem is true, even if we ignore orthonormality of the r_i's.

Our linear map $f/\sqrt{t}$ turns out to be a bit more convenient in some algorithmic applications, because we avoid the awkward Gram-Schmidt step.

1.2 The Main Idea

Consider the following problem: you have a vector $v \in \mathbb{R}^d$ and you want to compute $\|v\|$ by applying a *linear* function to v. Since $\|\cdot\|$ is certainly not a linear function, this seems impossible! Amazingly, randomness allows us to solve this problem. The main idea comes from the interesting algebraic properties of variance.

Fact 3 *Let $G_1, \ldots, G_d$ be independent random variables with finite variance. Let $\sigma_1, \ldots, \sigma_d \in \mathbb{R}$ be arbitrary. Then* $\mathrm{Var}\left[\sum_i \sigma_i G_i\right] = \sum_i \sigma_i^2 \mathrm{Var}\left[G_i\right]$.

References: Mitzenmacher-Upfal Corollary 3.4 and Exercise 3.4, Grimmett-Stirzaker Theorem 3.3.11

To see how this fact connects to our problem, let $G_1, \ldots, G_d$ satisfy $\mathrm{E}\left[G_i\right] = 0$ and $\mathrm{Var}\left[G_i\right] = 1$. Consider the random linear function g defined by $g(v) = \sum_i v_i G_i$. By Fact 3, we have

$$\mathrm{Var}\left[g(v)\right] \;=\; \mathrm{Var}\left[\sum_i v_i G_i\right] \;=\; \sum_i v_i^2 \;=\; \|v\|^2 .$$

So $g(v)^2$ gives a good estimate for $\|v\|^2$ because $\mathrm{E}\left[g(v)\right] = 0$, so $\mathrm{E}\left[g(v)^2\right] = \mathrm{Var}\left[g(v)\right] = \|v\|^2$.

We will use that idea in the special case where the G_i random variables are Gaussian. That turns out to be convenient because the sum of Gaussians is again Gaussian.

Fact 4 *Let $G_1, \ldots, G_d$ be independent random variables where G_i has distribution $N(0,1)$. Then, for any scalars $\sigma_1, \ldots, \sigma_d$, the sum $\sum_i \sigma_i G_i$ has distribution $N(0, \sum_{i=1}^d \sigma_i^2)$.*

References: Grimmett-Stirzaker Example 4.8.3, Durrett Theorem 2.1.13 and Exercise 3.3.4.

It is convenient to use Gaussian random variables in our context because then we know the distribution of $g(v)$ exactly: it is $N(0, \sum_i v_i^2) = N(0, \|v\|^2)$.

As explained above, the quantity $g(v)^2$ gives us a good estimate of $\|v\|^2$. Intuitively we should get an even better estimate of $\|v\|^2$ by computing t such estimates then averaging them. That is exactly what $f(v)$ does: each coordinate of $f(v)$ is of the form $g(v)$, and the average of their squares is is exactly $\|f(v)\|^2 / t$.

1.3 Proof of Lemma 2

We can use separate but similar arguments to analyze the upper and lower tails, as was the case with Chernoff bounds. We will prove only the upper tail. For convenience we square both sides, so our goal is to prove that

$$\Pr\left[\; \|f(v)\|^2 \;>\; (1+\epsilon)^2 t \;\right] \;\leq\; \delta. \tag{2}$$

Recall that $f(v) = Rv$ where the entries of R are independent Gaussians. It will be convenient for us to consider the rows of the matrix R, so let $r_i \in \mathbb{R}^d$ be the i^{th} row, for $i = 1, \ldots, t$. Define $X_i = r_i^{\mathsf{T}} v$, which is the i^{th} coordinate of $f(v)$. As discussed above, X_i is an estimate of the form $g(v)$ and it has the distribution $N(0, \|v\|^2) = N(0,1)$.

The goal in (2) is to prove an upper tail bound on $\|f(v)\|^2$. Expanding this, we have

$$\|f(v)\|^2 \;=\; \sum_{i=1}^{t} (r_i^T v)^2 \;=\; \sum_{i=1}^{t} X_i^2.$$

Fortunately, this random variable has a well-known distribution. We have just written $\|f(v)\|^2$ as the the sum-of-squares of t standard normal random variables, which is called the chi-squared distribution with parameter t. It is easy to see that

$$\mathrm{E}\left[\,\|f(v)\|^2\,\right] \;=\; \sum_{i=1}^{t} \mathrm{E}\left[\,X_i^2\,\right] \;=\; t,$$

since $\mathrm{E}\left[\,X_i^2\,\right]$ is the variance of X_i, which we have shown is 1.

So our desired inequality (2) is asking for a bound on the probability that a chi-squared random variable slightly exceeds its expectation. Claim 5 proves such a bound using a Chernoff-style approach. Applying Claim 5 to $Y = \|f(v)\|^2$ with $t = (4/3)\ln(1/\delta)/\epsilon^2$ completes the proof of (2).

Claim 5 *Let $X = \sum_{i=1}^{t} X_i^2$ have the chi-squared distribution with parameter t. Set $\alpha = t(1+\epsilon)^2$. Then* $\Pr[\,X > \alpha\,] \le \exp(-(3/4)t\epsilon^2)$.

Proof: Our proof will follow the Chernoff bound strategy. For any $\theta \in [0, 1/2)$, we have

$$\Pr[\,X > \alpha\,] \;=\; \Pr\left[\,e^{\theta X} > e^{\theta\alpha}\,\right] \;\le\; e^{-\theta\alpha}\,\mathrm{E}\left[\,e^{\theta X}\,\right]. \tag{3}$$

The quantity $\mathrm{E}\left[\,e^{\theta X}\,\right]$ is called the moment generating function, and for many standard distributions it has a known closed form. We now cheat by referring to Wikipedia, where we find that the moment generating function for the chi squared distribution is $\mathrm{E}\left[\,e^{\theta X}\,\right] = (1-2\theta)^{-t/2}$, so

$$\Pr[\,X > \alpha\,] \;\le\; e^{-\theta\alpha}(1-2\theta)^{-t/2}.$$

The next step is to plug in an appropriate choice of θ. We set $\theta = (1 - t/\alpha)/2$, giving

$$\Pr[\,X > \alpha\,] \;=\; e^{(t-\alpha)/2}(t/\alpha)^{-t/2}.$$

Plugging in $\alpha = t(1+\epsilon)^2$, this becomes

$$\exp\left(\frac{t}{2}\Big(1-(1+\epsilon)^2\Big) - \frac{t}{2}\ln\left(\frac{1}{(1+\epsilon)^2}\right)\right) \;=\; \exp\Big(-t\big(\epsilon+\epsilon^2/2-\ln(1+\epsilon)\big)\Big).$$

Using our usual techniques from Notes on Convexity Inequalities, one can show that $\ln(1+x) \le x - x^2/4$ for $x \in [0,1]$. So this shows that

$$\Pr[\,X > \alpha\,] \;\le\; \exp\Big(-t\big(\epsilon+\epsilon^2/2-(\epsilon-\epsilon^2/4)\big)\Big) \;\le\; \exp\big(-(3/4)t\epsilon^2\big).$$

□

2 Remarks

Recently there has been much progress on understanding the optimality of these results. The DJL lemma is actually optimal, up to constant factors.

Theorem 6 (Jayram-Woodruff 2013, Kane-Meka-Nelson 2011) *Any f satisfying the DJL lemma must satisfy $t = \Omega(\log(1/\delta)/\epsilon^2)$.*

But, this does not necessarily imply that Theorem 1 is optimal; perhaps the theorem can be proven without using the DJL lemma. Alon proved the following lower bound.

Theorem 7 (Alon) *Let $x_1, \ldots, x_{n+1} \in \mathbb{R}^n$ be the vertices of a simplex, i.e., $\|x_i - x_j\| = 1$ for all $i \neq j$. If $y_1, \ldots, y_{n+1} \in \mathbb{R}^t$ satisfy (1), then $t = \Omega(\frac{\log(n)}{\epsilon^2 \log(1/\epsilon)})$.*

This shows that Theorem 1 is almost optimal, up to the factor $\log(1/\epsilon)$ in the denominator. Actually, for this particular set of points (the vertices of a simplex), Theorem 1 is *not* optimal and Alon's bound is the right one. However, there is a different point set showing that Theorem 1 is in fact optimal.

Theorem 8 (Larsen-Nelson FOCS 2017) *There exist points $x_1, \ldots, x_n \in \mathbb{R}^d$ such that the following is true. Consider any map $L : \mathbb{R}^d \to \mathbb{R}^t$, let $y_j = L(x_j)$, and suppose that (1) is satisfied. Then $t = \Omega(\log(n)/\epsilon^2)$.*

Other norms and metrics. The Johnson-Lindenstrauss lemma very strongly depends on properties of the Euclidean norm. For other norms, this remarkable dimensionality reduction is not necessarily possible. For example, for the ℓ_1 norm $\|x\|_1 := \sum_i |x_i|$, it is known that any map into $\mathbb{R}^d$ that preserves pairwise ℓ_1-distances between n points up to a factor $c \geq 1$ must have $d = \Omega(n^{1/c^2})$. If $c = 1 + \epsilon$, then there are upper bounds of $d = O(n \log n/\epsilon^2)$ and $d = O(n/\epsilon^2)$.

References: Talagrand Proc. AMS 1990, Brinkman-Charikar FOCS 2003, Lee-Naor 2004, Newman-Rabinovich SODA 12.

For more on this subject, see the survey of Indyk and Matousek or the tutorial of Indyk.