Dimensionality Reduction — Notes 1

Jelani Nelson
minilek@seas.harvard.edu

August 10, 2015

1 Preliminaries

Here we collect some notation and basic lemmas used throughout this note.

Throughout, for a random variable X, $\|X\|_p$ denotes $(\mathbb{E}\,|X|^p)^{1/p}$. It is known that $\|\cdot\|_p$ is a norm for any $p \ge 1$ (Minkowski's inequality). It is also known $\|X\|_p \le \|X\|_q$ whenever $p \le q$. Henceforth, whenever we discuss $\|\cdot\|_p$, we will assume $p \ge 1$.

Lemma 1 (Khintchine inequality)**.** *For any $p \ge 1$, $x \in \mathbb{R}^n$, and (σ_i) independent Rademachers,*

$$\|\sum_i \sigma_i x_i\|_p \lesssim \sqrt{p} \cdot \|x\|_2$$

Proof. Without loss of generality we can assume p is an even integer. Consider (g_i) independent gaussians of mean zero and variance 1. Expand $\mathbb{E}(\sum_i \sigma_i x_i)^p$ into a sum of monomials. Any monomial with odd exponents vanishes, as in the gaussian case. Meanwhile other monomials are nonnegative with all Rademacher moments being 1, while in the gaussian case the moments are at least 1. Thus the Rademacher case is term-by-term dominated by the gaussian case and $\|\sum_i \sigma_i x_i\|_p \le \|\sum_i g_i x_i\|_p$. But $\sum_i g_i x_i$ is a gaussian with mean zero and variance $\|x\|_2^2$, and hence its p-norm is $\|x\|_2 \cdot (p!/(2^{p/2}(p/2)!))^{1/p}$. □

We often use Jensen's inequality below, especially for $F(x) = |x|^p$ ($p \ge 1$).

Lemma 2 (Jensen's inequality)**.** *For F convex, $F(\mathbb{E}\,X) \le \mathbb{E}\,F(X)$.*

Before proving a couple concentration inequalities, we prove a lemma now which lets us freely obtain tail bounds from moment bounds and vice versa (often we prove a moment bound and later invoke a tail bound, or vice versa, without even mentioning any justification).

Lemma 3. *Let Z be a scalar random variable. Consider the following statements:*

(1a) There exists $\sigma > 0$ s.t. $\forall p \geq 1$, $\|Z\|_p \leq C_1\sigma\sqrt{p}$.

(1b) There exists $\sigma > 0$ s.t. $\forall \lambda > 0$, $\mathbb{P}(|Z| > \lambda) \leq C_2 e^{-C_2'\lambda^2/\sigma^2}$.

(2a) There exists $K > 0$ s.t. $\forall p \geq 1$, $\|Z\|_p \leq C_3 K p$.

(2b) There exists $K > 0$ s.t. $\forall\ \lambda > 0$, $\mathbb{P}(|Z| > \lambda) \leq C_4 e^{-C_4'\lambda/K}$.

(3a) There exist $\sigma, K > 0$ s.t. $\forall\ p \geq 1$, $\|Z\|_p \leq C_5(\sigma\sqrt{p} + Kp)$.

(3b) There exist $\sigma, K > 0$ s.t. $\forall\ \lambda > 0$, $\mathbb{P}(|Z| > \lambda) \leq C_6(e^{-C_6'\lambda^2/\sigma^2} + e^{-C_6'\lambda/K})$.

Then 1a is equivalent to 1b, 2a is equivalent to 2b, and 3a is equivalent to 3b, where the constants C_i, C_i' in each case change by at most some absolute constant factor.

Proof. We will show only that 1a is equivalent to 1b; the other cases are argued identically.

To show that 1a implies 1b, by Markov's inequality

$$\mathbb{P}(Z > \lambda) \leq \lambda^{-p} \cdot \mathbb{E}\,|Z|^p \leq \left(\frac{C_1^2\sigma^2}{\lambda^2 p}\right)^{p/2}.$$

Statement 1b follows by choosing $p = \max\{1, 2C_1^2\lambda^2/\sigma^2\}$.

To show that 1b implies 1a, by integration by parts we have

$$\mathbb{E}\,|Z|^p = \int_0^\infty px^{p-1}\,\mathbb{P}(|Z| > \lambda)d\lambda \leq 2C_2 p \cdot \int_0^\infty px^{p-1} \cdot e^{-C_2'\lambda^2/\sigma^2}d\lambda.$$

The integral on the right hand side is exactly the pth moment of a gaussian random variable with mean zero and variance $\sigma'^2 = \sigma^2/(2C_2')$. Statement 1a then follows since such a gaussian has p-norm $\Theta(\sigma'\sqrt{p})$. □

Now, the following is a bread-and-butter trick for bounding pth moments of sums of independent random variables. A more general version of this lemma can be found as Lemma 6.3 in [LT91].

Lemma 4 (Symmetrization / Desymmetrization)**.** *Let $Z_1, \ldots, Z_n$ be independent random variables. Let $r_1, \ldots, r_n$ be independent Rademachers. Then*

$$\|\sum_i Z_i - \mathbb{E}\sum_i Z_i\|_p \le 2 \cdot \|\sum_i r_i Z_i\|_p \text{ \textit{(symmetrization inequality)}}$$

and

$$(1/2) \cdot \|\sum_i r_i(Z_i - \mathbb{E}\, Z_i)\|_p \le \|\sum_i Z_i\|_p \text{ \textit{(desymmetrization inequality).}}$$

Proof. For the first inequality, let $Y_1, \ldots, Y_n$ be independent of the Z_i but identically distributed to them. Then

$$\begin{aligned}
\|\sum_i Z_i - \mathbb{E}\sum_i Z_i\|_p &= \|\sum_i Z_i - \mathop{\mathbb{E}}_{Y}\sum_i Y_i\|_p \\
&\le \|\sum_i (Z_i - Y_i)\|_p \text{ (Jensen)} \\
&= \|\sum_i r_i(Z_i - Y_i)\|_p \qquad (1) \\
&\le 2 \cdot \|\sum_i r_i X_i\|_p \text{ (triangle inequality)}
\end{aligned}$$

(1) follows since the $X_i - Y_i$ are independent across i and symmetric.

For the second inequality, let Y_i be as before. Then

$$\begin{aligned}
\|\sum_i r_i(Z_i - \mathbb{E}\, Z_i)\|_p &= \|\mathop{\mathbb{E}}_{Y}\sum_i r_i(Z_i - Y_i)\|_p \\
&\le \|\sum_i r_i(Z_i - Y_i)\|_p \text{ (Jensen)} \\
&= \|\sum_i (Z_i - Y_i)\|_p \\
&\le 2 \cdot \|\sum_i Z_i\|_p \text{ (triangle inequality)}
\end{aligned}$$

□

Lemma 5 (Decoupling [dlPnG99])**.** *Let $x_1, \ldots, x_n$ be independent and mean zero, and $x'_1, \ldots, x'_n$ identically distributed as the x_i and independent of them. Then for any $(a_{i,j})$ and for all $p \geq 1$*

$$\| \sum_{i \neq j} a_{i,j} x_i x_j \|_p \leq 4 \| \sum_{i,j} a_{i,j} x_i x'_j \|_p$$

Proof. Let $\eta_1, \ldots, \eta_n$ be independent Bernoulli random variables each of expectation $1/2$. Then

$$\begin{aligned} \| \sum_{i \neq j} a_{i,j} x_i x_j \|_p &= 4 \cdot \| \mathop{\mathbb{E}}_{\eta} \sum_{i \neq j} a_{i,j} x_i x_j |\eta_i| |1 - \eta_j| \|_p \\ &\leq 4 \cdot \| \sum_{i \neq j} a_{i,j} x_i x_j \eta_i (1 - \eta_j) \|_p \text{ (Jensen)} \end{aligned} \tag{2}$$

Hence there must be some fixed vector $\eta' \in \{0, 1\}^n$ which achieves

$$\| \sum_{i \neq j} a_{i,j} x_i x_j \eta_i (1 - \eta_j) \|_p \leq \| \sum_{i \in S} \sum_{j \notin S} a_{i,j} x_i x_j \|_p$$

where $S = \{i : \eta'_i = 1\}$. Let x_S denote the $|S|$-dimensional vector corresponding to the x_i for $i \in S$. Then

$$\begin{aligned} \| \sum_{i \in S} \sum_{j \notin S} a_{i,j} x_i x_j \|_p &= \| \sum_{i \in S} \sum_{j \notin S} a_{i,j} x_i x'_j \|_p \\ &= \| \mathop{\mathbb{E}}_{x_S} \mathop{\mathbb{E}}_{x'_{\bar{S}}} \sum_{i,j} a_{i,j} x_i x'_j \|_p \text{ } (\mathbb{E}\, x_i = \mathbb{E}\, x'_j = 0) \\ &\leq \| \sum_{i,j} a_{i,j} x_i x'_j \|_p \text{ (Jensen)} \end{aligned}$$

□

The following proof of the Hanson-Wright was shared to me by Sjoerd Dirksen (personal communication).

Theorem 1 (Hanson-Wright inequality [HW71])**.** *For $\sigma_1, \ldots, \sigma_n$ independent Rademachers and $A \in \mathbb{R}^{n \times n}$ real and symmetric, for all $p \geq 1$*

$$\|\sigma^T A \sigma - \mathbb{E}\, \sigma^T A \sigma\|_p \lesssim \sqrt{p} \cdot \|A\|_F + p \cdot \|A\|.$$

Proof. Without loss of generality we assume in this proof that $p \geq 2$ (so that $p/2 \geq 1$). Then

$$\begin{align}
\|\sigma^T A\sigma - \mathbb{E}\,\sigma^T A\sigma\|_p &\lesssim \|\sigma^T A\sigma'\|_p \text{ (Lemma 5)} \tag{3}\\
&\lesssim \sqrt{p}\cdot \|\|Ax\|_2\|_p \text{ (Khintchine)} \tag{4}\\
&= \sqrt{p}\cdot \|\|Ax\|_2^2\|_{p/2}^{1/2} \tag{5}\\
&\leq \sqrt{p}\cdot \|\|Ax\|_2^2\|_p^{1/2} \notag\\
&\leq \sqrt{p}\cdot (\|A\|_F^2 + \|\|Ax\|_2^2 - \mathbb{E}\,\|Ax\|_2^2\|_p)^{1/2} \text{ (triangle inequality)} \notag\\
&\leq \sqrt{p}\cdot \|A\|_F + \sqrt{p}\cdot \|\|Ax\|_2^2 - \mathbb{E}\,\|Ax\|_2^2\|_p^{1/2} \notag\\
&\lesssim \sqrt{p}\cdot \|A\|_F + \sqrt{p}\cdot \|x^T A^T Ax'\|_p^{1/2} \text{ (Lemma 5)} \notag\\
&\lesssim \sqrt{p}\cdot \|A\|_F + p^{3/4}\cdot \|\|A^T Ax\|_2\|_p^{1/2} \text{ (Khintchine)} \notag\\
&\lesssim \sqrt{p}\cdot \|A\|_F + p^{3/4}\cdot \|A\|^{1/2}\cdot \|\|Ax\|_2\|_p^{1/2} \tag{6}
\end{align}$$

Writing $E = \|\|Ax\|_2\|_p^{1/2}$ and comparing (4) and (6), we see that for some constant $C > 0$,

$$E^2 - Cp^{1/4}\|A\|^{1/2}E - C\|A\|_F \leq 0.$$

Thus E must be smaller than the larger root of the above quadratic equation, implying our desired upper bound on E^2. □

Remark 1. The "square root trick" in the proof of the Hanson-Wright inequality above is quite handy and can be used to prove several moment inequalities (for example, you will see how to prove the Bernstein inequality with it in tomorrow's lecture). As far as I am aware, the trick was first used in a work of Rudelson [Rud99].

Remark 2. We could have upper bounded Eq. (5) by

$$\sqrt{p}\cdot \|A\|_F + \sqrt{p}\cdot \|\|Ax\|_2^2 - \mathbb{E}\,\|Ax\|_2^2\|_{p/2}^{1/2}$$

by the triangle inequality. Now notice we have bounded the pth central moment of a symmetric quadratic form (3) by the $p/2$th moment also of a symmetric quadratic form. Writing $p = 2^k$, this observation leads to a proof by induction on k, which was the approach used in [DKN10].

2 Johnson-Lindenstrauss (JL) lemma

First we prove the *Distributional JL Lemma (DJL).*

Lemma 6. *DJL Lemma For any integer $n > 1$ and $\varepsilon, \delta \in (0, 1/2)$, there exists a distribution $\mathcal{D}_{\varepsilon,\delta}$ over $\mathbb{R}^{m\times n}$ for $m \lesssim \varepsilon^{-2}\log(1/\delta))$ such that for any $x \in \mathbb{R}^n$ of unit Euclidean norm,*

$$\mathop{\mathbb{P}}_{\Pi\sim\mathcal{D}_{\varepsilon,\delta}}(|\|\Pi x\|_2^2 - 1| > \varepsilon) < \delta$$

Proof. Write $\Pi_{i,j} = \sigma_{i,j}/\sqrt{m}$, where the $\sigma_{i,j}$ are independent Rademachers. Also overload σ to mean these Rademachers arranged as a vector of length mn, by concatenating rows of Π. Note then

$$\|\Pi x\|_2^2 = \|A_x\sigma\|_2^2$$

where

$$A_x = \frac{1}{\sqrt{m}} \cdot \begin{bmatrix} -x^T- & 0 & \cdots & 0 \\ 0 & -x^T- & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & -x^T- \end{bmatrix}. \tag{7}$$

Thus

$$\mathbb{P}(|\|\Pi x\|_2^2 - 1| > \varepsilon) = \mathbb{P}(|\|A_x\sigma\|_2^2 - \mathbb{E}\,\|A_x\sigma\|_2^2| > \varepsilon),$$

where we see that the right-hand side is readily handled by the Hanson-Wright inequality Theorem 1 with $A = A_x^T A_x$. Now observe A is a block-diagonal matrix with each block equaling $(1/m)xx^T$, and thus $\|A\| = \|x\|_2^2/m = 1/m$. We also have $\|A\|_F^2 = 1/m$. Thus Hanson-Wright yields

$$\mathbb{P}(|\|\Pi x\|_2^2 - 1| > \varepsilon) \lesssim e^{-C\varepsilon^2 m} + e^{-C\varepsilon m},$$

which for $\varepsilon < 1$ is at most δ for $m \gtrsim \varepsilon^{-2}\log(1/\delta)$. □

The following is what is usually referred to as the *Johnson-Lindenstrauss (JL) lemma* [JL84]. In this note we typically refer to it as the *Metric* JL lemma (or MJL) to distinguish it from DJL above. At some points we simply say JL instead of DJL or MJL, but the version meant will be understood from context.

Corollary 1 (Metric JL lemma (MJL))**.** *For any* $X = \{x_1, \dots, x_N\} \subset \mathbb{R}^n$ *and* $0 < \varepsilon < 1/2$*, there exists* $f : X \to \mathbb{R}^m$ *for* $m = O(\varepsilon^{-2} \log N)$ *such that for all* $1 \le i < j \le N$*,*

$$(1-\varepsilon)\|x_i - x_j\|_2 \le \|f(x_i) - f(x_j)\|_2 \le (1+\varepsilon)\|x_i - x_j\|_2. \tag{8}$$

Proof. Let $\mathcal{D}_{\varepsilon,\delta}$ be as in DJL with $\delta < 1/\binom{N}{2}$. Consider a random f, where $f(x) = \Pi x$ for Π drawn from $\mathcal{D}_{\varepsilon,\delta}$. By DJL, each vector of the form $(x_i - x_j)/\|x_i - x_j\|_2$ has its norm preserved up to $1+\varepsilon$ with probability strictly larger than $1-1/\binom{N}{2}$. Thus by a union bound over all i, j, all such vectors are preserved with positive probability, showing existence of the desired f. □

2.1 Example application: k-means clustering

In the *k-means* clustering problem the input consists of $x_1, \dots, x_N \in \mathbb{R}^n$ and a positive integer k, and the goal is to output some partition $\mathcal{P}$ of $[n]$ into k disjoint subsets $P_1, \dots, P_k$ as well as some $y = (y_1, \dots, y_k) \in (\mathbb{R}^n)^k$ (the y_i need not be equal to any of the x_i and can be chosen arbitrarily) so as to minimize the cost function

$$\underset{\mathcal{P},y}{cost}(x_1, \dots, x_N) = \sum_{j=1}^{k} \sum_{i \in P_j} \|x_i - y_j\|_2^2.$$

That is, the x_i are clustered into k clusters according to $\mathcal{P}$, and the cost of a given clustering is the sum of squared Euclidean distances to the cluster centers (the y_j's).

Unfortunately finding the optimal clustering for k-means is NP-hard, however efficient approximation algorithms do exist which find a clusterings that are close to optimal.

It is easy to show, e.g. by taking the gradient of the cost function, that for a fixed partition $\mathcal{P}$ of $[n]$, the optimal choice of cluster centers y for that given $\mathcal{P}$ is the one where, for the P_j of positive size, $y_j = (1/|P_j|) \cdot \sum_{i \in P_j} x_i$. Thus we can restrict our attention to just optimizing over $\mathcal{P}$. For a set of input points X, we let $cost_{\mathcal{P}}(X)$ denote

$$\inf_{y} \underset{\mathcal{P},y}{cost}(X).$$

Lemma 7. *Let the input points to k-means be* $X = \{x_1, \dots, x_n\}$*. Then for any* $0 < \varepsilon < 1/2$*, if* $f : X \to \mathbb{R}^m$ *is such that*

$$\forall i, j \ (1-\varepsilon)\|x_i - x_j\|_2^2 \le \|f(x_i) - f(x_j)\|_2^2 \le (1+\varepsilon)\|x_i - x_j\|_2^2$$

then for $\hat{\mathcal{P}}$ a γ-approximate optimal clustering for $f(X)$ and $\mathcal{P}^*$ an optimal clustering for X, it holds that

$$\underset{\hat{\mathcal{P}}}{cost}(X) \le \gamma \cdot \left(\frac{1+\varepsilon}{1-\varepsilon}\right) \cdot \underset{\mathcal{P}^*}{cost}(X).$$

Proof. Fix a partition $\mathcal{P}$ of $[n]$ and write $\mathcal{P} = (P_1, \ldots, P_k)$. Then

$$\begin{aligned}
\underset{\mathcal{P}}{cost}(X) &= \sum_{j\in[k]} \sum_{i\in P_j} \|x_i - \frac{1}{|P_j|} \sum_{i'\in P_j} x_{i'}\|_2^2 \\
&= \sum_{j\in[k]} \frac{1}{|P_j|} \sum_{i\in P_j} \left(\sum_{i'\in P_j} \|x_i\|_2^2 - 2\langle x_i, \sum_{i'\in P_j} x_{i'}\rangle + \| \sum_{i'\in P_j} x_{i'}\|_2^2 \right) \\
&= \sum_{j\in[k]} \frac{1}{|P_j|} \sum_{i\in P_j} \sum_{i'\in P_j} \left(\frac{\|x_i\|_2^2 + \|x_{i'}\|_2^2}{2} - \langle x_i, x_{i'}\rangle \right) \\
&= \sum_{j\in[k]} \frac{1}{2|P_j|} \sum_{i\in P_j} \sum_{i'\in P_j} \|x_i - x_{i'}\|_2^2
\end{aligned}$$

Thus if f satisfies the condition of the lemma, then

$$(1-\varepsilon)\underset{\mathcal{P}}{cost}(X) \le \underset{\mathcal{P}}{cost}(f(X)) \le (1+\varepsilon)\underset{\mathcal{P}}{cost}(X)$$

for all partitions $\mathcal{P}$ simultaneously. Thus we have

$$(1-\varepsilon)\underset{\hat{\mathcal{P}}}{cost}(X) \le \underset{\hat{\mathcal{P}}}{cost}(f(X)) \le \gamma \cdot \underset{\mathcal{P}^*}{cost}(f(X)) \le \gamma \cdot (1+\varepsilon)\underset{\mathcal{P}^*}{cost}(X).$$

The lemma follows by comparing the right hand side with the left. □

References

[DKN10] Ilias Diakonikolas, Daniel M. Kane, and Jelani Nelson. Bounded independence fools degree-2 threshold functions. In *51th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 11–20, 2010.

[dlPnG99] Victor de la Peña and Evarist Giné. *Decoupling: From dependence to independence*. Probability and its Applications. Springer-Verlag, New York, 1999.

[HW71] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Statist.*, 42:1079–1083, 1971.

[JL84] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Springer-Verlag, Berlin, 1991.

[Rud99] Mark Rudelson. Random vectors in the isotropic position. *J. Functional Analysis*, 164(1):60–72, 1999.