STAT 535 Final Project
Random Projection for High-Dimensional Statistical Learning

Ming-Chun Wu
Department of Statistics
University of Washington

1 Introduction

Nowadays high-dimensional statistical learning is getting more and more important since we have more and more high-dimensional data like gene microarray. The main challenging is the so called curse of dimensionality. Briefly speaking, this says many learning algorithms are not scalable in dimension. For example, although kNN is very powerful in low-dimension, it usually fails in high-dimensional scenarios. One possible solution is to first apply dimension reduction methods then to apply learning algorithms. Or, an even more attractive way is to do dimension reduction and statistical learning in the same time. In both case, we need a effective and scalable dimension reduction method for high-dimensional data. It seems random projection is an useful tool. Therefore, the main purpose of this report is to first review the theorems of random projection and then to explore its applications and potential extensions.

2 Random Projection

Given a set of data $\mathcal{X} = \{x^1, x^2, \ldots, x^N\}$ where $x^i \in \mathbb{R}^d$ we want to find a mapping that transforms data into a low dimensional space, say $\mathbb{R}^k, k < d$. Moreover, we also want the compressed data to be as similar as the original one in some sense. Since many statistical learning methods only depend on the distance between data points, it is attractive to have a transform preserving the distance between data points. However, it is impractical that we can find such a good transform. Fortunately, if we have some tolerance of the aliasing caused by the transform, we can find a simple transform that works with high probability. Specifically, give the tolerance parameter $\epsilon \in (0, 1)$, we want to find a transform Φ such that with high probability

$$(1-\epsilon)\left\|x^i - x^j\right\|_2^2 \leq \left\|\Phi(x^i) - \Phi(x^j)\right\|_2^2 \leq (1+\epsilon)\left\|x^i - x^j\right\|_2^2, \forall x^i, x^j \in \mathcal{X} \subseteq \mathbb{R}^d \tag{1}$$

We call such a mapping *ϵ-faithful* Johnson-Lindenstrauss (JL) transform. Now the question is whether there exists such a transform. If the answer is yes, how can we find such a transform? The Johnson-Lidenstrauss (JL) lemma answer these two question.

2.1 The Johnson-Lindenstrauss Transform

Surprisingly, if we choose the projection dimension k large enough, linear random projection is good enough to achieve ϵ-faithful. If we let $\Phi(x) = \frac{1}{\sqrt{k}}Px$ and generate the projection matrix $P \in \mathbb{R}^{k\times d}$ with P_{ij}s being i.i.d. $N(0,1)$, the JL lemma tells us this transform can be ϵ-faithful. The proof of JL lemma is quite constructive, which is an example of the usefulness of concentration inequality.

Lemma 1 (JL transform)**.** *Given $\epsilon \in (0,1)$ and $k > 32\frac{\log N}{\epsilon^2}$, if we construct $P \in \mathbb{R}^{k\times d}$ such that $P_{ij} \overset{i.i.d.}{\sim} N(0,1)$ then, $\Phi(x) = \frac{1}{\sqrt{k}}Px$ is an ϵ-faithful JL transform with probability at least $1 - e^{-k\epsilon^2/16}$. That is,*

$$(1-\epsilon)\left\|x^i - x^j\right\|_2^2 \leq \frac{1}{k}\left\|Px^i - Px^j\right\|_2^2 \leq (1+\epsilon)\left\|x^i - x^j\right\|_2^2, \forall x^i, x^j \in \mathcal{X} \subseteq \mathbb{R}^d \tag{2}$$

Proof. Note that (1) can be rewrite as

$$\sup_{x^i, x^j \in \mathcal{X}} \left| \frac{1}{k}\left\| P\frac{(x^i - x^j)}{\|x^j - x^j\|_2}\right\|_2^2 - 1 \right| \leq \epsilon \tag{3}$$

and we want to show this holds with high probability. If we can bound

$$\mathbb{P}\left(\left|\frac{1}{k}\left\|P\frac{(x^i-x^j)}{\|x^j-x^j\|_2}\right\|_2^2-1\right|\geq\epsilon\right) \tag{4}$$

for each pair of x^i, x^j, then we can apply union bound to finish the proof. Moreover, it suffices to bound

$$\mathbb{P}\left(\left|\frac{1}{k}\|Pu\|_2^2-1\right|\geq\epsilon\right), \forall \|u\|_2=1 \tag{5}$$

If we let p_i^T be the ith row of P and notice that $p_i \sim N(0, I_d)$, then $\|Pu\|_2 = \sum_{i=1}^k (u^T p_i)^2$ and $Z_i := u^T p_i \sim N(0, \|u\|_2) \sim N(0,1)$. So, we need to bound

$$\mathbb{P}\left(\left|\frac{1}{k}\sum_{i=1}^k Z_i^2-1\right|\geq\epsilon\right)=\mathbb{P}\left(\left|\frac{1}{k}\sum_{i=1}^k (Z_i^2-\mathbb{E}Z_i^2)\right|\geq\epsilon\right) \tag{6}$$

where $Z_i^2 \overset{i.i.d.}{\sim} \mathcal{X}_1^2$ and $\mathbb{E}Z_i^2 = 1$. This is the standard form to apply Chernoff bound. That is, if we can bound the MGF of $\mathcal{X}_1^2$ then we can bound the tail probability. Now, apply the chi-square concentration Lemma 2, we get

$$\mathbb{P}\left(\left|\frac{1}{k}\sum_{i=1}^k (Z_i^2-\mathbb{E}Z_i^2)\right|\geq\epsilon\right)\leq 2e^{-k\epsilon^2/8} \tag{7}$$

By union bound, we can prove that when $|\mathcal{X}| = N$

$$\mathbb{P}\left(\sup_{x^i,x^j\in\mathcal{X}}\left|\frac{1}{k}\left\|P\frac{(x^i-x^j)}{\|x^j-x^j\|_2}\right\|_2^2-1\right|\geq\epsilon\right)\leq\sum_{ij}\mathbb{P}\left(\left|\frac{1}{k}\left\|P\frac{(x^i-x^j)}{\|x^j-x^j\|_2}\right\|_2^2-1\right|\geq\epsilon\right)\leq\binom{|\mathcal{X}|}{2}2e^{-k\epsilon^2/8} \tag{8}$$

$$\leq N^2 e^{-k\epsilon^2/8}=\exp\left(2\log N-k\epsilon^2/8\right) \tag{9}$$

$$\leq\exp\left(-k\epsilon^2/16\right), k>32\frac{logN}{\epsilon^2} \tag{10}$$

□

Lemma 2 (Chi-square concentration)**.** *If* $Z_i^2 \overset{i.i.d.}{\sim} \mathcal{X}_1^2$*, then* $\mathbb{P}\left(\left|\frac{1}{K}\sum_{i=1}^k Z_i^2-1\right|\geq\epsilon\right)\leq 2e^{-k\epsilon^2/8}$

Proof.

$$\mathbb{E}e^{\lambda(Z_i^2-1)}=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{\lambda(z^2-1)}e^{-z^2/2}dz \tag{11}$$

$$\overset{t=z^2}{=}\frac{e^{-\lambda}}{\sqrt{1-2\lambda}}\int_0^\infty\frac{1}{\Gamma(1/2)(2/(1-2\lambda))^{1/2}}t^{1/2-1}e^{-(1/2-\lambda)t}dt \tag{12}$$

$$=\frac{e^{-\lambda}}{\sqrt{1-2\lambda}}, \frac{1}{2}-\lambda\geq 0 \qquad \text{pdf of Gamma}\left(\frac{1}{2},\frac{2}{1-2\lambda}\right) \tag{13}$$

$$\leq e^{2\lambda^2}, \lambda<\frac{1}{4} \tag{14}$$

Then apply Chernoff bound

$$\mathbb{P}\left(\frac{1}{k}\sum_{i=1}^k (Z_i^2-\mathbb{E}Z_i^2)\geq\epsilon\right)\leq\frac{\mathbb{E}e^{\lambda\sum_{i=1}^k(Z_i^2-1)}}{e^{\lambda k\epsilon}}\leq e^{2k\lambda^2-k\epsilon\lambda}\overset{\lambda=\epsilon/4}{\leq} e^{-k\epsilon^2/8} \tag{15}$$

□

Now, we have proved the JL lemma for the Gaussian random matrix case. The key step is to bound (5). If we can do so, then the JL lemma holds. In fact we can show that if p_{ij} are i.i.d. sub-Ggaussian, the JL lemma still holds.

Corollary 1 (JL for sub-Gaussian)**.** *If p_{ij} are i.i.d. sub-Gaussian with sub-Gaussian constant* 1*, then Lemma 1 still holds.*

Note that WLOG we let the sub-Gaussian constant equal to 1. Otherwise, one can scale the projection matrix with a constant to make it equal to 1.

Proof. Firstly, we can show that $p_i^T u$, the weighted sum of i.i.d. sub-Gaussian random variables is still sub-Gaussian,

$$\mathbb{E}e^{tp_i^T u} = \Pi_j \mathbb{E}e^{tp_{ij}u_j} \leq \Pi_j e^{u_j^2 t^2/2} \qquad \text{sub-Gaussian} \tag{16}$$

$$= e^{\|u\|_2^2 t^2/2} = e^{t^2/2} \qquad \|u\|_2 = 1 \tag{17}$$

Use the fact that if $Z_i = p_i^T u$ is sub-Gaussian, then Z_i^2 is sub-exponential [7], then we apply the tail bound for sub-exponential random variable [8](similar to Lemma 2)

$$\mathbb{P}\left(\frac{1}{k}\sum_{i=1}^{k}(Z_i^2 - \mathbb{E}Z_i^2) \geq \epsilon\right) \leq e^{-Ck\epsilon^2} \tag{18}$$

for some constant C. Now, we need one more step, since $\mathbb{E}Z_i^2$ is not necessary equal to 1 (this is true if Z_i is $N(0,1)$). It is not difficult to show that $\mathbb{E}Z_i^2 \leq 1$ (use the result of HW2 problem 3(d)), then

$$\mathbb{P}\left(\frac{1}{k}\sum_{i=1}^{k}Z_i^2 - 1 \geq \epsilon\right) \leq \mathbb{P}\left(\frac{1}{k}\sum_{i=1}^{k}(Z_i^2 - \mathbb{E}Z_i^2) \geq \epsilon\right) \leq e^{-Ck\epsilon^2} \tag{19}$$

By symmetry, we can get

$$\mathbb{P}\left(\left|\frac{1}{k}\sum_{i=1}^{k}Z_i^2 - 1\right| \geq \epsilon\right) \leq 2e^{-Ck\epsilon^2} \tag{20}$$

and then finish the proof by following the same reasoning of the proof in Lemma 1. □

2.2 General Result via Gaussian Complexity

In the JL lemma, we aim to bound the extreme value statistic

$$\mathcal{Z}(\mathcal{K}) = \sup_{u\in\mathcal{K}}\left|\frac{1}{k}\sum_{i=1}^{k}(p_i^T u)^2 - 1\right| \tag{21}$$

for the special case when $\mathcal{K}$ is a finite subset of the unit sphere in $\mathbb{R}^d$. If we use the Gaussian random projection matrix, we can extend such that $\mathcal{K}$ (can be uncountable) is any subset of the unit sphere. For finite sets, we can use union bound, but it is not applicable in general case. Since we want to prove a uniform property over the set $\mathcal{K}$, the scenario is similar to the proof of Glivenko-Cantelli (GC) theorem. Recall from the lectures, we also want to bound the extreme value statistic

$$\sup_{g\in\mathcal{G}}|\mathbb{E}_n g - \mathbb{E}g| = \sup_{g\in\mathcal{G}}\left|\frac{1}{N}\sum_{i=1}^{N}g(X_i) - \mathbb{E}g(X)\right| \tag{22}$$

in the proof of GC theorem. Since the scenarios are quite similar, to generalize JL lemma we may follow the structure of the proof of GC theorem. Briefly speaking, the proof of GC theorem is as follows

- Step 1: Show that $\sup_{g\in\mathcal{G}}|\mathbb{E}_n g - \mathbb{E}g|$ is concentrates around its mean (use McDiarmid inequality)
- Step 2: Approximate $\mathbb{E}\sup_{g\in\mathcal{G}}|\mathbb{E}_n g - \mathbb{E}g|$ by *Rademacher complexity*
- Step 3: Use Massart lemma to bound the Rademacher complexity

We now introduce the Gaussian complexity, which is the counterpart of Rademacher complexity.

Definition 1 (Gaussian complexity)**.** *The Gaussian complexity, or Gaussian width, of a set $\mathcal{K}$ is defined as*

$$\mathcal{W}(\mathcal{K}) = \mathbb{E}[\sup_{u\in\mathcal{K}} u^T g] \tag{23}$$

where g follows $N(0, I)$.

Note that if g is the Rademacher sequence, $\mathcal{W}$ is the Rademacher complexity. Now, we want to show that

Theorem 1 (Gaussian random projection)**.** *For all $\epsilon \in (0, 1/2)$ and $\mathcal{K}$ is a subset of the unit sphere in $\mathbb{R}^d$. If the projection dimension k is large enough such that $\frac{\mathcal{W}(\mathcal{K})}{\sqrt{k}} \leq 1$, we have*

$$\mathbb{P}\left(\mathcal{Z}(\mathcal{K}) \geq 4\left(\frac{\mathcal{W}(\mathcal{K})}{\sqrt{k}} + \epsilon\right)\right) \leq 2e^{-k\epsilon^2/2} \tag{24}$$

where $\mathcal{Z}(\mathcal{K}) = \sup_{u\in\mathcal{K}} \left|\frac{1}{k}\sum_{i=1}^k (p_i^T u)^2 - 1\right|$.

Before proving this theorem, let us try to understand it. We can consider this theorem as the counterparts of the first two steps in the proof of GC theorem since it relates the extreme value statistic to the Gaussian complexity (an average). Furthermore, let us try to use this general result to get Lemma 1.

In Lemma 1, we have $\{x^1, x^2, \ldots, x^N\}$, and thus $\binom{N}{2}$ pairs of x^i, x^j. Since $u = \frac{x^i - x^j}{\|x^i - x^j\|_2}$, there are $\binom{N}{2}$ possible values of u. Then $\mathcal{W} = \mathbb{E}\max_u u^T g = \mathbb{E}\max_{i=1,\ldots\binom{N}{2}} Z_i$ where $Z_i = u_i^T g$ are i.i.d. $N(0, 1)$ (recall $\|u_i\|_2 = 1$ and g is $N(0, I)$). To bound the expectation of the maximum of $\binom{N}{2}$ i.i.d. $N(0, 1)$, we can apply the Massart lemma (the result of HW2 problem 5)

$$\mathcal{W} \leq \sqrt{2\log\binom{N}{2}} \leq \sqrt{4\log N} \tag{25}$$

Therefore,

$$\mathbb{P}\left(\mathcal{Z}(\mathcal{K}) \geq 4\left(\sqrt{\frac{4\log N}{k}} + \epsilon\right)\right) \leq 2e^{-k\epsilon^2/2} \tag{26}$$

Choose $k = C\frac{\log N}{\epsilon^2}$ for some constant C to have the result of Lemma 1.

2.3 Proof of Theorem 1

Now, we start to prove Theorem 1. Let us first outline the proof. The first step is try to relate $\mathcal{Z}(\mathcal{K}) = \sup_{u\in\mathcal{K}} \left|\frac{1}{k}\sum_{i=1}^k (p_i^T u)^2 - 1\right|$ to the Gaussian complexity $\mathcal{W}$. We can do this by showing that with high probability, say $1 - 2e^{-k\epsilon^2/2}$, the following relation holds

$$-\frac{\mathcal{W}(\mathcal{K})}{\sqrt{k}} - \epsilon \leq \frac{\|Pu\|_2}{\sqrt{k}} - 1 \leq \frac{\mathcal{W}(\mathcal{K})}{\sqrt{k}} + \epsilon, \forall u \in \mathcal{K} \tag{27}$$

The reason is as follows. If (27) holds, we have

$$0 \leq \frac{\|Pu\|_2}{\sqrt{k}} + 1 \leq \frac{\mathcal{W}(\mathcal{K})}{\sqrt{k}} + \epsilon + 2 \tag{28}$$

$$\left|\frac{\|Pu\|_2}{\sqrt{k}} - 1\right| \leq \frac{\mathcal{W}(\mathcal{K})}{\sqrt{k}} + \epsilon \tag{29}$$

Therefore,

$$\mathcal{Z}(\mathcal{K}) = \sup_{u\in\mathcal{K}} \left| \frac{1}{k}\sum_{i=1}^{k}(p_i^T u)^2 - 1 \right| = \sup_{u\in\mathcal{K}} \left| \left(\frac{\|Pu\|_2}{\sqrt{k}} + 1\right)\left(\frac{\|Pu\|_2}{\sqrt{k}} - 1\right) \right| \tag{30}$$

$$\leq \sup_{u\in\mathcal{K}} \left| \frac{\|Pu\|_2}{\sqrt{k}} + 1 \right| \sup_{u\in\mathcal{K}} \left| \frac{\|Pu\|_2}{\sqrt{k}} - 1 \right| \tag{31}$$

$$\leq \left(\frac{\mathcal{W}(\mathcal{K})}{\sqrt{k}} + \epsilon + 2\right)\left(\frac{\mathcal{W}(\mathcal{K})}{\sqrt{k}} + \epsilon\right) \tag{32}$$

$$\leq 4\left(\frac{\mathcal{W}(\mathcal{K})}{\sqrt{k}} + \epsilon\right) \qquad \epsilon, \frac{\mathcal{W}(\mathcal{K})}{\sqrt{k}} \leq 1 \tag{33}$$

The remaining is to prove (27) but only upper is proved in this report. The idea to prove the lower bound is similar to the idea used to prove upper bound while we need the more involved Gordon inequality [3].

2.3.1 Proof of the upper bound (27)

Now, our goal is to show $\sup_{u\in\mathcal{K}} \frac{\|Pu\|_2}{\sqrt{k}} \leq 1 + \frac{\mathcal{W}(\mathcal{K})}{\sqrt{k}}$. The idea is similar to step 1 and 2 in the proof of GC theorem. We first show that the extreme value statistic $\sup_{u\in\mathcal{K}} \frac{\|Pu\|_2}{\sqrt{k}}$ concentrates around its mean, then show that its mean can be upper bounded by the Gaussian complexity $\frac{\mathcal{W}(\mathcal{K})}{\sqrt{k}} + 1$. We need the following concentration inequality.

Theorem 2 (Gaussian dimension-free concentration)**.** *If $X \sim N(0, I)$ is a random vector in $\mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz, then we have the concentration inequality*

$$\mathbb{P}\left(g(X) - \mathbb{E}g(X) \geq t\right) \leq e^{-t^2/2L^2}, \forall t > 0 \tag{34}$$

Proof. See [9] or lecture notes. □

In this proof, the Gaussian dimension-free concentration is the counterpart of McDiarmid inequality. To use this concentration, we need to check that $f(P) = \|Pu\|_2 \sqrt{k}$ is L-Lipschitz. Let u^* be the maximizer such that $\frac{\|Pu^*\|_2}{\sqrt{k}} = \sup_{u\in\mathcal{K}} \frac{\|Pu\|_2}{\sqrt{k}}$

$$f(P) - f(P') = \frac{\|Pu^*\|_2}{\sqrt{k}} - \sup_{u\in\mathcal{K}} \frac{\|P'u\|_2}{\sqrt{k}} \leq \frac{\|Pu^*\|_2}{\sqrt{k}} - \frac{\|P'u^*\|_2}{\sqrt{k}} \leq \frac{\|(P-P')u^*\|_2}{\sqrt{k}} \tag{35}$$

$$\leq \frac{\|P - P'\|_{op} \|u^*\|_2}{\sqrt{k}} \tag{36}$$

$$\leq \frac{\|P - P'\|_F}{\sqrt{k}} \tag{37}$$

We get (36) by the definition of operator norm $\|P\|_{op} = \sup_{\|x\|=1} \frac{\|Px\|_2}{\|x\|_2}$. And (37) is by the fact that operator norm is smaller than Frobenius norm. Similarly for $f(P') - f(P)$ we can show that f is $\frac{1}{\sqrt{k}}$-Lipschitz. Then by the dimension-free concentration

$$\mathbb{P}\left(\sup_{u\in\mathcal{K}} \frac{\|Pu\|_2}{\sqrt{k}} \geq \mathbb{E}\sup_{u\in\mathcal{K}} \frac{\|Pu\|_2}{\sqrt{k}} + \epsilon\right) \leq e^{-k\epsilon^2/2} \tag{38}$$

Now, we need show that $\mathbb{E}\sup_{u\in\mathcal{K}} \frac{\|Pu\|_2}{\sqrt{k}}$ is upper bounded by $\frac{\mathcal{W}(\mathcal{K})}{\sqrt{k}} + 1$. To do this, we need the following Gaussian comparison inequality

Theorem 3 (Sudakov-Fernique)**.** *Give a index set T and consider two zero-mean Gaussian processes $\{X_t\}, \{Y_t\}, t \in T$, if $\mathbb{E}(X_t - X_{t'})^2 \leq \mathbb{E}(Y_t - Y_{t'})^2, \forall t, t' \in T$, then $\mathbb{E}[\sup_{t\in T} X_t] \leq \mathbb{E}[\sup_{t\in T} Y_t]$*

Proof. This is a special case of Gordon's theorem see [3]. □

To apply the Gaussian comparison inequality, we need to first use the following trick

$$\sup_{u\in\mathcal{K}}\frac{\|Su\|_2}{\sqrt{k}}=\sup_{u\in\mathcal{K}}\sup_{\|v\|_2=1}\frac{v^TPu}{\sqrt{k}}=\sup X_{u,v} \tag{39}$$

since $\sup_{\|v\|_2=1} v^Ty=\frac{y^T}{\|y\|_2}y=\|y\|_2$. Define $X_{u,v}=\frac{v^TPu}{\sqrt{k}}\sim N(0,1/k)$ then $\{X_{u,v}\}$ is a double indexed Gaussian process with index set $\mathcal{K}\times S$ where S is the unit sphere in $\mathbb{R}^k$. Now we need to find another Gaussian process $Y_{u,v}$ on $\mathcal{K}\times S$ such that we can bound the expectation of our extreme value statistic

$$\mathbb{E}\sup_{u\in\mathcal{K}}\frac{\|Su\|_2}{\sqrt{k}}=\mathbb{E}\sup X_{u,v}\le\mathbb{E}\sup Y_{u,v} \tag{40}$$

It turns out that $Y_{u,v}=\frac{v^Th}{\sqrt{k}}+\frac{u^Tg}{\sqrt{k}}$ where $h\sim N(0,I_k), g\sim N(0,I_d)$ may be a good choice since

$$\mathbb{E}\sup Y_{u,v}=\mathbb{E}\sup_{\|v\|_2=1}\frac{v^Th}{\sqrt{k}}+\mathbb{E}\sup_{u\in\mathcal{K}}\frac{u^Tg}{\sqrt{k}}\le 1+\frac{\mathcal{W}(\mathcal{K})}{\sqrt{k}} \tag{41}$$

In the last inequality, we use

$$\mathbb{E}\sup_{\|v\|_2=1}\frac{v^Th}{\sqrt{k}}=\mathbb{E}\frac{\|h\|_2}{\sqrt{k}}=\mathbb{E}\sqrt{\frac{\|h\|_2^2}{k}}\overset{Jensen}{\le}\sqrt{\mathbb{E}\,\|h\|_2^2/k}=\sqrt{\mathbb{E}\mathcal{X}_k^2/k}=1 \tag{42}$$

Now, we need to check to condition of the Gaussian comparison inequality

$$\mathbb{E}\,(X_{u,v}-X_{u',v'})^2=\frac{1}{k}\mathbb{E}\left(tr(v^TPu-\tilde{v}^TP\tilde{u})\right)^2=\frac{1}{k}\mathbb{E}\left(P^T(vu^T-\tilde{v}\tilde{u}^T)\right)^2 \tag{43}$$

$$=\frac{1}{k}\mathbb{E}\left(P^TA\right)^2=\frac{1}{k}\mathbb{E}(\sum_i(P^TA)_{ii})^2 \qquad A=vu^T-\tilde{v}\tilde{u}^T \tag{44}$$

$$=\frac{1}{k}\sum_i\mathbb{E}(S^TA)_{ii}^2 \qquad (S^TA)_{ii}\text{'s are independent} \tag{45}$$

$$=\frac{1}{k}\sum_i\mathbb{E}(\sum_j P_{ij}A_{ji})^2 \qquad P_{ij}\overset{i.i.d.}{\sim}N(0,1) \tag{46}$$

$$=\frac{1}{k}\sum_{ij}A_{ij}^2\mathbb{E}P_{ij}^2=\frac{1}{k}\sum A_{ij}^2=\frac{1}{k}\left\|vu^T-\tilde{v}\tilde{u}^T\right\|_F^2 \tag{47}$$

$$\mathbb{E}\,(Y_{u,v}-Y_{u',v'})^2=\frac{1}{k}\left(\|v-\tilde{v}\|_2^2-\|u-\tilde{u}\|_2^2\right) \qquad \text{similar to } X \tag{48}$$

Moreover,

$$\left\|vu^T-\tilde{v}\tilde{u}^T\right\|_F^2=2-tr(\tilde{u}^Tu\tilde{v}^Tv)=2(1-(\tilde{u}^Tu)(\tilde{v}^Tv)) \tag{49}$$

$$\|v-\tilde{v}\|_2^2-\|u-\tilde{u}\|_2^2=4-2(v^T\tilde{v}+u^T\tilde{u}) \tag{50}$$

$$\|v-\tilde{v}\|_2^2-\|u-\tilde{u}\|_2^2-\left\|vu^T-\tilde{v}\tilde{u}^T\right\|_F^2=2(a-1)(b-1)\ge 0 \qquad a=\tilde{v}^Tv\le 1, b=\tilde{u}^Tu\le 1 \tag{51}$$

and thus we prove

$$\mathbb{E}\sup_{u\in\mathcal{K}}\frac{\|Su\|_2}{\sqrt{k}}\le 1+\frac{\mathcal{W}(\mathcal{K})}{\sqrt{k}} \tag{52}$$

Along with (38) we have with probability at least $1-e^{-k\epsilon^2/2}$,

$$\frac{\|Pu\|_2}{\sqrt{k}}-1\le\frac{\mathcal{W}(\mathcal{K})}{\sqrt{k}}+\epsilon,\forall u\in\mathcal{K} \tag{53}$$

We can get similar result for lower bound using stronger inequality [3], and with probability at least $1-e^{-k\epsilon^2/2}$

$$-\frac{\mathcal{W}(\mathcal{K})}{\sqrt{k}}-\epsilon\le\frac{\|Pu\|_2}{\sqrt{k}}-1,\forall u\in\mathcal{K} \tag{54}$$

Combine them together to finish the proof.

So far, we have two main results of random projection. The first one is when $\mathcal{K}$ is finite set, then the JL lemma holds for sub-Gaussian random projection matrices. The proof is neat and is developed using sub-exponential probability tail bound and union bound. When $\mathcal{K}$ is uncountable, we can extend the result like what we do in the GC theorem. We first develop concentration of the extreme value statistic $\sup_{u\in\mathcal{K}} \frac{\|Pu\|_2}{\sqrt{k}}$ around its mean, then bound its mean using Gaussian complexity. The result says that the required projection dimension to have ϵ-faithful JL transform with high probability is $k \geq C\frac{\mathcal{W}(K)}{\epsilon^2}$ for some constant C.

3 The Fast Johnson-Lindenstrauss Transform

In the previous section, we have introduced random projection and its theoretical guarantee as a dimension reduction method. In many applications, random projection is coupled with other statistical method. In such scenarios, although random projection reduces the dimension of data, it does not necessary speed up the computation. The reason is that all entries of the projection matrix are nonzero, and doing computation with big nonsparse matrix is usually computational burdensome. Therefore, one may try to sparsify the projection matrix [1, 2]. Then, we naturally arise concerns like how to construct such a sparse projection matrix and whether such transformation is can satisfy the JL lemma. That is, to find a sparse random projection such that with high probability, it is a ϵ-faithful JL transform. One sparse projection matrix is defined as follows.

Definition 2 (FJLT [2])**.** *The Fast Johnson-Lidenstrauss Transform (FJLT) is a mapping $\Phi : \mathbb{R}^d \to \mathbb{R}^k$ with $\Phi(x) = PHDx$ and*

- *P is a $k \times d$ spares random projection matrix with $P_{ij} = b_{ij}r_{ij}$ and b_{ij}s are binomial with successful probability q and r_{ij}s are i.i.d. $N(0, \frac{1}{q})$*
- *H is a $d \times d$ Walsh-Hammard matrix with $H_{ij} = d^{-1/2}(-1)^{\langle i-1, j-1\rangle}$. The $\langle a, b\rangle$ is the inner product of the binary representations of a, b. For example $\langle 2, 3\rangle = \langle [1, 0], [1, 1]\rangle = 1$.*
- *D is a $d \times d$ diagonal matrix with $D_{ii} = -1, 1$ with equal probability.*

Theorem 4 (FJLT lemma)**.** *For $\epsilon \in (0, 1)$ and let $q = \min\left(\frac{\log^2 N}{d}, 1\right)$, the FJLT defined in Definition 2 satisfies, with non-vanishing probability*

$$(1-\epsilon)\|x\|_2 \leq \frac{1}{k}\|\Phi x\|_2 \leq (1+\epsilon)\|x\|_2, \forall x \in \mathcal{X} = \{x^1, x^2, \ldots, x^N\} \tag{55}$$

when $k = c\frac{\log N}{\epsilon^2}$.

Proof. (sketch) The original proof can be found in [2]. Instead of proving it thoroughly, let us try to get the idea from the sketch of the proof. Our goal is to show that, with non vanishing probability

$$\left|\frac{1}{k}\left\|PHD\frac{x}{\|x\|_2}\right\|_2 - 1\right| \leq \epsilon, \forall x \in \mathcal{X} \tag{56}$$

If for each $x \in \mathcal{X}$ we can bound

$$\mathbb{P}\left(\left|\frac{1}{k}\left\|PHD\frac{x}{\|x\|_2}\right\|_2 - 1\right| \geq \epsilon\right) \tag{57}$$

, then we can apply union bound to finish the proof. WLOG, we can proceed with the assumption that $\|x\|_2 = 1$.

Step 1. Show that HDx behaves well with high probability. If HDx is too sparse, we will lose too much information when applying the sparse projection matrix P. So, we want to show that HDx is not sparse with high probability. Specifically, we can show that $\max_{x\in\mathcal{X}} \|HDx\|_\infty = O(\sqrt{\frac{\log N}{d}})$. Intuitively, we can try to bound each entry of HDx. Let $u_i = (HDx)_i$, by the definition of H, D we have $u_i = \sum_j a_j x_j$ where a_j are i.i.d. random

variables only taking values of $-d^{-1/2}$ or $d^{-1/2}$ with probability 1/2. Then the MGF of du_i is

$$\mathbb{E}e^{s(du_i)} = \Pi_j \mathbb{E}e^{sda_j x_j} = \Pi_j \frac{e^{s\sqrt{d}x_j} + e^{-s\sqrt{d}x_j}}{2} \tag{58}$$

$$\leq \Pi_j \left(1 + \frac{s^2 d x_j^2}{2}\right) \qquad \text{Taylor expansion} \tag{59}$$

$$\leq \Pi_j \left(e^{s^2 d x_j^2/2}\right) \qquad 1 + x \leq e^x \tag{60}$$

$$= e^{s^2 d/2} \qquad \|x_i\|_2 = 1 \tag{61}$$

This says du_i is sub-Gaussian with sub-Gaussian constant d. Apply sub-Gaussian tail bound (see lecture notes) we have $\mathbb{P}(|u_i| \geq s) = \mathbb{P}(|du_i| \geq ds) \leq 2e^{-s^2 d/2}$. Use union bound

$$\mathbb{P}\left(\max_x \|HDx\|_\infty \geq s\right) \leq \sum_{i=1}^{N}\sum_{j=1}^{d} \mathbb{P}(|(HDx_i)_j| \geq s) \leq 2Nde^{-s^2 d/2} \tag{62}$$

$$\leq \frac{1}{20}, \quad s = \sqrt{\frac{C\log N}{d}}, \text{ for some constant } C \tag{63}$$

This says with high probability, the entries HDx are bounded. Moreover, one can argument that the transform HD preserves signal energy, i.e. $\|x\|_2 = \|HDx\|_2$ [2]. Then the entries of $u = HDx$ seem to have similar magnitudes.

Step 2 Construct the concentration of $\|P(HDx)\|_2$. Fix $x \in \mathcal{X}$, let $y = P(HDx)$ then $y_i = [Pu]_i = \sum_{j=1}^{d} r_{ij} b_{ij} u_j \sim N(0, q^{-1}\sum_j b_{ij} u_j^2)$. Let $z_i = \sum_j b_{ij} u_j^2$ we know that given z_i, $y_i \sim N(0, q^{-1} z_i)$ and $\frac{q}{z_i} \sim X_1^2$. Apply the chi-square concentration inequality Lemma 2 we can show that $\frac{1}{k}\|y\|_2 = \frac{1}{k}\|PHDx\|_2 \approx \frac{\sum z_i}{kq}$ with high probability given z_is. Then with union bound we can show that given z_is, $\frac{1}{k}\|PHDx\|_2 \approx \frac{\sum z_i}{kq}, \forall x \in \mathcal{X}$ with high probability.

Step 3 Show that $\frac{\sum z_i}{kq} \approx 1$ with high probability when HDx is well behaved (well behaved means $\|HDx\|_\infty \leq \frac{1}{m^2}$ for some constant m). As shown in step 1, it is likely that HDx satisfies this condition $\|HDx\|_\infty \leq \frac{1}{m^2}$. In step 2 we show $\frac{1}{k}\|PHDx\|_2 \approx \frac{\sum z_i}{kq}, \forall x \in \mathcal{X}$. So if $\frac{\sum z_i}{kq} \approx 1$ we have $\frac{1}{k}\|PHDx\|_2 \approx 1$, which is what FJLT lemma says. Specifically, we want to show that with high probability

$$\left|\frac{\sum z_i}{kq} - 1\right| \leq \epsilon \tag{64}$$

Again, this is standard concentration inequality type argument, so we may try to upper bound the MGF then apply Chernoff inequality. The details of how to bound the MGF can be found in [2]. □

4 Applications and Extensions

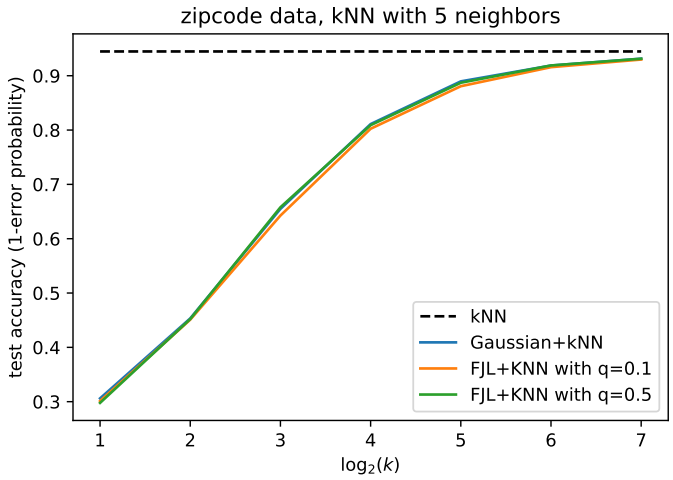
After reviewing the theory side of random projection, let us explore some applications or related topics.

4.1 kNN

We first do a little experiment on the `zipcode` dataset, which can be downloaded from
`https://web.stanford.edu/ hastie/ElemStatLearn/data.html`
The dataset is consist of images of handwritten digits automatically scanned from envelopes by the U.S. Postal Service. Each image is normalize and processed to be a 16 x 16 grayscale image. The dataset is split into a 7091 training set and a 2007 testing data. Our goal is to do classification (recognize digits from images). Classification algorithms are trained based on the training data and their performances are assessed based on the testing accuracy (1-misclassification error). The benchmark in our experiment is the kNN algorithm applied on the original data. In the other methods, kNN is applied on the projected low-dimensional data where the projection matrices are either Gaussian or FJL as described in the previous sections. The result is summarized in Figure 1.

It seems random projection is quite useful for this example. If do a 50% compression, that is let $k = 2^7 = 128$ (original dimension is $16 \times 16 = 256$), the degradation of performance is neglectable. When $k = 128$, all the random projection based algorithms lead to about 0.93 accuracy while kNN is about 0.945. This suggests the usefulness

Figure 1: Random projected kNN



of random projection for some high-dimensional problem. The other interesting observation is that all the three random projected kNN have almost the same performance. Recall that q is defined such that each entry of P only has probability q to be nonzero. So we may just use small $q = 0.1$ in this application and this speeds up the computation $P(HDx)$. This observation suggests that if the data is good enough, a very sparse transform $q = 0.1$ is good enough for accuracy.

4.2 Ensemble learning

Beyond first applying random projection then statistical learning, we may be able to develop good algorithms that do random projection and training at the same time. To get the idea, let us recall random forest, which is one of the most successful classification method. In random forest, we repeatedly draw bootstrap sample from the dataset, then train a decision tree (base classifier) for each bootstrap sample. When training the tree, only part of the the features (dimensions) are randomly selected to decide the best split, and this is a sort of random projection (although a little bit arbitrary). At the end, we combined all the trees together to output the final classifier. Follow this idea, we may combine base classifier, random projection and bootstrap together. One recent work can be found in [6].

4.3 Compressed sensing

On the theory side, the idea of JL transform is extended to many topics. From my personal viewpoint, the most successful one is compressed sensing, a series of seminal works proposed by brilliant statisticians [5, 4]. Roughly speaking, compressed sensing shows that we can exactly recover sparse signals after compressed them using random matrices (random projection) when the sensing matrices are good enough. More precisely, the sensing matrices are required to have *Restricted Isometric Property* (RIP) and many important theory has been developed under the RIP assumption.

Definition 3 (RIP)**.** *We call a vector x s-sparse if it has at most s nonzero entries. A matrix Φ has the s-restricted isometry property if there exist $\epsilon_s \in (0, 1)$ such that*

$$(1 - \epsilon_s) \|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \epsilon_s) \|x\|_2^2, \text{ for all } s\text{-sparse } x \tag{65}$$

And the smallest ϵ_s is defined as the restricted isometry constant.

As we can see, the criteria of being a good matrix in compressed sensing is very similar to that in JL lemma. Therefore, the successful compressed sensing can be considered as an extension of JL transform.

5 Concluding Remarks

The JL lemma itself can be used for dimension reduction and be used to deal with the curse of dimensionality in high-dimensional statistical learning. An intuitive approach, first random projection then classification, may be useful as shown in the `zipcode` example in the last section. Moreover, we can also design learning algorithms that implement training and random projection at the same time. The idea is suggested by the JL transform. If we give up some accuracy, we can possibly get a scalable (in dimension) approach that is good enough for most of the time. This opens a break point to deal with the curse of dimensionality. More precisely, we may give up some statistical accuracy (not to find the minimizer of loss function), but have a solvable and scalable optimization problem for high-dimensional statistical learning (one more example is the sketched least square problem [10]). This idea is very interesting for modern applications, since it bridges the gap between statistics and computation (optimization) and tends to offer a framework of making trade-off between statistical accuracy and computation complexity.

Besides the idea behind the JL lemma and its implications, the proofs shown in this report are quite constructive. These proofs are good applications of concentration inequalities and offer a good opportunity to practice what we have learned in the lectures.

References

[1] D. Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, June 2003.

[2] N. Ailon and B. Chazelle. The fast johnson-lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, May 2009.

[3] A. S. Bandeira. Johnson-lindenstrauss lemma and gordons theorem.

[4] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Inf. Theor.*, 51(12):4203–4215, Dec. 2005.

[5] E. J. Cands. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589 – 592, 2008.

[6] T. I. Cannings and R. J. Samworth. Random-projection ensemble classification.

[7] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices.

[8] M. Wainwright. Basic tail and concentration bounds.

[9] M. Wainwright. Nachdiplom lecture 2.

[10] M. Wainwright. Nachdiplom lecture 3.