JOHNSON-LINDENSTRAUSS TRANSFORMATION AND RANDOM PROJECTION

LONG CHEN

ABSTRACT. We give a brief survey of Johnson-Lindenstrauss lemma.

CONTENTS

1. INTRODUCTION

In big data era, fast algorithms are needed to manipulate massive data with large number of features or dimensions. One example is the word vector model. A document can be covert to a vector. The length of the vector is the number of (English) words which is around $25,000$. The ith component is the count of the ith word. By computing the angle or distance of corresponding vectors, we can then analyze a large collection of documents. A convenient way is to form a $d \times n$ matrix A with each column represents a document. Then $A^T A$ will give the pairwise inner products of these n-vectors.

As $n, d \gg 1$, the computation of $A^T A$ is time consuming. A fundamental tool to speed up the computation is the random projection to a k-dimensional subspace such that the pairwise distance is preserved within certain error. The remarkable fact is that the dimension of the lower dimensional space depends only on the logarithmic of the size of the data, i.e., $k = \mathcal{O}(\ln n)$. Then the pairwise distance or relative ordering can be computed efficiently. The dimension reduction from d to k is a great advantage since complexity of many algorithms grows exponentially as a function of d which is frequently termed as the "curse of dimensionality". The storage of n column vectors is also reduced from dn to $kn = \mathcal{O}(n \log n)$. The mathematical theory behind this dimension reduction is the following Johnson-Lindenstrauss Lemma [9].

Theorem 1.1 (Johnson-Lindenstrauss Lemma). *For any $0 < \epsilon < 1$ and any integer $n > 1$, let k be a positive integer such that $k \geq k_0$ with $k_0 = C\epsilon^{-2} \ln n$, where C is a suitable*

Date: December 23, 2015.

constant. Then for any set V of n points in $\mathbb{R}^d$, there exists a map $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $u, v \in V$,

$$(1-\epsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\epsilon)\|u-v\|^2.$$

Constant C in the lower bound k_0 is of practical importance. It is not specified in Johnson and Lindenstrauss' paper [9] (not surprising for pure mathematicians). In the simplified analysis give by Frankl and Maehara [6], $C \approx 9$, and in a further simplified proof by Dasgupta and Gupta [5], $C \approx 8$. In practice, $C = 2$ is good enough; see computational results in [16]. The gap between the theoretic bound and practical one could be due to the union bound used in the proof.

One of such mapping is the projection to a random subspace of dimension k. Algorithmically the projection is realized by the multiplication of a random matrix T of size $k \times d$ to the left of A, i.e., TA, which will be called a JL transformation (JLT). (Draw a picture on the product of matrices.) A related result is the following random projection theorem.

Theorem 1.2 (Random Projection)**.** *For any $0 < \epsilon, \delta < 1/2$ and positive integer d, there exists a random matrix T of size $k \times d$ such that for $k \geq k_0$ with $k_0 = C\epsilon^{-2}\ln(1/\delta)$ and for any unit-length vector $x \in \mathbb{R}^d$,*

$$\Pr\left\{\left|\|Tx\|^2 - 1\right| > \epsilon\right\} < \delta. \tag{1}$$

The probability estimate (1) can be easily derived from the exponential tail bound

$$\Pr\left\{\left|\|Tx\|^2 - 1\right| > \epsilon\right\} < e^{-Ck\epsilon^2}. \tag{2}$$

Several proofs will be developed by passing the estimate for one random variable to summation of independently identical distributed (i.i.d) random variables.

Remark 1.3. In the power of the exponential function in (2), the dependence of k is desirable which implies $k = \mathcal{O}(\ln n)$ while ϵ^2 leads to the factor ϵ^{-2} in the lower bound k_0. When $n = 10^3 \sim 10^6$, $\ln n = 7 \sim 14$. For $\epsilon = 0.25 \sim 0.1$, the factor $\epsilon^{-2} = 16 \sim 100$. So ϵ cannot be too small. A factor $C = 2$ is further multiplied to yield $k = 224 \sim 2800$.

It is interesting to note that k, however, is independent of d. The projection is beneficial only if $k \ll d$. □

One can derive J-L lemma from the random projection theorem as follows. Choosing $\delta = 1/n^2$ and, for one pair $u, v \in V$, setting $x = (u-v)/\|u-v\|$, we obtain

$$\Pr\left\{\left|\|T(u-v)\|^2/\|u-v\|^2 - 1\right| > \epsilon\right\} < \frac{1}{n^2}.$$

Using a union bound over all $\binom{n}{2} = n(n-1)/2$ pairs, we obtain, for all $u, v \in V$,

$$\Pr\left\{\left|\|T(u-v)\|^2/\|u-v\|^2 - 1\right| > \epsilon\right\} < 1/2.$$

Here we use the union bound

$$\Pr\left\{\bigcup_i E_i\right\} \leq \sum_i \Pr\left\{E_i\right\}$$

and the equality holds if events E_i are mutually independent.

Define $f(x) = Tx$. We then obtain the probabilisitic version of J-L lemma

$$\Pr\left\{(1-\epsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\epsilon)\|u-v\|^2\right\} > 1/2.$$

Since the probability is strictly greater than zero, the existence of such a map is proved. Repeating the projection p times can boost the success probability to $1 - 1/2^p$.

Remark 1.4. The random projection is oblivious to V, i.e., it can be applied to the whole space $\mathbb{R}^d$ not depending on the data set V. If *a prior* information of V is known, one may find a better mapping adapt to V.

The mapping f is defined as a linear mapping. Again in some cases, one may find a nonlinear mapping to do a better job, e.g., when the data set V is near a low-dimensional manifold [14, 15]. □

A statistic interpretation of the transform by Achlioptas [1] is as follows. Recall that a data is represented by one column, say x, of the matrix A. Each row of T acts as an estimator of the length of x by computing the inner product. To reach consensus we take an average (squared sum) of k estimators. By Central Limit Theorem, if we have sufficiently many estimators (i.e. as $k \to \infty$), we can get an arbitrarily good estimate of $\|x\|$. But using the randomness of estimator, a Gaussian type concentration will be reached for a remarkably small samples $k = \mathcal{O}(\ln n)$.

The transformed data TA is a $k \times n$ matrix and can be thought as a low rank approximation of A. A related dimension reduction technique is using A_k, the space spanned by the singular vectors corresponding to the k-largest singular values of A. Then A_k is the best rank k approximation of A for any rotationally invariant norm of matrices. So in the average sense, the information of A is preserved as much as possible in A_k. This optimality, however, implies no guarantees regarding local properties such as pairwise distance (or in general local geometry). In addition, the computation of SVD is not cheap. In contrast, JLT can be used to speed up the computation of SVD.

In practice, it is important to generate and evaluate the random matrix T fast. We end the introduction with a discussion on several choices of T.

(1) Orthonormal vectors. In the original proof of Johnson and Lindenstrauss [9], the rows of T are chosen as a random k-tupe of orthonormal vectors in $\mathbb{R}^d$ (with a proper scaling $\sqrt{d/k}$). Then T is the projection to the row space of T. From statistical point of view, requiring the k-vectors to be orthonormal has the pleasant statistical overtone of "maximizing mutual information". The resulting matrix is dense and orthogonalization process is needed.

(2) Gaussian vectors. Indyk and Motwani [8] and Dasgupta and Gupta [5] choose the entries of T as independent random variables with the standard normal (Gaussian) distribution. The orthogonalization and normlization of row vectors is thus dropped. The matrix is easier and faster to generate by skipping the orthogonalization procedure. The relaxation of orthogonality is due to the large dimension. As $d \gg 1$, random unit vectors are mutually orthogonal with high probability. The relaxation of normalization is a consequence of the exponential concentration bound. But the matrix is still dense.

(3) Sub-Gaussian vectors. Achlioptas [1] showed the Gaussian distribution can be relaxed to a much simpler ± 1 random distribution with zero mean and variance one. Matoušek [13] generalizes such results to sub-Gaussian vectors; see Theorem 2.16 for details. A computationally efficient choice will be: $T_{ij} = 0$ with probability $2/3$ and ± 1 with probability $1/6$ each (with a proper scaling so that the variance is one). The matrix is sparser and requires only summation and subtraction operations (no multiplication since the scaling can be post-ponded). Therefore the computation is much faster. But the number of non-zero is still in the order of $\mathcal{O}(kd)$.

(4) Fast Johnson-Lindenstrauss Transform. Ailon and Chazelle [2] propose a discrete Fourier transform to enlarge the support of sparse vectors. To prevent the sparsification of dense vectors, the Fourier transform is randomized. Then a sparse T can be used. They call the resulting transform as Fast-Johnson-Lindenstrauss-Transform (FJLT). The computational cost is reduced to $\mathcal{O}(d\ln d + k\ln^2 n)$.

2. JL Transform

We will consider probability proofs for the random projection theorem. We first follow Dasgupta and Gupta [5] to give a simple proof for random projections using Gaussian i.i.d variables and then Matoušek [13] for random projections using sub-Gaussian entries.

We shall write $X \sim Y$ to denote that the random variable X is distributed as random variable Y. We use $N(0,1)$ to denote the standard normal (Gaussian) distribution with mean 0 and variance 1.

2.1. An Elementary Proof.

We first flip the randomness. The length of a unit vector u in $\mathbb{R}^d$ when projected to a random k-dimensional subspace has the same distribution as the length of a random unit vector projected down onto a fixed k-dimensional subspace, say, the space spanned by the first k coordinate vectors.

This fact can be mathematically justified as follows. Let $SO(d)$ be the group of all rotation of $\mathbb{R}^d$ equipped with a measure σ. A random projection can be realized as $Tu = U^T P_k U u$ where $U \in SO(d)$ is chosen randomly w.r.t. σ and P_k is the projection matrix to the first k coordinates. Since U is unitary and $P_k^2 = P_k$, we have $\|Tu\|^2 = \|P_k U u\|^2 = \|P_k x\|^2$ and the vector $x = Uu \in S^{d-1}$ is a random unit vector. The randomness in T is then switched to the randomness in x.

How to realize *a random unit vector* mathematically? That is how to generate the random variable with uniform density on the sphere S^{d-1}. We can achieve this through the normalization of a random variable which is centrally symmetric (i.e. rotation invariant) around the origin. One convenient way is through Gaussian distribution. Let $X_1, \dots, X_d$ be i.i.d. $N(0,1)$ random variables, and let $X = (X_1, \dots, X_d)$. Since X_i are independent, the probability density function of X is $(2\pi)^{-d/2} e^{-|\boldsymbol{x}|^2/2}$, i.e., $X \sim N^d(0,1)$ which is rotation invariant. The normalized vector $Y = X/\|X\|$ is then a random unit vector.

Exercise 2.1. Consider generation of points uniformly at random on the sphere. Show that independently generate each coordinate uniformly at random from the interval $[-1,1]$ and project to the sphere will not give uniformly distributed points on sphere S^{d-1}.

Exercise 2.2. Show that if $X \sim N^d(0,1)$ and u is a unit vector, then $X \cdot u \sim N(0,1)$.

Let $Z \in \mathbb{R}^k$ be the projection of Y onto its first k-coordinates, i.e. $Z = P_k Y$. As $X_i \sim N(0,1)$ and independent, $\mathbb{E}\left[\sum_{i=1}^d X_i^2\right] = \sum_{i=1}^d \mathbb{E}[X_i^2] = d$. Consequently $\mathbb{E}[\|Z\|^2] = k/d$. Indeed it is highly concentrated around the mean. A Chernoff tail bound can be derived using the moment generating function.

Let X be a random variable with density function $f(x)$. The moment generating function $M_X(t)$ of X is defined for an open interval of t containing 0

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \int_{-\infty}^{\infty} e^{tx} f(x)\,\mathrm{d}x.$$

It is named after the fact that

$$M_X^{(n)}(0) = \mathbb{E}[X^n] \quad n \geq 1.$$

We will skip the subscript X, i.e., use $M(t)$ only, when it is clear from the context.

Exercise 2.3. Let $X \sim N(0,1)$ be the standard normal distribution. Show that

(1) $M_X(t) = e^{t^2/2}$ for all $t \in \mathbb{R}$;
(2) $M_{X^2}(t) = (1-2t)^{-1/2}$ for $-\infty < t < 1/2$.

Exercise 2.4. Prove the following inequalities

(1) $1 + x \le e^x$ for all $x \in \mathbb{R}$;
(2) $\ln(1-x) \le -x - x^2/2$ for all $x \in [0,1]$;
(3) $\ln(1+x) \le x - x^2/2 + x^3/3$ for all $x \ge 0$;
(4) $e^x \le 1 + 2x$ for all $x \in [0,1]$;
(5) $e^x \le 1 + x + x^2$ for all $x \le 1$;
(6) $(e^x + e^{-x})/2 \le e^{x^2/2}$ for all $x \in \mathbb{R}$.

Lemma 2.5 (Dasgupta and Gupta [5]). *Let $k < d$ and let $X_i \sim N(0,1)$, for $i = 1, \ldots, d$, be i.i.d standard normal random varibles.*
a) If $\beta < 1$, then

$$\Pr\left\{d\sum_{i=1}^{k} X_i^2 \le k\beta \sum_{i=1}^{d} X_i^2\right\} \le \beta^{k/2}\left(1 + \frac{k(1-\beta)}{d-k}\right)^{(d-k)/2} \le e^{k(1-\beta+\ln\beta)/2}.$$

b) If $\beta > 1$, then

$$\Pr\left\{d\sum_{i=1}^{k} X_i^2 \ge k\beta \sum_{i=1}^{d} X_i^2\right\} \le \beta^{k/2}\left(1 + \frac{k(1-\beta)}{d-k}\right)^{(d-k)/2} \le e^{k(1-\beta+\ln\beta)/2}.$$

Proof. For part a), for any $t > 0$,

$$\begin{aligned}
\Pr\left\{d\sum_{i=1}^{k} X_i^2 \le k\beta \sum_{i=1}^{d} X_i^2\right\} &= \Pr\left\{k\beta\sum_{i=1}^{d} X_i^2 - d\sum_{i=1}^{k} X_i^2 \ge 0\right\} \\
&= \Pr\left\{e^{t\left(k\beta\sum_{i=1}^{d} X_i^2 - d\sum_{i=1}^{k} X_i^2\right)} \ge 1\right\} \\
&\le \mathbb{E}\left[e^{t\left(k\beta\sum_{i=1}^{d} X_i^2 - d\sum_{i=1}^{k} X_i^2\right)}\right] \\
&= \mathbb{E}\left[e^{tk\beta X_1^2}\right]^{(d-k)} \mathbb{E}\left[e^{t(k\beta-d)X_1^2}\right]^{k} \\
&= (1-2tk\beta)^{-(d-k)/2}\left[1 - 2t(k\beta - d)\right]^{-k/2}.
\end{aligned}$$

Define

$$g(t) = (1-2tk\beta)^{-(d-k)/2}\left[1-2t(k\beta-d)\right]^{-k/2}.$$

It gives us two more constraints that $tk\beta < \frac{1}{2}$ and $t(k\beta - d) < \frac{1}{2}$. We may combine these two constraints together to get $0 < t < 1/2k\beta$. In order to minimize $g(t)$, it is equivalent to maximize

$$f(t) = (1-2tk\beta)^{(d-k)}\left[1-2t(k\beta-d)\right]^{k}$$

in the interval $0 < t < 1/2k\beta$ which is achieved at

$$t^* = \frac{1-\beta}{2\beta(d-k\beta)}.$$

Therefore, we have

$$g(t) \ge g(t^*) = \frac{1}{\sqrt{f(t^*)}} = \beta^{k/2}\left(\frac{d-k\beta}{d-k}\right)^{(d-k)/2} = \beta^{k/2}\left[1 + \frac{(1-\beta)k}{d-k}\right]^{(d-k)/2}.$$

For part b), the proof is almost identical. □

The estimate can be rewritten as

$$\Pr\left\{d/k\|Z\|^2 \leq \beta\right\} \leq e^{k(1-\beta+\ln\beta)/2}.$$

If we define the projection as $\sqrt{d/k}\,Z$ and chose $\beta = 1-\epsilon$, Lemma 2.4 and inequality (2) in Exercise 2.4 implies the desired tail bound: for a unit vector x and $k \geq 8\epsilon^{-2}\ln n$,

$$\Pr\left\{\|Tx\|^2 \leq 1-\epsilon\right\} \leq e^{-k\epsilon^2/4} \leq e^{-2\ln n} = 1/n^2.$$

Similarly choosing $\beta = 1+\epsilon$ and using the estimate in part b) and (3) in Exercise 2.4, we have

$$\Pr\left\{\|Tx\|^2 \geq 1+\epsilon\right\} \leq e^{k[1-(1+\epsilon)+\ln(1+\epsilon)]/2} \leq e^{-k(\epsilon^2/2-\epsilon^3/3)/2} \leq e^{-2\ln n} = 1/n^2$$

if $k \geq k_0 = 4(\epsilon^2/2-\epsilon^3/3)^{-1}\ln n$.

We can thus prove the random projection theorem and consequently the JL lemma.

2.2. **T-bound and M-bound.** The technique used in the above proof is known as Chernoff tail bound:

$$\Pr\{X \geq a\} \leq e^{-ta}M(t) \quad \text{for all } t > 0, \tag{3}$$

$$\Pr\{X \leq a\} \leq e^{-ta}M(t) \quad \text{for all } t < 0. \tag{4}$$

Proof of Chernoff tail bound is straightforward. For example, to prove (3), we use monotonicity of the exponential function and Markov inequality to get, for $t > 0$,

$$\Pr\{X \geq a\} = \Pr\left\{e^{tX} \geq e^{ta}\right\} \leq \mathbb{E}[e^{tX}]e^{-ta} = M(t)e^{-ta}.$$

We obtain the best bound by minimizing the upper bound $e^{-ta}M(t)$.

Example 2.6 (Chernoff bounds for the standard normal distribution)**.** Let $X \sim N(0,1)$ be the standard normal distribution. Then $M(t) = e^{t^2/2}$. So the Chernoff bound is given by

$$\Pr\{X \geq a\} \leq e^{-ta}e^{t^2/2} \quad \text{for all } t > 0.$$

The minimum is achieved at $t = a$ which gives the exponential decay tail bound

$$\Pr\{X \geq a\} \leq e^{-a^2/2} \quad \text{for all } a > 0. \tag{5}$$

Similarly, we get the tail bound in the left

$$\Pr\{X \leq a\} \leq e^{-a^2/2} \quad \text{for all } a < 0. \tag{6}$$

To generalize to random variables other than Gaussian, let us introduce the following concepts. The presentation mainly follows Matoušek [13].

Definition 2.7. *Let X be a random variable with zero mean $\mathbb{E}[X] = 0$. X has a* sub-Gaussian upper tail *if there exists a constant $C > 0$ such that for all $a > 0$,*

$$\Pr\{X \geq a\} \leq e^{-Ca^2}. \tag{7}$$

X has a sub-Gaussian upper tail up to a_0 if the bound (7) holds for all $a \leq a_0$. X has a sub-Gaussian tail if both X and $-X$ have sub-Gaussian upper tail.

Definition 2.8. *A random variable X is called* sub-Gaussian *up to t_0 if there exists a constant C such that for all $t \leq t_0$*

$$M_X(t) \leq e^{Ct^2} \tag{8}$$

We will simply call (7) "T-bound" and (8) "M-bound".

M-bound $\longrightarrow$ *T-bound.* By checking the proof of Example 2.6, we see that if we have a bound on the moment generating function $M(t) \leq e^{Ct^2}$, then we could have a sub-Gaussian upper tail bound.

We summarize as the following lemma.

Lemma 2.9. *For a random variable X with $\mathbb{E}[X] = 0$, if X is sub-Gaussian, i.e., $M_X(t) \leq e^{Ct^2/2}$ for some constant C and for all $t > 0$, then X has a sub-Gaussian upper tail $\Pr\{X \geq a\} \leq e^{-a^2/(2C)}$. If $M_X(t) \leq e^{Ct^2/2}$ holds for all $t \in (0, t_0)$, then X has a sub-Gaussian upper tail up to Ct_0.*

T-bound $\longrightarrow$ *M-bound.* We have kind of converse of the above result.

Lemma 2.10 (Matoušek [13])**.** *If $\mathbb{E}[X] = 0, \mathrm{Var}[X] = \mathbb{E}[X^2] = 1$ and X has a sub-Gaussian upper tail, i.e. $\Pr\{X \geq a\} \leq e^{-C_a a^2}$, then X is sub-Gaussian, i.e., $M(t) \leq e^{Ct^2}$ for all $t > 0$, where the constant C depends only on C_a in the sub-Gaussian tail.*

Proof. Let $F(x) = \Pr\{X < x\}$ be the distribution function of X. Then we have

$$\mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx}\, dF(x) = \int_{-\infty}^{1/t} e^{tx}\, dF(x) + \int_{1/t}^{\infty} e^{tx}\, dF(x)$$

For the first term, we have the following inequalities:

$$\begin{aligned}\int_{-\infty}^{1/t} e^{tx}\, dF(x) \leqslant \int_{-\infty}^{1/t} (1 + tx + t^2x^2)\, dF(x) &\leqslant \int_{-\infty}^{\infty} (1 + tx + t^2x^2)\, dF(x) \\ = 1 + t\mathbb{E}[X] + t^2\mathbb{E}[X^2] = 1 + t^2&\end{aligned}$$

where in the first step we use the fact that $e^x \leqslant 1 + x + x^2$ when $x \leqslant 1$.

For the second term, we can rewrite the integral by the sum:

$$\begin{aligned}\int_{1/t}^{\infty} e^{tx}\, dF(x) = \sum_{k=1}^{\infty} \int_{k/t}^{k+1/t} e^{tx}\, dF(x) &\leqslant \sum_{k=1}^{\infty} e^{t\frac{k+1}{t}} \Pr\left\{X \geqslant \frac{k}{t}\right\} \\ &\leqslant \sum_{k=1}^{\infty} e^{2k} e^{-C_a k^2/t^2} = \sum_{k=1}^{\infty} e^{k(2 - C_a k/t^2)}.\end{aligned}$$

Now consider the quadratic function $x(2 - xC_a/t^2)$. When k is in the negative part of the above quadratic function. The summation is a geometrical decay series. More precisely, if $t \leqslant \sqrt{C_a}/2$, then we have $2 - C_a k/t^2 \leqslant -C_a/2t^2$ and thus

$$\begin{aligned}\sum_{k=1}^{\infty} e^{k(2 - C_a k/t^2)} &\leqslant \sum_{k=1}^{\infty} e^{-kC_a/t^2} = \frac{e^{-C_a/t^2}}{1 - e^{-C_a/t^2}} \\ &\leqslant \frac{e^{-C_a/t^2}}{1 - e^{-4}} \leqslant 2e^{-C_a/t^2} \\ &\leqslant \frac{2t^2}{C_a} = O(t^2)\end{aligned}$$

the last line is due to the fact that $e^{-x} \leqslant 1/x$ for all $x > 0$. Hence

$$\mathbb{E}[e^{tx}] \leqslant 1 + O(t^2) \leqslant e^{O(t^2)}.$$

If $t > \sqrt{C_a}/2$, the largest terms in the sum are those k near t^2/a, and the sum is still $O(e^{t^2/2C_a})$. So we also get $\mathbb{E}[e^{tx}] \leqslant e^{O(t^2)}$. □

Exercise 2.11. Let X be the random variable taking values ± 1 with probability $1/2$. Prove the M-bound directly

$$\mathbb{E}\left[e^{tX}\right] \le e^{t^2/2}.$$

2.3. **Sum of independent random variables.** Consider n-independent random variables. We want to pass properties of individual random variable to the sum. A sequence of random variables $X_1, X_2, \dots, X_n$ have a *uniform sub-Gaussian tail* if all of them have sub-Gaussian tails with the same constant.

The target is a tail bound for the sum of X_i. But when passing the properties of each random variable to the summation, working on the M-bound is much easier. So we first apply "T-bound to M-bound" procedure to each X_i and then "M-bound to T-bound"; see the diagram below

$$\begin{tikzcd}
\text{T-bound} \arrow[r] \arrow[d, "\Sigma"'] & \text{M-bound} \arrow[d, "\Sigma"'] \\
\text{T-bound} & \text{M-bound} \arrow[l]
\end{tikzcd}$$

Lemma 2.12. *Let $X_1, \dots, X_n$ be independent random variables satisfying $\mathbb{E}[X_i] = 0$, $\operatorname{Var}(X_i) = 1$, and having a uniform sub-Gaussian tail. Let $\alpha_1, \dots, \alpha_n$ be real coefficients satisfying $\sum_{i=1}^n \alpha_i^2 = 1$. Then the sum $Y = \sum_{i=1}^n \alpha_i X_i$ has $\mathbb{E}[Y] = 0, \operatorname{Var}(Y) = 1$, and a sub-Guassian tail.*

Proof. By the linearity of expectation, we get $\mathbb{E}[Y] = 0$. For independent random variables, the variance is additive and thus $\operatorname{Var}(Y) = \sum_{i=1}^n \alpha_i^2 \operatorname{Var}(X_i) = \sum_{i=1}^n \alpha_i^2 = 1$.

By Lemma 2.10, $M_{X_i}(t) \le e^{Ct^2}$ for all $t > 0$. Then

$$M_Y(t) = \mathbb{E}\left[e^{tY}\right] = \mathbb{E}\left[\prod_{i=1}^n e^{t\alpha_i X_i}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{t\alpha_i X_i}\right] \le e^{Ct^2 \sum_{i=1}^n \alpha_i^2} = e^{Ct^2}.$$

Therefore by Lemma 2.9, the M-bound of Y implies the desired tail bound. □

Exercise 2.13 (The 2-stability of Gaussian distribution)**.** Let $X_1, \dots, X_n$ be i.i.d $N(0,1)$. Let $\alpha_1, \dots, \alpha_n$ be real coefficients satisfying $\sum_{i=1}^n \alpha_i^2 = 1$. Prove that $Y = \sum_{i=1}^n \alpha_i X_i$ is still the standard normal distribution, i.e., $Y \sim N(0,1)$.

Let T be a random matrix with each entry R_{ij} satisfying $\mathbb{E}[R_{ij}] = 0$, $\operatorname{Var}(R_{ij}) = 1$, and having a uniform sub-Gaussian tail. Then for a unit vector x, by Lemma 2.12, $Y_i := (Tx)_i = \sum_{j=1}^d R_{ij} x_j$ has properties $\mathbb{E}[Y_i] = 0, \operatorname{Var}(Y_i) = 1$, for $i = 1, 2, \dots, k$ and has a sub-Guassian tail. Next we consider the squared length of $(Y_1, Y_2, \cdots, Y_k)$.

Lemma 2.14 (Matoušek [13])**.** *Let $k \ge 1$ be an integer. Let $Y_1, \dots, Y_k$ be independent random variables with $\mathbb{E}[Y_i] = 0$, $\operatorname{Var}(Y_i) = 1$, and having a uniform sub-Gaussian tail. Then*

$$Z = \frac{1}{\sqrt{k}}(Y_1^2 + Y_2^2 + \cdots + Y_k^2 - k)$$

has a sub-Gaussian tail up to $\sqrt{k}$.

Proof. By the linearity of expectation, we have

$$\mathbb{E}\left[\sum_{i=1}^k Y_i^2\right] = \sum_{i=1}^k \mathbb{E}\left[Y_i^2\right] = \sum_{i=1}^k \operatorname{Var}(Y_i) = k.$$

The scaling $1/\sqrt{k}$ is introduced to normalize the variance of $\sum_{i=1}^k Y_i^2$ as that in Central Limit Theorem.

We calculate the moment generating function of Z first:

$$\mathbb{E}\left[e^{tZ}\right] = \mathbb{E}\left[\exp\left\{\frac{t}{\sqrt{k}}\sum_{i=1}^{k}\left(Y_i^2 - 1\right)\right\}\right] = \prod_{i=1}^{k}\mathbb{E}\left[\exp\left\{\frac{t}{\sqrt{k}}(Y_i^2 - 1)\right\}\right].$$

So we need a M-bound of $Y_i^2 - 1$ derived from the tail bound of Y_i^2.

Following the proof of Lemma 2.10, we can show that: There exist constant $C \geq 1/2$ and t_0 such that for all $t \in [0, t_0]$, we have

$$\mathbb{E}\left[e^{t(Y_i^2-1)}\right] \leq e^{Ct^2} \quad \text{and} \quad \mathbb{E}\left[e^{t(1-Y_i^2)}\right] \leq e^{Ct^2}. \tag{9}$$

Then for $t \in (0, t_0\sqrt{k})$, we obtain the M-bound of Z

$$\mathbb{E}\left[e^{tZ}\right] \leq \left(e^{Ct^2/k}\right)^k = e^{Ct^2},$$

and a sub-Gaussian upper tail up to $2C\sqrt{k} \geq \sqrt{k}$ follows from Lemma 2.9. □

Exercise 2.15. Let Y be a random variable like Y_i in Lemma 2.14.

(1) Show that $\mathbb{E}[Y^4]$ is finite and thus $\mathrm{Var}(Y^2)$ is finite;
(2) Prove M-bound (9) for $Y^2 - 1$.

2.4. **JL Transform by sub-Gaussian entries.** The following theorem shows the random projection matrix can be composed by entries with i.i.d. sub-Gaussian random variables [10, 1, 13].

Theorem 2.16 (Matoušek [13])**.** *Let $n \geq 1$ be an integer, $\epsilon \in (0,1)$, and $\delta \in (0,1)$, and let $k = C\epsilon^{-2}\log(\delta/2)$ with a suitable constant C. Define a random linear map $T : \mathbb{R}^n \to \mathbb{R}^k$ by*

$$(Tx)_i = \frac{1}{\sqrt{k}}\sum_{j=1}^{n} R_{ij}x_j, \quad i = 1, 2, \ldots, k,$$

where R_{ij} are independent random variables with $\mathbb{E}[R_{ij}] = 0$, $\mathrm{Var}(R_{ij}) = 1$, and a uniform sub-Gaussian tail or R_{ij} are independent sub-Gaussian random variables. Then for every $x \in \mathbb{R}^n$ we have

$$\Pr\left\{(1-\epsilon)\|x\|^2 \leq \|Tx\|^2 \leq (1+\epsilon)\|x\|^2\right\} \geq 1 - \delta.$$

Proof. Since T is linear, we can assume x is a unit vector. Then for a unit vector x, by Lemma 2.12, $Y_i := (Tx)_i = \sum_{j=1}^{d} R_{ij}x_j$ has properties $\mathbb{E}[Y_i] = 0, \mathrm{Var}(Y_i) = 1$, for $i = 1, 2, \ldots, k$ and has a sub-Guassian tail.

Next we consider the squared length of $(Y_1, Y_2, \cdots, Y_k)$. Note that the random variable $\|Tx\|^2 - 1$ with a fixed x (the randomness is from T) has the same distribution of $Z/\sqrt{k}$. So

$$\Pr\left\{\|Tx\|^2 \geq 1 + \epsilon\right\} = \Pr\left\{Z \geq \epsilon\sqrt{k}\right\} \leq e^{-C\epsilon^2 k} \leq \delta/2,$$

for $k \geq C\epsilon^{-2}\log(\delta/2)$. Repeat the same argument for $\Pr\left\{\|Tx\|^2 \leq 1 - \epsilon\right\}$ to finish the proof. □

Example 2.17. According to Theorem 2.16, the Gaussian is not essential. A very simple choice could be $R_{ij} = \pm 1$ with probability $1/2$. Then by Exercise 2.11, we have M-bound of R_{ij}, i.e., R_{ij} are sub-Gaussian.

Example 2.18. Achlioptas [1] proposed a computationally efficient choice: $R_{ij} = 0$ with probability $2/3$ and ± 1 with probability $1/6$ each (with a proper scaling $\sqrt{3}$ so that the variance is 1). The sub-Gaussian tail bound can be obtained by Hoeffding's inequality for bounded random variables. The matrix contains only $1/3$ non-zeros and requires only summation and subtraction operations (no multiplication since the scaling can be postponded). Therefore the computation is much faster.

Here we follow [11] to prove Hoeffding's inequality to verify the sub-Gaussian M-bound for a bounded random variable.

Proposition 2.19 (Hoeffding's inequality)**.** *Let X be a random variable with $\mathbb{E}[X] = 0$ and $a \le X \le b$. Then for $t > 0$,*

$$\mathbb{E}[e^{tX}] \le e^{t^2(b-a)^2/8}.$$

Proof. We use the convexity of the exponential function to get the inequality

$$e^{tx} \le w_1(x)e^{ta} + w_2(x)e^{tb}$$

with $w_1(x) = (x-a)/(b-a)$ and $w_2 = (b-x)/(b-a)$. Apply the expectation operator and notice that $\mathbb{E}[X] = 0$ to get

$$\mathbb{E}[e^{tX}] \le w_2(0)e^{tb} - w_1(0)e^{ta} = (1 - p + pe^{t(b-a)})e^{-pt(b-a)} = e^{\Phi(u)},$$

where $p = -a/(b-a), u = t(b-a)$, and $\Phi(u) = -pu + \log(1 - p + pe^u)$. Now it is a calculus problem to show $\Phi(u) \le t^2(b-a)^2/8$, e.g. by Taylor series. □

Exercise 2.20. Consider a random matrix T with i.i.d. entries $R_{ij} \sim N(0,1)$. Prove

$$\Pr\left\{(1-\epsilon)\|x\|^2 \le \|Tx\|^2 \le (1+\epsilon)\|x\|^2\right\} \ge 1 - 2e^{-(\epsilon^2-\epsilon^3)k/4},$$

directly using the moment generating function of the square of the standard normal distribution.

Exercise 2.21 (Preservation of inner products [12])**.** Let $f = 1/\sqrt{k}Tx$ where T is a J-L transform satisfying $\Pr\left\{|\|Tx\|^2 - 1| > \varepsilon\right\} \le \delta$ for a unit vector x. Prove for any two unit vectors $u, v \in \mathbb{R}^d$:

$$\Pr\left\{|u \cdot v - f(u) \cdot f(v)| > \epsilon\right\} \le 2\delta.$$

3. Fast JL Transform

The JL transforms discussed before require a dense $k \times d$ matrix with $k = \mathcal{O}(\epsilon^{-2}\ln n)$ and thus need $\mathcal{O}(kd) = \mathcal{O}(\epsilon^{-2}d\ln n)$ operations for transforming one data. Can we reduce the computational cost? The lower bound of $k_0 \ge \mathcal{O}(\epsilon^{-2}/|\log\epsilon|\ln n)$ by Alon [4] eliminates the possibility of using less number of rows. Then the only hope is to make the $k \times d$ matrix sparser, i.e., with more zero entries. One attempt is the approach by Achlioptas [1] which reduces the number of non-zero entry to $1/3$ of a full matrix, c.f. Example 2.18. Can we go further?

A naive choice would the coordinate axis vectors, e.g., randomly choosing k-coordinate axis vectors of the original $\mathbb{R}^d$. The corresponding matrix is a permutation matrix. If the transform matrix is sparse, however, then some vectors will be totally missed, e.g., the non-zero pattern of the vector coincides to be the zero pattern of rows. In the example of choosing k-coordinate axis, data points could be aligned to the other $n-k$ axis. If the matrix is considerably sparse, then there are considerably portion of such vectors. From this point of view, the randomization in the original JL transform can be thought as insurance against axis-alignment.

A sparse transform cannot be oblivious. It can be only applied to a special class of data. For example, if the d components of the vector x contribute almost equally to the norm $\|x\|$, randomly chose k components would give a good approximation of the norm.

Another choice would be the generalization of Achlioptas' idea. We could use the random variable: $R = 0$ with probability $1-q$ and $R = \pm q^{-1/2}$ with probability $q/2$. Then the transformation matrix contains $\mathcal{O}(kdq)$ non-zeros. When $q \ll 1$, e.g. $q = \mathcal{O}(1/d)$, then only $\mathcal{O}(k)$ operations is needed for one transformation. However, by Hoeffding's inequality, the M-bound will hold with a constant depending on $1/q$ and consequently k should be larger.

We shall follow Matoušek [13] to present a precise characterization of the trade-off between the sparsity and the tail bound. In the sequel, we use short nation $p_q = \Pr\{X = q\}$ for a discrete random variable X.

Lemma 3.1 (Matoušek [13])**.** *Let $\alpha^2 \le q \le 1$ and let $x \in \mathbb{R}^n$ be a unit vector with $\|x\|_\infty \le \alpha$. Chose i.i.d random variables $S_1, S_2, \ldots, S_d$ using the distribution $p_0 = 1-q, p_{q^{-1/2}} = q/2, p_{-q^{-1/2}} = q/2$ and let $Y = \sum_{i=1}^d S_i x_i$. Then Y has a sub-Gaussian tail up to $\sqrt{2q}/\alpha$.*

Proof. For such distribution, it is easy to check the M-bound. For each S_i, using the last inequality in Ex. 2.4, we have

$$\mathbb{E}\left[e^{tS}\right] = \frac{q}{2}\left(e^{t/\sqrt{q}} + e^{-t/\sqrt{q}}\right) + 1 - q \le qe^{t^2/(2q)} + 1 - q.$$

Then by the independence of S_i, we have

$$\mathbb{E}\left[e^{tY}\right] = \prod_{i=1}^d \mathbb{E}\left[e^{tx_iS_i}\right] \le \prod_{i=1}^d \left(qe^{t^2x_i^2/(2q)} + 1 - q\right).$$

When $\|x\|_\infty \le \alpha$, for $t^2 \le 2q/\alpha^2$, the power $t^2x_i^2/(2q) \le 1$. We can use the inequality of exponential functions to bound as

$$qe^{t^2x_i^2/(2q)} + 1 - q \le q(1 + t^2x_i^2/q) + 1 - q = 1 + t^2x_i^2 \le e^{t^2x_i^2}.$$

Consequently $\mathbb{E}\left[e^{tY}\right] \le \prod_{i=1}^d e^{t^2x_i^2} = e^{t^2}$ for $t^2 \le 2q/\alpha^2$. From the M-bound, we easily get the T-bound for $a \le a_0 = \sqrt{2q}/\alpha$. □

The parameter q in the above lemma can be thought of as a measure of the sparsity. The smaller q is, the sparser the matrix is. The non-zero value $\pm q^{-1/2}$ of the random variable is to satisfy the condition: mean zero and variance one. If no control of the maximum norm, the sub-Gaussian tail holds only up to $\sqrt{2q}$ which is small for a sparse transformation. Choosing $q = \mathcal{O}(\alpha^2)$ leads to a tail-bound with $O(1)$ range. Therefore we would like to control $\|x\|_\infty$. Note that for a unit vector x, $1/\sqrt{d} \le \|x\|_\infty \le 1$ and if $\|x\|_\infty = \mathcal{O}(1/\sqrt{d})$, then components of x are almost equi-distributed or in other words more spread out.

How to apply a sparse JL transform for a general set of data without such control of maximum norm? Ailon and Chazelle [2] propose an ingenious idea of combination of a sparse transformation with fast Fourier transform. They use the *Uncertainty Principle* in harmonic analysis: *A nonzero function and its Fourier transform cannot both be sharply localized*; see Appendix for a concise description. Based on that, they propose a discrete Fourier transform to enlarge the support of any sparse vector. To prevent the sparsification of dense vectors, the Fourier transform is randomized.

This procedure can be thought as precondition. That is preconditioned the $d \times n$ matrix by a $d \times d$ matrix such that the transformed data will have a lower maximum norm while the l^2-norm is preserved. Preconditioner proposed by Ailon and Chazelle [2] is the composition of the following matrices:

- H is the $d \times d$ normalized Hadamard matrix:

$$H_{ij} = d^{-1/2}(-1)^{(i-1,j-1)},$$

  where (i,j) is the dot-product of index i,j expressed in binary. The factor $d^{-1/2}$ is introduced such that the norm of row vectors is unit.
- D is a $d \times d$ diagonal matrix, where each D_{ii} is drawn independently from $\{-1,1\}$ with probability $1/2$.

The multiplication of the preconditioner HD matrix, which will be called *randomized Hardmard transformation*, can be computed fast, say, in $\mathcal{O}(d \ln d)$ operations using Walsh-Hadamard transform. It is easy to see HD is l^2-norm preserving. The following lemma shows the maximum norm is reduced and the 'energy' is more spread out.

Lemma 3.2 (Randomized Hardmard transformation [2])**.** *Let H, D be defined above. Then for any set V of n vectors in $\mathbb{R}^d$, with probability at least* $1 - 1/20$*,*

$$\max_{x \in V} \|HDx\|_\infty \leq \left(\frac{2\ln(40nd)}{d}\right)^{1/2} \|x\|_2. \tag{10}$$

Proof. Since HDx is linear in x, w.l.o.g. we assume $\|x\|_2 = 1$. Let $u = (u_1, u_2, \ldots, u_d)^T = HD(x_1, x_2, \ldots, x_d)^T$. Then $u_1 = \sum_{i=1}^d R_i x_i$, where each $R_i = \pm d^{-1/2}$ is chosen uniformly and independently. We compute the M-bound of u_1 as

$$\mathbb{E}\left[e^{tdu_1}\right] = \prod_{i=1}^d \mathbb{E}\left[e^{tdR_ix_i}\right] \prod_{i=1}^d \frac{1}{2}\left(e^{tx_i\sqrt{d}} + e^{-tx_i\sqrt{d}}\right) \leq \prod_{i=1}^d e^{t^2dx_i^2/2} = e^{t^2d/2}.$$

From the M-bound, we can derive the T-bound

$$\Pr\{|u_1| \geq a\} \leq 2e^{-a^2d/2} \leq \frac{1}{20nd},$$

by choosing $a^2 = 2\ln(40nd)/d$. Using the union bound over all nd coordinates of $HD(V)$, we obtain the desired result. □

With sub-Gaussian tail of Y and the bound of the maximum norm of preconditioned data, we can follow the previous proof to show the high concentration of the random variable $Z = \left(\sum_{i=1}^k Y_i^2 - k\right)/\sqrt{k}$. Note that Lemma 2.14 cannot be applied directly since the sub-Gaussian tail only holds up to a finite value $\sqrt{2q}/\alpha$. Further trick (using elementary calculus and probability estimate) is needed and can be found in [13].

If we denote the sparse projection matrix as P. Then $T = PHD$ is called Fast-Johnson-Lindenstrauss-Transform (FJLT). The projection matrix P is i.i.d. Gaussian in the original paper by Ailon and Chazelle [2] and improved to sub-Gaussian by Matoušek [13] which is more computationally efficient.

The computational cost of HDx is $\mathcal{O}(d\ln d + qdk)$. With the choice of $q = \mathcal{O}(\alpha^2) = \mathcal{O}(1/d\ln n)$, the cost becomes $\mathcal{O}(d\ln d + k\ln n)$. As only k-components of HDx is needed in $PHDx$, the first term $\mathcal{O}(d\ln d)$ can be further reduced to $\mathcal{O}(d\ln k)$, c.f. [3]. Recall that the dense J-L transform requires $\mathcal{O}(kd)$ operations. The fast J-L transform essentially changes the product kd to the addition $k+d$ penalized with logarithmic factors.

APPENDIX: THE UNCERTAINTY PRINCIPLE FOR FOURIER TRANSFORMS

We denoted by $\mathcal{S}(\mathbb{R})$ the Schwartz functions on the real line which is, roughly speaking, smooth functions and all derivatives decay faster than polynomials approaching to the infinity. Although the Fourier transform

$$\hat{f}(\xi) = \int_{\mathbb{R}} f(x) e^{-2\pi i x \xi} \, \mathrm{d}x$$

is well defined for $f \in L^1(\mathbb{R})$, restricted to $\mathcal{S}(\mathbb{R})$, the transform is isomorphism. And more importantly, restricted to $\mathcal{S}(\mathbb{R})$, the differentiation changes to a multiplication, e.g. $\widehat{f'(x)} = i\xi \hat{f}$ which can be proved using integration by parts. Details can be found in [7].

Theorem 3.3. *For any $f \in \mathcal{S}(\mathbb{R})$ and any $x_0, \xi_0 \in \mathbb{R}$, we have the following inequality:*

$$\|f(x)\|^2 \leq 4\pi \|(x - x_0) f(x)\| \|(\xi - \xi_0) \hat{f}(\xi)\|. \tag{11}$$

Proof. For simplicity we consider real functions. We use integration by parts to get

$$\int_{\mathbb{R}} |f(x)|^2 \, \mathrm{d}x = -2 \int_{\mathbb{R}} x f(x) f'(x) \, \mathrm{d}x,$$

where the boundary term at $-\infty$ and ∞ are vanished since $f \in \mathcal{S}(\mathbb{R})$. Therefore

$$\|f\|^2 \leq 2 \|x f(x)\| \|f'(x)\|.$$

Now we can compute the norm $\|f'(x)\|$ by its Fourier transform and obtain $\|f'(x)\| = 2\pi \|\xi \hat{f}(\xi)\|$. This finishes the proof for $x_0 = \xi_0 = 0$. For general cases, we can apply the zero case to $g(x) = e^{-2\pi i x \xi_0} f(x + x_0)$. □

Consider the case $|f|^2$ and $|\hat{f}|^2$ are probability density functions of random variables X and $\hat{X}$, and x_0 and ξ_0 are the expectation of X and $\hat{X}$, respectively. Then the right hand side are standard deviation. So the inequality implies that if one random variable is tightly concentrated in a small region then the other will have a wider spread. Smaller standard deviation implies a more precise prediction (or measurement). So if two probability density functions are Fourier transforms of one another, we can accurately predict at best one event. This is known as 'Uncertainty Principle' in quantum physics. Physical quantities of such conjugate pairs include: position and momentum, time and frequency.

REFERENCES

[1] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003.

[2] N. Ailon and B. Chazelle. The Fast JohnsonLindenstrauss Transform and Approximate Nearest Neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.

[3] N. Ailon and E. Liberty. Fast dimension reduction using Rademacher series on dual BCH codes. *Symposium on Discrete Algorithms*, pages 1–9, 2008.

[4] N. Alon. Problems and results in extremal combinatorics - I. *Discrete Mathematics*, 273(1-3):31–53, 2003.

[5] S. Dasgupta and A. Gupta. An Elementary Proof of a Theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65, 2003.

[6] P. Frankl and H. Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory, Series B*, 44(3):355–362, 1988.

[7] M. Hill. The uncertainty principle for fourier transforms on the real line. pages 1–17, 2013.

[8] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, 126:604–613, 1998.

[9] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability*, pages 189–206, 1984.

[10] E. Kushilevitz, R. Ostrovsky, and Y. Rabani. Efficient Search for Approximate Nearest Neighbor in High Dimensional Spaces. *SIAM J. Comput.*, 30(2):457–474, 2000.

[11] G. Lugosi. Concentration-of-measure inequalities. *Lecture Notes*, XXXIII(2):81–87, 2014.

[12] A. Magen. Dimensionality reductions that preserve volumes and distance to affine spaces. *Discrete and Computational Geometry*, 38(1):139–153, 2007.

[13] J. Matousek. On Variants of the JohnsonLindenstrauss Lemma. *Rand. Struct. Algor.*, 33(2):142–156, 2008.

[14] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science (New York, N.Y.)*, 290(5500):2323–2326, 2000.

[15] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science (New York, N.Y.)*, 290(5500):2319–2323, 2000.

[16] S. Venkatasubramanian and Q. Wang. The Johnson-Lindenstrauss Transform : An Empirical Study. pages 164–173, 2011.