Rescaled JL Embedding*

Erik Lindgren
erikml@utexas.edu
Department of Electrical and Computer Engineering
The University of Texas at Austin

Shanshan Wu
shanshan@utexas.edu
Department of Electrical and Computer Engineering
The University of Texas at Austin

December 5, 2016

Abstract

In this paper we present *Rescaled JL*, a new data-oblivious dimension reduction scheme that can potentially better preserve pairwise geometry. The key idea is based on a novel observation that removing norm distortions is much easier than removing distortions from angles. We prove high probability bounds for Rescaled JL in terms of pairwise Euclidean distances and dot products. We also demonstrate the practical gains of Rescaled JL over standard JL through simulations.

*This is course project for 2016 Fall UT-Austin graduate course CS 395T Sublinear Algorithm

1 Introduction

Dimensionality-reducing maps are widely used as a pre-processing step in modern large-scale machine learning applications. They transform high-dimensional data to a low-dimensional space while preserving certain geometric quantities (e.g., pairwise distance, subspace, etc.) of the original data. Compared to directly analyzing the original data, mining the transformed low-dimensional data has the benefit of small storage consumption and fast runtime. A powerful technique for generating *data-oblivious*[1] dimensionality reduction maps is the Johnson-Lindenstrauss (JL) lemma [JL84].

Theorem 1.1 (JL lemma [JL84]). *For any subset $X \subset \mathbb{R}^d$ of size n, and any $\epsilon \in (0, 1/2)$, there exists a map $f : X \to \mathbb{R}^m$ with $m = O(\log n/\epsilon^2)$ such that*

$$\forall x, y \in X, \quad (1-\epsilon)\|x-y\|_2^2 \le \|f(x)-f(y)\|_2^2 \le (1+\epsilon)\|x-y\|_2^2$$

Most data-oblivious dimensionality reduction maps considered in the literature are *linear* maps, i.e., $f(x) := Gx$, where $G \in \mathbb{R}^{m\times d}$ is a random matrix drawn from certain distribution. A simple example is a random Gaussian matrix.

Theorem 1.2 (see, e.g., Theorem 2.1 in [Woo14]). *Let $G \in \mathbb{R}^{m\times d}$, where each entry of G is drawn independently from $N(0, 1/m)$ Gaussian distribution. For any subset $X \subset \mathbb{R}^d$ of size n, and any $\epsilon, \delta \in (0, 1/2)$, if $m = O(\frac{1}{\epsilon^2}\log\frac{n}{\delta})$, then with probability at least $1-\delta$, $\forall x, y \in X$ we have*

$$\langle x, y\rangle - \epsilon\|x\|_2\|y\|_2 \le \langle Gx, Gy\rangle \le \langle x, y\rangle + \epsilon\|x\|_2\|y\|_2.$$

$$(1-\epsilon)\|x-y\|_2^2 \le \|Gx-Gy\|_2^2 \le (1+\epsilon)\|x-y\|_2^2,$$

In this project we consider a new *nonlinear* data-oblivious dimensionality reduction map called *Rescaled JL embedding*.

Definition 1.3. *Let $G \in \mathbb{R}^{m\times d}$, where each entry of G is drawn independently from a $N(0, 1/m)$ Gaussian distribution*[2]*. Then for any $x \in \mathbb{R}^d$, we define the Rescaled JL embedding $f : \mathbb{R}^d \to \mathbb{R}^m$ as*

$$f(x) := \frac{Gx}{\|Gx\|_2}\|x\|_2. \tag{1}$$

It is easy to check that $\|f(x)\|_2 = \|x\|_2$, i.e., the length of the vector is preserved after embedding, which is done by a simple rescaling operation. The idea of rescaling is first proposed in our recent paper [WBSD16], where we have empirically demonstrated the superior performance of rescaled JL embedding over normal JL embedding for preserving inner products. However, we do not provide theoretical justifications.

The goal of this course project is to advance our theoretical understanding of Rescaled JL embedding. In Section 3, we present our main theorem and its proof. In Section 4, we present experimental results using random Gaussian matrices and subsampled Hadamard matrices.

2 Notations

Let $\mathbb{N}_+$ denote the set of positive integers. For any $n \in \mathbb{N}_+$, let $[n]$ denote the set $\{1, 2, \cdots, n\}$. The ℓ_2-norm of a vector $x \in \mathbb{R}^d$ is defined as $\|x\|_2 = \left(\sum_{i=1}^d |x_i|^2\right)^{1/2}$. Let $\mathrm{F}(d_1, d_2)$ be the F-distribution with parameters d_1 and d_2. Let $\mathrm{Beta}(\alpha, \beta)$ be the Beta-distribution with parameters α and β.

[1] By data-oblivious, we mean that the randomness of the map does not depend on the input data.

[2] Although we define the Rescaled JL embedding for random Gaussian matrices, the same idea also works for other random matrices (see our experiments in Section 4).

3 Main Result

We now present a high probability bound for Rescaled JL embedding, which is the main theorem of this project. Note that we use the median trick here in order to obtain a high probability result.

Theorem 3.1. *Define t independent Rescaled JL maps: for $i \in [t]$, $f_i(x) = \frac{G_i x}{\|G_i x\|_2}\|x\|_2$, where each entry of $G_i \in \mathbb{R}^{k\times d}$ is drawn independently from a $N(0, 1/k)$ Gaussian distribution. For any subset $X \subset \mathbb{R}^d$ of size n, and any $\epsilon, \delta \in (0, 1/2)$, if $k = \lceil \frac{4}{\epsilon^2} \rceil$ and $t = O(\log \frac{n}{\delta})$, then with probability at least $1 - \delta$, $\forall x, y \in X$ we have*

$$\langle x, y\rangle - p(\theta_{x,y})\|x\|_2\|y\|_2 \le \underset{i\in[t]}{\text{median}} \langle f_i(x), f_i(y)\rangle \le \langle x, y\rangle + p(\theta_{x,y})\|x\|_2\|y\|_2,$$

$$\|x - y\|_2^2 - 2p(\theta_{x,y})\|x\|_2\|y\|_2 \le \underset{i\in[t]}{\text{median}} \|f_i(x) - f_i(y)\|_2^2 \le \|x - y\|_2^2 + 2p(\theta_{x,y})\|x\|_2\|y\|_2,$$

where $\theta_{x,y}$ is the actual angle between x and y, i.e., $\langle x, y\rangle = \|x\|_2\|y\|_2 \cos\theta_{x,y}$.

The error term $p(\theta)$ is a function of $\theta \in [0, \pi]$, where $\epsilon(\pi/2) \le \epsilon$ and $\epsilon(0) = \epsilon(\pi) = 0$. It is defined as an expectation over two independent random variables:

$$p(\theta) = 2\sqrt{\underset{g,u}{\mathbf{E}}[\gamma^2 - 2\cos\theta\gamma] + \cos^2\theta}, \tag{2}$$

where

$$g \sim F(k, k), \quad \frac{u+1}{2} \sim Beta(\frac{k-1}{2}, \frac{k-1}{2}), \quad \gamma = \frac{u\sin\theta + \sqrt{g}\cos\theta}{\sqrt{g\cos^2\theta + \sin^2\theta + 2u\sin\theta\cos\theta\sqrt{g}}}$$

Table 1 gives a comparison of the guarantees provided by Theorem 3.1 and Theorem 1.2. We see that the output dimensions m and kt have the same dependence on $O(\frac{1}{\epsilon^2}\log\frac{n}{\delta})$. A key difference is that, JL lemma produces an additive error for inner product and a multiplicative error for squared distance, while our Rescaled JL gives additive errors to both inner product and squared distance.

Table 1: A comparison of the guarantees provided by Standard JL and Rescaled JL.

Methods	Standard JL	Rescaled JL
Reduced dimension	$O(\frac{1}{\epsilon^2}\log\frac{n}{\delta})$	$O(\frac{1}{\epsilon^2}\log\frac{n}{\delta})$
Dot product	$\langle x, y\rangle \pm \epsilon\|x\|_2\|y\|_2$	$\langle x, y\rangle \pm p(\theta_{x,y})\|x\|_2\|y\|_2$
Euclidean distance	$(1+\epsilon)\|x - y\|_2^2$	$\|x - y\|_2^2 \pm 2p(\theta_{x,y})\|x\|_2\|y\|_2$

The additive error term $p(\theta)$ in Theorem 3.1 depends on the actual angle between the input vectors. To see how $p(\theta)$ varies with θ, we plot $p(\theta)$ as a function of θ for $k = 400$ ($\epsilon = 0.1$) in Figure 1. We see that $p(\theta)$ has a bell-shaped curve, which archives maximum at $\pi/2$ and minimum at 0 and π. It is also easy to check that $p(\pi/2) \le \epsilon$ and $p(0) = p(\pi) = 0$ by substituting $\theta = 0$, $\pi/2$, and π into (2):

$$\epsilon(0) = 2\sqrt{\underset{g,u}{\mathbf{E}}[\gamma^2 - 2\gamma] + 1} = 2\sqrt{\underset{g,u}{\mathbf{E}}[0]} = 0, \quad \epsilon(\pi/2) = 2\sqrt{\underset{g,u}{\mathbf{E}}[\gamma^2]} = 2\sqrt{\underset{u}{\mathbf{E}}[u^2]} = 2\sqrt{1/k} \le \epsilon,$$

where the last inequality follows from $k = \lceil \frac{4}{\epsilon^2} \rceil$.

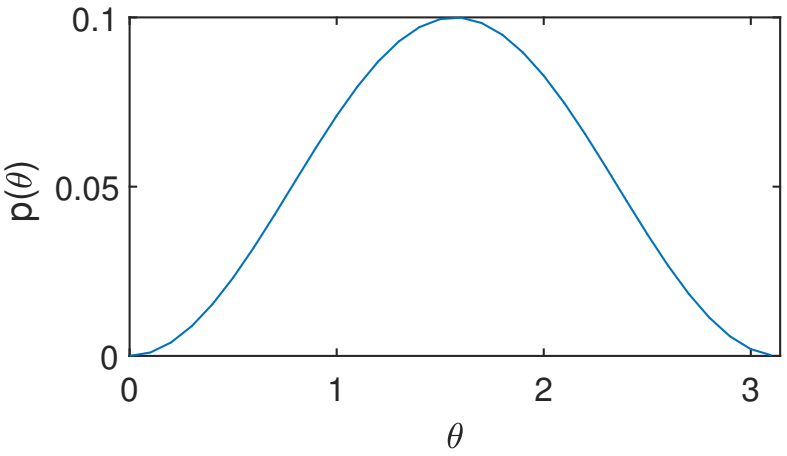

Figure 1: We use Matlab to evaluate (2) and plot $p(\theta)$ as a function of θ for $k = 400$ ($\epsilon = 0.1$).

3.1 Proof of Theorem 3.1

We first prove a constant probability bound (Lemma 3.3) for Rescaled JL embedding, and then use it to derive a high probability bound (Theorem 3.1). The proof of Lemma 3.3 uses the result of Lemma 3.2.

Lemma 3.2. *Let u be the inner product of two vectors that are independently uniformly distributed on the unit sphere S^{k-1}, then $(u+1)/2 \sim Beta(\frac{k-1}{2}, \frac{k-1}{2})$.*

Proof. A proof of this lemma can be found at [Hub14]. □

Lemma 3.3. *Given two vectors $x, y \in \mathbb{R}^d$, define a Rescaled JL mapping: $f(x) = \frac{Gx}{\|Gx\|_2}\|x\|_2$, where each entry of $G \in \mathbb{R}^{k\times d}$ is drawn independently from $N(0, 1/k)$ Gaussian distribution. For any $\epsilon \in (0, 1/2)$, if $k = \lceil \frac{4}{\epsilon^2} \rceil$, then with probability at least 3/4, we have*

$$\langle x, y\rangle - p(\theta_{x,y})\|x\|_2\|y\|_2 \le \langle f(x), f(y)\rangle \le \langle x, y\rangle + p(\theta_{x,y})\|x\|_2\|y\|_2,$$

where $\theta_{x,y}$ is the angle between x and y, and $p(\cdot)$ is defined in (2).

Proof. We will show that

$$\mathbf{E}[(\langle f(x), f(y)\rangle - \langle x, y\rangle)^2] = p(\theta_{x,y})^2\|x\|_2^2\|y\|_2^2/4. \tag{3}$$

If (3) holds, then Markov's inequality will give the desired bound:

$$\mathbb{P}(|\langle f(x), f(y)\rangle - \langle x, y\rangle| \ge p(\theta_{x,y})\|x\|_2\|y\|_2) \le \frac{\mathbf{E}[(\langle f(x), f(y)\rangle - \langle x, y\rangle)^2]}{p(\theta_{x,y})^2\|x\|_2^2\|y\|_2^2} \le \frac{1}{4}. \tag{4}$$

To prove (3), note that

1. Since $\langle f(x), f(y)\rangle = \left\langle f(\frac{x}{\|x\|_2}), f(\frac{y}{\|y\|_2})\right\rangle \|x\|_2\|y\|_2$, we only need to prove Lemma 3.3 for unit vectors. In other words, we can assume $\|x\|_2 = \|y\|_2 = 1$ and prove that

$$\mathbf{E}[(\langle f(x), f(y)\rangle - \langle x, y\rangle)^2] = p(\theta_{x,y})^2/4.$$

2. The random Gaussian matrix G will project x and y on a random subspace, so without loss of generality we can assume that $x = e_1$ and $y = e_1 \cos\theta_{x,y} + e_2 \sin\theta_{x,y}$, where e_1 and e_2 are standard basis vectors in $\mathbb{R}^d$.

Let $g_1, g_2 \in \mathbb{R}^k$ be the first two column vectors of G. Let $u = \left\langle \frac{g_1}{\|g_1\|_2}, \frac{g_2}{\|g_2\|_2} \right\rangle$, then

$$\langle Gx, Gy\rangle = \langle g_1, g_1\cos\theta_{x,y} + g_2 \sin\theta_{x,y}\rangle = \|g_1\|_2^2 \cos\theta_{x,y} + \|g_1\|_2\|g_2\|_2 u \sin\theta_{x,y}.$$

$$\begin{aligned}\|Gx\|_2\|Gy\|_2 &= \|g_1\|_2\|g_1\cos\theta_{x,y} + g_2\sin\theta_{x,y}\|_2 \\ &= \|g_1\|_2\sqrt{\|g_1\|_2^2\cos^2\theta_{x,y} + 2\|g_1\|_2\|g_2\|_2 u\cos\theta_{x,y}\sin\theta_{x,y} + \|g_2\|_2^2\sin^2\theta_{x,y}}.\end{aligned}$$

Let $g = \frac{\|g_1\|_2^2}{\|g_2\|_2^2}$, then

$$\langle f(x), f(y)\rangle = \frac{\langle Gx, Gy\rangle}{\|Gx\|_2\|Gy\|_2} = \frac{\sqrt{g}\cos\theta_{x,y} + u\sin\theta_{x,y}}{\sqrt{g\cos^2\theta_{x,y} + \sin^2\theta_{x,y} + 2u\sin\theta_{x,y}\cos\theta_{x,y}\sqrt{g}}} = \gamma.$$

Here γ has the same definition as in Theorem 3.1, where $g \sim \mathrm{F}(k,k)$, and $\frac{u+1}{2} \sim \mathrm{Beta}(\frac{k-1}{2}, \frac{k-1}{2})$ (Lemma 3.2). Now we can compute $\mathbf{E}[(\langle f(x), f(y)\rangle - \langle x, y\rangle)^2]$ as

$$\begin{aligned}\mathbf{E}[(\langle f(x), f(y)\rangle - \langle x, y\rangle)^2] &= \mathbf{E}[\langle f(x), f(y)\rangle^2] - 2\cos\theta_{x,y}\,\mathbf{E}[\langle f(x), f(y)\rangle] + \cos^2\theta_{x,y} \\ &= \mathop{\mathbf{E}}_{g,u}\gamma^2 - 2\cos\theta_{x,y}\mathop{\mathbf{E}}_{g,u}\gamma + \cos^2\theta_{x,y} \\ &= p(\theta_{x,y})^2/4,\end{aligned}$$

where the last inequality follows from the definition of $p(\theta)$ in (2). We have thus proved (3) under the assumption of unit vectors. The lemma then holds because of (4). □

Proof of Theorem 3.1. For a fixed pair $x, y \in X$, we define t independent binary random variables $Z_1, ..., Z_t$, where $Z_i = 1$ if the Rescaled Jl mapping $f_i(\cdot)$ satisfies

$$\langle x, y\rangle - p(\theta_{x,y})\|x\|_2\|y\|_2 \le \langle f_i(x), f_i(y)\rangle \le \langle x, y\rangle + p(\theta_{x,y})\|x\|_2\|y\|_2.$$

Lemma 3.3 says that for any $i \in [t]$, $\mathbb{P}[Z_i = 1] \ge 3/4$. Let Z be 1 if

$$\langle x, y\rangle - p(\theta_{x,y})\|x\|_2\|y\|_2 \le \operatorname*{median}_{i\in[t]} \langle f_i(x), f_i(y)\rangle \le \langle x, y\rangle + p(\theta_{x,y})\|x\|_2\|y\|_2,$$

and 0 otherwise. Then

$$\mathbb{P}[Z = 0] \le \mathbb{P}[\frac{1}{t}\sum_i^t Z_i \le 1/2] \le e^{-O(t)} \le \frac{\delta}{n^2}, \tag{5}$$

where the last two inequalities follow from the fact that $\frac{1}{t}\sum_i^t Z_i$ is subguassian with variance $O(1/t)$, and $t = O(\log\frac{n}{\delta})$. Eq. (5) holds for any fixed pair $x, y \in X$. There are no more than n^2 different pairs, so $\forall x, y \in X$, the following is true with probability at least $1 - \delta$,

$$\langle x, y\rangle - p(\theta_{x,y})\|x\|_2\|y\|_2 \le \operatorname*{median}_{i\in[t]} \langle f_i(x), f_i(y)\rangle \le \langle x, y\rangle + p(\theta_{x,y})\|x\|_2\|y\|_2.$$

We have proved a high probability bound for inner products. The high probability bound for squared distances is straightforward because

$$\begin{aligned}\|f_i(x) - f_i(y)\|_2^2 &= \|f_i(x)\|_2^2 + \|f_i(x)\|_2^2 - 2\langle f_i(x), f_i(y)\rangle \\ &= \|x\|_2^2 + \|y\|_2^2 - 2(\langle x, y\rangle \pm p(\theta_{x,y})\|x\|_2\|y\|_2) \\ &= \|x - y\|_2^2 \pm 2p(\theta_{x,y})\|x\|_2\|y\|_2.\end{aligned}$$

□

4 Experiments

In this section we will compare the empirical performance of JL and Rescaled JL. The goal is to see how well the dot product as well as Euclidean distance computed in the low-dimension space approximates those in the original high-dimension space.

4.1 Matlab Simulations

We conduct the following experiment in Matlab:

- For each $\theta \in [0, \pi]$ with interval 0.05, we generate 10 random pairs of vectors $x, y \in \mathbb{R}^{100}$ such that $\|x\|_2 = \|y\|_2 = 1$ and $\langle x, y \rangle = \|x\|_2 \|y\|_2 \cos\theta$.
- For each pair of vectors x, y generated in the previous step, we independently construct a random Gaussian matrix $G \in \mathbb{R}^{10\times 100}$, where each entry $G_{ij} \sim N(0, 1/10)$. Then we use JL ($f : x \to Gx$) and Rescaled JL ($f : x \to \frac{Gx}{\|Gx\|_2}\|x\|_2$) to transform x, y into $f(x), f(y) \in \mathbb{R}^{10}$.
- We generate scatter plots of $(\langle f(x), f(y)\rangle, \langle x, y\rangle)$ and $(\|f(x) - f(y)\|_2^2, \|x - y\|_2^2)$ in Figure 2. We see that Rescaled JL (shown in red dots) seems to have smaller variance than JL (shown in blue dots). Besides, the shape of red dots matches the shape of $p(\theta)$ shown in Figure 1.

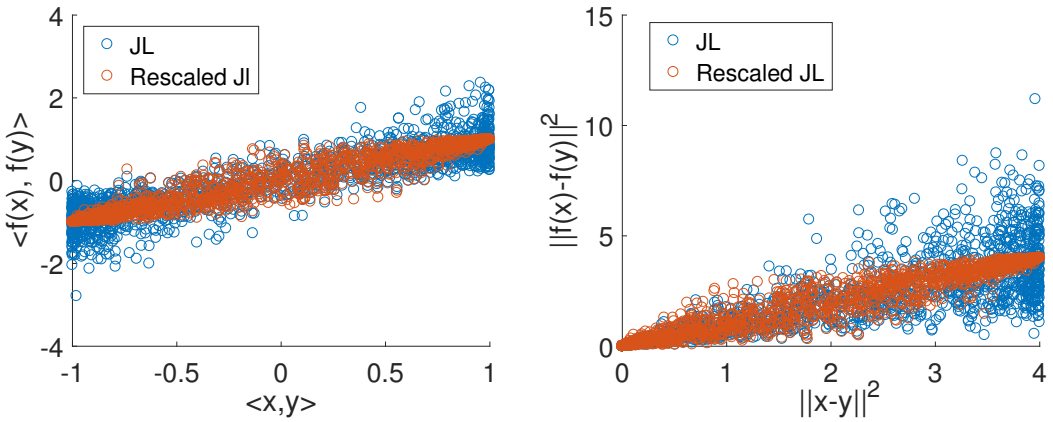

Figure 2: Scatter plots of $(\langle f(x), f(y)\rangle, \langle x, y\rangle)$ (left) and $(\|f(x) - f(y)\|_2^2, \|x - y\|_2^2)$ (right).

4.2 Mean Squared Error

The Mean Squared Error (MSE) for estimating dot product and Euclidean distance is defined as

$$\mathbf{E}[(\langle f(x), f(y)\rangle - \langle x, y\rangle)^2], \quad \mathbf{E}[(\|f(x) - f(y)\|_2^2 - \|x - y\|_2^2)^2],$$

where the expectation is over the randomness of the map f.

We have already computed an exact form of MSE for Rescaled JL in the proof Lemma 3.3 (see Eq. (3)). Similar computation can be done for standard JL. In Figure 3 we compare the MSE as a function of θ, where we see that Rescaled JL has smaller MSE than JL for dot product estimation, but can perform worse for Euclidean distance estimation.

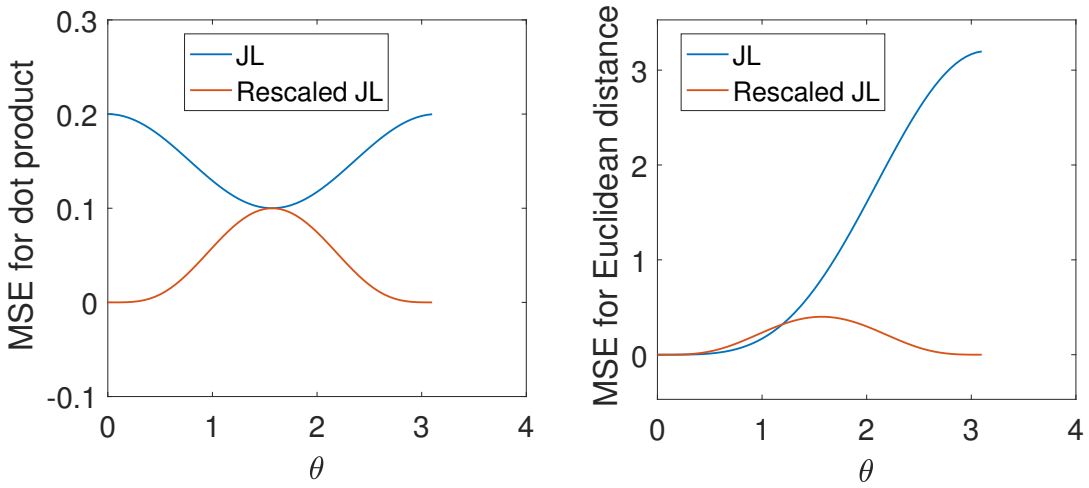

Figure 3: MSE of dot product estimation (left) and Euclidean distance estimation (right). We use the same parameters in Section 4.1: the reduced dimension is 10, and $\|x\|_2 = \|y\|_2 = 1$.

Note that in Figure 3, we assume that $\|x\|_2 = \|y\|_2 = 1$, so a natural question is to ask if the MSE curve has the same trend when $\|x\|_2 \neq \|y\|_2$. This is true for dot product estimation since $\langle f(x), f(y)\rangle = \left\langle f(\frac{x}{\|x\|_2}), f(\frac{y}{\|y\|_2})\right\rangle \|x\|_2\|y\|_2$, which indicates that the ratio between the MSE for Rescaled JL and JL will not change if we choose different $\|x\|_2, \|y\|_2$. However, it is not true for Euclidean distance estimation. For example, in Figure 4 we plot the MSE for $\|x\|_2 = 1, \|y\|_2 = 2$. We see that the Rescaled JL always has smaller MSE than JL, which is different from the case when $\|x\|_2 = \|y\|_2 = 1$ (shown in Figure 3).

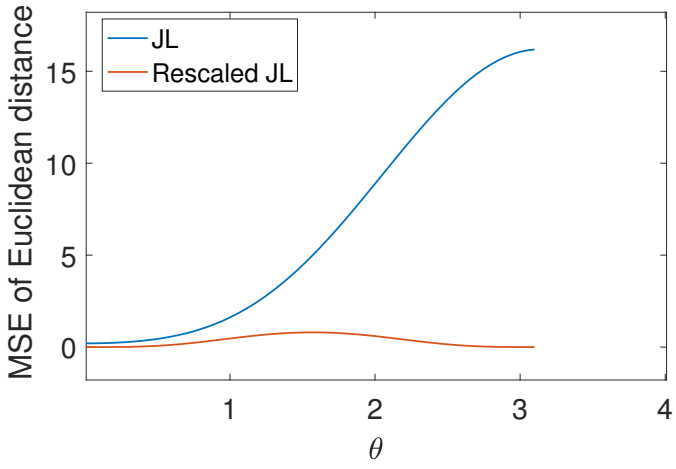

Figure 4: MSE of Euclidean distance estimation for $\|x\|_2 = 1, \|y\|_2 = 2$.

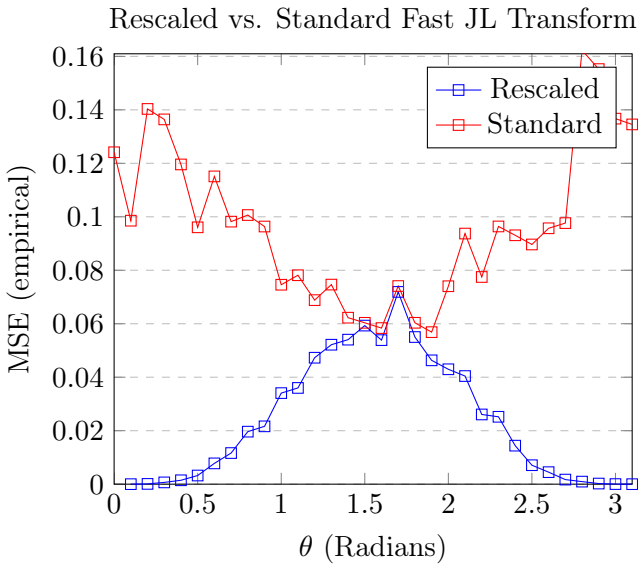

Figure 5: We simulate the effect rescaling has on the MSE of dot products using the Fast JL transform. We see that rescaling drastically improves performance for lower values of the angle θ.

4.3 Fast JL vs Rescaled Fast JL

So far we only consider random Gaussian matrices. In practice, we would want to use random matrices with special structures in order to speed up the computation. Although our theoretical analysis technique does not immediately allow us to analyze these variants, we experimentally see that rescaling does improve performance greatly. In Figure 5 we simulate the Fast JL Transform [AC09] and its Rescaled version. To simulate data, for each angle θ we generate 50 random pairs of unit vectors in 1024 dimensions with an angle θ. The sparsity was set to 0.01 and the dimension after projection is 20. We empirically compute the MSE for dot products. Figure 5 indicates that rescaling still improves the performance for Fast JL Transform.

5 Conclusion

In this project we present a new idea of using rescaling to improve the accuracy of estimating dot product and Euclidean distance (Eq. (1)). Theoretical guarantee is derived for Rescaled JL (Theorem 3.1), which is then compared with standard JL (Table 1). We compute the exact form of mean squared error (MSE) for Rescaled JL and JL. We observe that for dot product estimation, Rescaled JL has smaller MSE than JL (Figure 3). However, this is not true for Euclidean distance estimation, which depends on the norms of the input vectors (Figure 4). While our analysis focuses on the simple random Gaussian matrices, we also demonstrated the effect of rescaling on the Fast JL transform (Figure 5).

References

[AC09] Nir Ailon and Bernard Chazelle. The fast johnson-lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.

[Hub14] William A. Huber. Distribution of a scalar product of two random unit vectors in $\mathbb{R}^d$, 2014. URL: http://stats.stackexchange.com/questions/85916/distribution-of-a-scalar-product-of-two-random-unit-vectors-in-mathbbrd (version: Feb 9, 2014).

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

[WBSD16] Shanshan Wu, Srinadh Bhojanapalli, Sujay Sanghavi, and Alexandros G. Dimakis. Single pass pca of matrix products. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[Woo14] David P Woodruff. Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*, 2014.