The Johnson-Lindenstrauss Lemma

Theory, Extensions, and Applications

David Veitch
University of Toroonto
david.veitch@mail.utoronto.ca
December 2020

***Abstract*—The Johnson-Lindenstrauss Lemma serves as the theoretical basis for several random projection dimensionality reduction methods. This paper examines the theory behind random projections, investigates extensions that have recently been made including database friendly random projections, and then explores applying random projections in a regression framework, and a data visualization task.**

I. Introduction

In today's 'big data' era, the importance of statistical methods which are robust to the curse of dimensionality has never been greater. Datasets where the number of features d is large (and particularly the case where d is much larger than the number of samples n) present a number of unique problems for analysis. For one it becomes difficult to visualize the structure of a dataset for d much greater than 3. As well, algorithms such as k-nearest neighbours can have computational complexities of $O(dn)$, making computation intractable for large datasets.

A number of methods have been proposed to address the above issues. One such method is random projections, the theory of which is provided by the Johnson-Lindenstrauss (JL) Lemma [1]. While on one hand this Lemma is surprisingly simple, as we shall see in this paper it is quite powerful, and made even more so through a number of recent extensions.

This paper shall be organized as follows. First, I will state and prove the 'standard' JL Lemma where the random projection matrix is made up of Gaussian random variables. Second, I will prove the database friendly case, where the projection matrix is of a simpler form. Third, I will discuss Compressed Least-Squares Regression, which applies the JL Lemma to regression problems where $d > n$. Finally, I will discuss applications of random projections in regression and data visualization.

II. Statement of JL Lemma

Let $\epsilon \in (0, \frac{1}{2})$. Let $Q \subset \mathbb{R}^d$ be a set of n points and $k = \frac{20 \log n}{\epsilon^2}$. There exists a Lipschitz mapping $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $u, v \in Q$

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

III. Proof of JL Lemma

The proof in this paper follows the argument found in the Lecture Notes of Kakade and Shakhnarovich [2]. However, before proving the JL lemma we must establish some preliminary technical lemmas.

A. Lemma 1 - Norm Preservation Lemma (Gaussian Case)

For any $x \in \mathbb{R}^d$, and A which is a $k \times d$ matrix where each entry $a_{i,j}$ is a randomly generated $\mathcal{N}(0,1)$ random variable

$$\mathbb{E}[||\frac{1}{\sqrt{k}}Ax||^2] = \mathbb{E}[||x||^2]. \quad (2)$$

B. Proof of Lemma 1

Denoting $[Ax]_i$ as the ith entry of the vector Ax we see that

$$\mathbb{E}[[Ax]_i^2] = \mathbb{E}[(\sum_{j=1}^{d} a_{i,j}x_j)^2] \quad (3)$$

$$= \mathbb{E}[\sum_{j=1}^{d}\sum_{j'=1}^{d} a_{i,j}a_{i,j'}x_j x_{j'}] \quad (4)$$

$$= \sum_{j=1}^{d}\mathbb{E}[a_{i,j}^2]x_j^2 + \sum_{j\neq j'}\mathbb{E}[a_{i,j}]\mathbb{E}[a_{i,j'}]x_j x_{j'} \quad (5)$$

$$= \sum_{j=1}^{d} x_j^2 \quad (6)$$

$$= ||x||^2. \quad (7)$$

In Equation 5 we used the independence of the entries of A and in Equation 6 we used the fact for a standard normal random variable Z, $\mathbb{E}[Z^2] = \sigma_z^2 = 1$. Next we see that

$$||\frac{1}{\sqrt{k}}Ax||^2 = \frac{1}{k}\sum_{i=1}^{k}[Ax]_i^2 \quad (8)$$

implies

$$\mathbb{E}[||\frac{1}{\sqrt{k}}Ax||^2] = \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[[Ax]_i^2]] = \frac{1}{k}\sum_{i=1}^{k}||x||^2 = ||x||^2. \quad (9)$$

C. Lemma 2

For any $\epsilon \in (0, \frac{1}{2})$

$$1+\epsilon \leq \exp(\epsilon - (\epsilon^2 - \epsilon^3)/2). \quad (10)$$

D. Proof of Lemma 2

For this I use the Power Series expansion of $\log(1+\epsilon)$ for $-1 < \epsilon \leq 1$, and then in the bound I use the fact that for our purposes $\epsilon \in (0, \frac{1}{2})$.

$$\log(1+\epsilon) = \sum_{n=1}^{\infty}(-1)^{n-1}\frac{\epsilon^n}{n} \tag{11}$$

$$= \epsilon - \frac{\epsilon^2}{2} + \frac{\epsilon^3}{3} + \sum_{n=1}^{\infty}\underbrace{-\epsilon^{2n+2}}_{<0}\underbrace{\Big(\frac{1}{2n+2} - \frac{\epsilon}{2n+3}\Big)}_{\geq 0 \ \forall \ n, \ \epsilon \in (0,\frac{1}{2})} \tag{12}$$

$$\leq \epsilon - \frac{\epsilon^2 - \epsilon^3}{2}. \tag{13}$$

Then by the monotonicity of the exponential

$$1 + \epsilon \leq \exp(\epsilon - (\epsilon^2 - \epsilon^3)/2). \tag{14}$$

E. Lemma 3 - χ^2 Concentration Inequality

For a chi-squared random variable with k degrees of freedom

$$\mathbb{P}(\chi_k^2 \geq (1+\epsilon)k) \leq \exp(-\frac{k}{4}(\epsilon^2 - \epsilon^3)) \tag{15}$$

$$\mathbb{P}(\chi_k^2 \leq (1-\epsilon)k) \leq \exp(-\frac{k}{4}(\epsilon^2 - \epsilon^3)). \tag{16}$$

F. Proof of Lemma 3

For standard normal random variables $Z_1, \ldots, Z_k$, and some $\lambda > 0$

$$\mathbb{P}(\chi_k^2 \geq (1+\epsilon)k) = \mathbb{P}(\sum_{i=1}^{k} Z_i^2 > (1+\epsilon)k) \tag{17}$$

$$= \mathbb{P}\Big(\exp(\lambda\sum_{i=1}^{k} Z_i^2) \geq \exp((\lambda(1+\epsilon)k)\Big) \tag{18}$$

$$\leq \frac{\mathbb{E}[\exp(\lambda\sum_{i=1}^{k} Z_i^2)]}{\exp((\lambda(1+\epsilon)k)} \tag{19}$$

$$= \frac{\mathbb{E}[\exp(\lambda Z_1^2)]^k}{\exp((\lambda(1+\epsilon)k)} \tag{20}$$

$$= (1-2\lambda)^{-\frac{k}{2}}\exp((-\lambda(1+\epsilon)k). \tag{21}$$

In the above, Equation 19 uses Markov's Inequality, Equation 20 uses the $Z_i's$ independence, and Equation 21 uses the fact that $\mathbb{E}[\exp(\lambda Z_1^2)]$ is the moment-generating function of a χ^2 random variable with one degree of freedom.

Substituting the value of k stated in the JL Lemma into Equation 21 leads to $\lambda = \frac{\epsilon}{2(1+\epsilon)}$ minimizing the equation, and also fulfilling the requirement that $\lambda < \frac{1}{2}$ (a requirement for the moment generating function of a χ^2 random variable). Substituting this optimal λ into Equation 21 yields

$$\mathbb{P}(\chi_k^2 \geq (1+\epsilon)k) \leq ((1+\epsilon)e^{-\epsilon})^{\frac{k}{2}}. \tag{22}$$

Combining Equation 22 with Equation 14 yields the result stated in Equation 15. The bound of $\mathbb{P}(\chi_k^2 \leq (1-\epsilon)k)$ can be proven in a nearly identical fashion [3].

G. Proof of JL Lemma (Gaussian Case)

Since a linear combination of Gaussian random variables is Gaussian, and the fact $\mathbb{E}[[Ax]_i] = 0$, $\mathbb{E}[[Ax]_i^2] = \text{Var}[[Ax]_i] = ||x||^2$

$$\tilde{Z}_i = [Ax]_i/||x|| \sim \mathcal{N}(0,1). \tag{23}$$

Combining this fact with Lemma 3

$$\mathbb{P}(||\frac{1}{\sqrt{k}}Ax||^2 > (1+\epsilon)||x||^2) = \mathbb{P}\Big(\sum_{i=1}^{k}\Big(\frac{[Ax]_i}{||x||}\Big)^2 > (1+\epsilon)k\Big) \tag{24}$$

$$= \mathbb{P}(\sum_{i=1}^{k}\tilde{Z}_i^2 > (1+\epsilon)k) \tag{25}$$

$$= \mathbb{P}(\chi_k^2 > (1+\epsilon)k) \tag{26}$$

$$\leq \exp(-\frac{k}{4}(\epsilon^2 - \epsilon^3)). \tag{27}$$

Now to complete the proof, I use subadditivity and the fact there are n^2 pairs of $u, v \in Q$. That is

$$\mathbb{P}(\exists \ u, v \text{ such that Equation 1 fails}) \tag{28}$$

$$\leq \sum_{u,v \in Q}\mathbb{P}(\text{Equation 1 fails for specific } u, v) \tag{29}$$

$$\leq 2n^2\exp(-\frac{k}{4}(\epsilon^2 - \epsilon^3)) \tag{30}$$

$$< 1. \tag{31}$$

where the final step in Equation (31) occurs if k is chosen to be $\frac{20}{\epsilon^2}\log n$.

To recap, what has been shown is that by defining the Lipschitz mapping f as the randomly generated projection matrix A, there is a nonzero probability that Equation 1 holds. Therefore since this probability is nonzero, there must exist some A^* which satisfies Equation 1, and hence such an A^* is the Lipschitz mapping I sought to prove existed.

As well by using Equation 30 it is possible for a fixed n to choose ϵ (which then determines k) to bound the probability that the A matrix that is generated leads Equation 1 to fail. In the case where conducting the matrix multiplication associated with the projection has a high computational cost, or the dataset is large so that it is costly to check if the projection did in fact preserve distances, this can be useful in ensuring with high probability that the projection 'gets it right' on the first shot.

IV. Database Friendly Random Projections

Note that in the proof of the JL Lemma certain properties of normal random variables were used to establish the probability bounds in Lemma 3, but there was nothing preventing one from using other random variables to construct the matrix A. Therefore, provided the same types of bounds can be established with other types of random variables it should be

possible to construct the projection matrix A in alternative ways. This observation is behind several extensions to the JL Lemma which utilize sparse A matricies. Namely A is constructed in such a way that $a_{i,j} \in \{0, \pm c\}\ c \in \mathbb{R}\ \forall\ i, j$.

One such construction of the A matrix, a database friendly random projection, is proposed by Achiloptas [4]. It is database friendly in the sense that the operations required can be implemented in SQL (a database programming language) using only its base functionality. This enables the use of random projections in database settings, a setting where often d and n are very large. As well, in the version of this random projection where 2/3 of the entries in the A matrix is equal to 0, this results in a substantial improvement in computation time since there are optimized algorithms for sparse matrix multiplication.

A. Theorem 1 - Database Friendly Random Projections

Let $Q \subset \mathbb{R}^d$ be a set of n points. For a given $\epsilon, \beta > 0$ let

$$k_0 = \frac{4 + 2\beta}{\epsilon^2/2 - \epsilon^3/3} \log n \tag{32}$$

and for some integer $k \geq k_0$ let $A \in \mathbb{R}^{d \times k}$ be a random matrix where its entries are independent random variables from either one of the following two probability distributions

$$a_{i,j} = \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases} \tag{33}$$

$$a_{i,j} = \begin{cases} \sqrt{3} & \text{with probability } \frac{1}{6} \\ 0 & \text{with probability } \frac{2}{3} \\ -\sqrt{3} & \text{with probability } \frac{1}{6}. \end{cases} \tag{34}$$

Let

$$E = \frac{1}{\sqrt{k}} QA \tag{35}$$

and $f : \mathbb{R}^d \to \mathbb{R}^k$ map the i^{th} row of Q to the i^{th} row of E. Then with probability of at least $1 - n^{-\beta}$ for all $u, v \in Q$

$$(1 - \epsilon)||u - v||^2 \leq ||f(u) - f(v)||^2 \leq (1 + \epsilon)||u - v||^2. \tag{36}$$

B. Preliminaries

Here I will follow the notational style of Achiloptas by scaling A by $\frac{1}{\sqrt{d}}$, and correspondingly to calculate the matrix E I scale E by $\sqrt{\frac{d}{k}}$. Then for some $x \in \mathbb{R}^d$, $f(x) = \sqrt{\frac{d}{k}}([Ax]_1, \ldots, [Ax]_k)$. Next define for an arbitrary x

$$S = S(x) = \sum_{j=1}^{k} [Ax]_j^2. \tag{37}$$

Using the same arguments of Lemma 1 yields that $\mathbb{E}[[Ax]_i] = 0$ and $\mathbb{E}[[Ax]_i^2] = \frac{1}{d}||x||^2$ (note here the $\frac{1}{d}$ comes from the previously mentioned scaling of the A matrix).

A few technical lemmas are now required before proving the main theorem.

C. Lemma 4

For all $\lambda \in [0, d/2)$ and all $d \geq 1$

$$\mathbb{E}[\exp(\lambda [Ax]_i^2)] \leq \frac{1}{\sqrt{1 - 2\lambda/d}} \tag{38}$$

$$\mathbb{E}[[Ax]_i^4] \leq \frac{3}{d^2}. \tag{39}$$

D. Proof of Lemma 4

The proof of Lemma 4, which can be found in the original paper, is fairly involved, and does not particularly contribute to our understanding of random projections. Therefore I will not repeat these steps here.

E. Lemma 5

For a_{ij} being distributed as in Equation 33 or Equation 34 then for any $\epsilon > 0$ and any unit vector $x \in \mathbb{R}^d$

$$\mathbb{P}(S > (1 + \epsilon)k/d)) < \exp(-\frac{k}{2}(\epsilon^2/2 - \epsilon^3/3)) \tag{40}$$

$$\mathbb{P}(S < (1 - \epsilon)k/d)) < \exp(-\frac{k}{2}(\epsilon^2/2 - \epsilon^3/3)). \tag{41}$$

F. Proof of Lemma 5

The Proof of Lemma 5 begins in a similar fashion to the proof of the JL Lemma, utilizing Markov's inequality and moment generating functions. For some arbitrary $\lambda > 0$

$$\mathbb{P}\Big(S > (1 + \epsilon)\frac{k}{d}\Big) = \mathbb{P}\Big(\exp(\lambda S) > \exp(\lambda(1 + \epsilon)\frac{k}{d})\Big) \tag{42}$$

$$< \mathbb{E}[\exp(\lambda S)] \exp(-\lambda(1 + \epsilon)\frac{k}{d}) \tag{43}$$

Next using the fact that $\{[Ax]_i\}_{i=1}^k$ are iid the expectation can be factored as

$$\mathbb{E}[\exp(\lambda S)] = \mathbb{E}\Big(\prod_{i=1}^{k} \exp(\lambda [Ax]_i^2)\Big) \tag{44}$$

$$= \mathbb{E}[\exp(\lambda [Ax]_i^2)]^k. \tag{45}$$

Combining Equation 43 and Equation 45 along with Lemma 4 yields

$$\mathbb{P}\Big(S > (1 + \epsilon)\frac{k}{d}\Big) < \mathbb{E}[\exp(\lambda [Ax]_i^2]^k \exp(-\lambda(1 + \epsilon)\frac{k}{d}) \tag{46}$$

$$< \Big(\frac{1}{\sqrt{1 - 2\lambda/d}}\Big)^k \exp(-\lambda(1 + \epsilon)\frac{k}{d}) \tag{47}$$

Next, and similar to the JL Lemma proof I can choose $\lambda = \frac{d}{2}\frac{\epsilon}{1+\epsilon}$ and then bound $(1 + \epsilon)$ using Lemma 2 to get

$$\mathbb{P}\Big(S > (1 + \epsilon)\frac{k}{d}\Big) < ((1 + \epsilon)\exp(-\epsilon))^{k/2} \tag{48}$$

$$< \exp(-\frac{k}{4}(\epsilon^2 - \epsilon^3)). \tag{49}$$

For determining $\mathbb{P}(S < (1-\epsilon)\frac{k}{d})$ I use similar ideas, including Markov's inequality, and the fact $xe^x \geq x+x^2+x^3/2 \; \forall \; x \in \mathbb{R}$ to get

$$\mathbb{P}(S < (1-\epsilon)\frac{k}{d}) \tag{50}$$

$$< \Big(E\Big[1 - \lambda[Ax]_i^2 + \frac{(-\lambda[Ax]_i)^2}{2}\Big]\Big)^k \exp(\lambda(1-\epsilon)\frac{k}{d}) \tag{51}$$

$$= (1 - \frac{\lambda}{d} + \frac{\lambda^2}{2}\mathbb{E}[[Ax]_i^4])^k \exp(\lambda(1-\epsilon)\frac{k}{d}). \tag{52}$$

Using the bound on $\mathbb{E}[[Ax]_i^4]$ from Lemma 4 and choosing $\lambda = \frac{d}{2}\frac{\epsilon}{1+\epsilon}$ yields

$$\mathbb{P}(S < (1-\epsilon)\frac{k}{d}) \tag{53}$$

$$\leq \Big(1 - \frac{\lambda}{d} + \frac{3}{2}(\frac{\lambda}{d})^2\Big)^k \exp(\lambda(1-\epsilon)\frac{k}{d}) \tag{54}$$

$$= (1 - \frac{\epsilon}{2(1+\epsilon)} + \frac{3\epsilon^2}{8(1+\epsilon)^2})^k \exp(\frac{\epsilon(1-\epsilon)k}{2(1+\epsilon)}) \tag{55}$$

$$< \exp(-\frac{k}{2}(\epsilon^2/2 - \epsilon^3/3)) \tag{56}$$

where the final step comes from series expansion inequalities.

G. Further Extensions

So as witnessed, with database friendly random projections one is not limited to only Gaussian random variables to serve as entries in projection matrix. Interestingly the quality of these projections is similar to the Gaussian case. Other extensions have been proposed to make the projection matrix even sparser to decrease computational complexity. Notably, Li, Hastie, and Church [5] propose that the entries of the projection matrix can be generated as

$$a_{ij} = \begin{cases} \sqrt{s} & \text{with probability } \frac{1}{2s} \\ 0 & \text{with probability } 1 - \frac{1}{s} \\ -\sqrt{s} & \text{with probability } \frac{1}{2s}. \end{cases} \tag{57}$$

They find that using $s >> 3$, such as $s = \sqrt{d}$ or $s = \frac{d}{\log d}$ significantly speeds up the computation time of the algorithm, with only a small loss of accuracy. This is quite a remarkable result given it would imply for a $d = 1000$ problem one could set 99.7% of the entries of the projection matrix to 0.

V. Applications of JL Lemma - Compressed Least-Squares Regression [6]

Previous sections have shown that random projections facilitate reducing the dimension of data while preserving distances. This preservation of distances is important because in a sense it means the 'structure' of the data is preserved. Compressed Least-Squares Regression exploits this lower-dimensional preserved structure in the linear regression framework.

Linear regression lies at the heart of much of data analysis due to its simplicity and theoretical guarantees. However, linear regression methods can break down when d is large. On one hand, for large d linear regression has a tendency to overfit to the training data, and hence fail to generalize to unseen data. Even more concerning is that in the case of $d > n$ the least-squares solution for the regression problem is no longer unique, and therefore it is difficult to determine which of infinitely many models most faithfully represents the true data generating process. Random projections serve as one means to address these problems.

A. Problem Setup

Let $\{x_j, y_j\}_{j=1}^n$ where $x_j \in \mathcal{X}$ and $y_j \in \mathbb{R}$ be a dataset, and assume x_j i.i.d., $x_j \sim P_{\mathcal{X}}$, and $y_j = f^*(x_j) + \eta_j(x_j)$ where f^* is an unknown function and η_j is centred independent noise with variance $\sigma^2(x_j)$. For a specific class of functions $\mathcal{F}$ and $f \in \mathcal{F}$ the empirical quadratic error is

$$L_n(f) = \frac{1}{n}\sum_{j=1}^n (y_j - f(x_j))^2 \tag{58}$$

and the generalization quadratic error is

$$L(f) = \mathbb{E}_{(X,Y)\sim P}[(Y - f(X))^2]. \tag{59}$$

Clearly the goal is to discover a regression function $\hat{f} \in \mathcal{F}$ with the minimum generalization error (i.e. it is able to predict a new y based on a new x).

Now since y_j is the combination of both a deterministic $f^*(x_j)$ and random noise $\eta_k(x_j)$, even if we knew the exact f^* we would still have $L(f^*) > 0$. Therefore the correct measurement of the ability of $\hat{f}$ to generalize is its excess risk, defined as

$$L(\hat{f}) - L(f^*) = ||\hat{f} - f^*||_P^2. \tag{60}$$

This measure of excess risk can further be decomposed into the sum of estimation error $L(\hat{f}) - \inf_{f\in\mathcal{F}} L(f)$ and approximation error $\inf_{f\in\mathcal{F}} L(f) - L(f^*)$.

Denote a class of linear functions $\mathcal{F}_d$, which also can be viewed as the span of a set of d functions $\{\varphi_i\}_{1\leq i\leq d}$ called features. Therefore

$$\mathcal{F}_d = \{f_\alpha = \sum_{i=1}^d \alpha_i\varphi_i, \; \alpha \in \mathbb{R}^d\}. \tag{61}$$

Now in the 'standard' case where the size of the dataset n is larger than the number of features d the least squares regression solution $f_{\hat{\alpha}}$ minimizes empirical risk. And the expected estimation error can be bounded by

$$\mathbb{E}[L(f_{\hat{\alpha}}) - \inf_{f\in\mathcal{F}_d} L(f)] \leq c\sigma^2\frac{d\log n}{n} \tag{62}$$

where c a constant, $\sigma = \sup_{x\in\mathcal{X}}\sigma(x)$. Now for the case of the $d >> n$ problem, since there are so many features the approximation error is small (since $\mathcal{F}_d$ becomes large), but the bound in Equation 62 is not tight. Intuitively this occurs because $f_{\hat{\alpha}}$ will overfit to the data and will have limited ability to generalize.

B. Previous Approaches to Solving the $d >> n$ Problem

Typically the approach to finding an appropriate f is to apply a penalty term in the quadratic error function which will favour less complex functions

$$\hat{f} = \arg\min_{f \in \mathcal{F}_d} L_n(f) + \lambda ||f||_p^p. \tag{63}$$

Generally $p = 1$ (LASSO) and $\hat{f}$ is sparse, or $p = 2$ (ridge-regression).

Another approach is to take a sequence of function classes $\{\mathcal{G}_k\}_{k \geq 1}$ and then solve for $\hat{g}_k = \arg\min_{g \in \mathcal{G}_k, k \geq 1} L_k(g) +$ penalty(k, n). An example of this would be choosing a subsets of the k features and basing $\hat{g}$ off of only this subset, while penalizing each feature that is added.

The issues with these approaches is that one needs to search a potentially large function space $\mathcal{F}_d$ which could be computationally costly. As well, by penalizing the size of the regression coefficients one is introducing bias into the fitted model and having to select an unknown hyperparameter λ.

C. Overview of the Compressed Least-Squares Regression Approach

To avoid the potentially costly search over $\mathcal{F}_d$, compressed least-squares regression projects the data onto a lower dimensional subspace and then searches for the optimal $\hat{g}$ regression function within $\mathcal{G}_k \subset \mathcal{F}_d$ (where $k < d$). The key finding of the authors of this paper is that it is possible to bound the excess risk of this lower dimensional regression function. The use of the JL Lemma is quite central to bounding the approximation error (i.e. how 'good' $\mathcal{G}_k$ is at approximating $\mathcal{F}_n$), therefore I will examine how it is used in its proof. Before jumping into this proof, a corollary of the JL Lemma must be stated.

D. Corollary 1 - Preservation of Inner Products

Let $\{u_j\}_{1 \leq j \leq n}$ and v be vectors of $\mathbb{R}^d$. Let A be a $k \times d$ matrix of iid elements drawn from distributions $\mathcal{N}(0, 1/k)$, or those in Equations 33 and 34. For $\epsilon > 0, \delta > 0, k \geq \frac{1}{\epsilon^2/4 - \epsilon^3/6} \log \frac{4n}{\delta}$ we have that with probability of at least $1-\delta$ that for all $j \leq n$

$$|Au_j \cdot Av - u_j \cdot v| \leq \epsilon ||u_j|| ||v||. \tag{64}$$

E. Proof of Corollary 1

From a slightly modified version of Lemma 3 we have

$$\mathbb{P}(||Au||^2 \geq (1+\epsilon)||u||^2) \leq \exp(-k(\epsilon^2/4 - \epsilon^3/6)) \tag{65}$$
$$\mathbb{P}(||Au||^2 \leq (1-\epsilon)||u||^2) \leq \exp(-k(\epsilon^2/4 - \epsilon^3/6)). \tag{66}$$

Next take vectors u, w of norm 1 and apply the parallelogram law

$$4Au \cdot Aw = ||Au + Aw||^2 - ||Au - Aw||^2. \tag{67}$$

Then consider the following event

$$\mathbb{P}(4Au \cdot Aw \leq (1+\epsilon)||u+w||^2 - (1-\epsilon)||u-w||^2) \tag{68}$$
$$= \mathbb{P}(||Au + Aw||^2 - ||Au - Aw||^2 \leq (1+\epsilon)||u+w||^2 - (1-\epsilon)||u-w||^2) \tag{69}$$
$$= \mathbb{P}(||Au + Aw||^2 - ||Au - Aw||^2 \leq 4u \cdot w + 4\epsilon) \tag{70}$$

which using Lemma 3 shows the complement of Equation 69 (the probability the even fails) is bounded by $2\exp(-k(\epsilon^2/4 - \epsilon^3/6))$.

Therefore for each u_j

$$Au_j \cdot Av - u_j \cdot v \leq \epsilon ||u_j|| ||v|| \tag{71}$$

and since this inequality holds with similar probability for $-u_j, -v$, we get that

$$\mathbb{P}(|Au_j \cdot Av - u_j \cdot v| \leq \epsilon ||u_j|| ||v||) \leq 1 - 4n\exp(-k(\epsilon^2/4 - \epsilon^3/6)). \tag{72}$$

Corollary 1 then immediately follows.

F. Bounding the Low Dimension Approximation Error

Before proving the bound on the approximation error of the low dimension regression function I introduce some new notation. For a dataset of d features and $k < d$ the set of k compressed features $\{\psi_h\}_{1 \leq h \leq k}$ where $\psi_h(x) = \sum_{i=1}^N a_{h,i}\varphi_i(x)$ (recall $\varphi_i(x)$ is the i^{th} original feature), and $\psi(x)$ is a k-vector of features for a given x (i.e. $\psi(x) = A\varphi(x)$). With this, we can then define the compressed domain $\mathcal{G}_k = \{g_\beta = \sum_{h=1}^k \beta_h \psi_h,\ \beta \in \mathbb{R}^k\}$ (i.e. the set of compressed regression functions). Finally let $\alpha^+ = \arg\min_{\alpha \in \mathbb{R}^d} L(f_\alpha) - L(f^*)$ be the parameters associated with the best regression function in $\mathcal{F}_d$.

G. Theorem 2 - Approximation Error Bound

For any $\delta > 0$, $k \geq 15\log(8n/\delta)$ let A be a $k \times d$ matrix with entries as defined in Corollary 1, and $\mathcal{G}_k$ be the compressed domain from this A. Then with probability of at least $1 - \delta$

$$\inf_{g \in \mathcal{G}_k} ||g - f^*||_P^2 \leq \frac{8\log(8n/\delta)}{k} ||\alpha^+||^2 \Big(\mathbb{E}[||\varphi(X)||^2] + 2 \sup_{x \in \mathcal{X}} ||\varphi(x)||^2 \sqrt{\frac{\log 4/\delta}{2n}} \Big) + \inf_{f \in \mathcal{F}_d} ||f - f^*||_P^2 \tag{73}$$

H. Proof of Theorem 2

Let $f^+ = f_{\alpha^+} = \arg\min_{f \in \mathcal{F}_d} ||f - f^*||_P$ and $g^+ = g_{A\alpha^+}$. Then the approximation error can be bounded

$$\inf_{g \in \mathcal{G}_k} ||g - f^*||_P^2 \leq ||g^+ - f^*||_P^2 = ||g^+ - f^+||_P^2 + ||f^+ - f^*||_P^2 \tag{74}$$

which can be done since f^+ is the orgthogonal projection of f^* onto $\mathcal{F}_d$ and $g^+ \in \mathcal{F}_d$. Now to bound $||g^+ - f^+||_P^2$ define $Z(x) = A\alpha^+ \cdot A\varphi(x) - \alpha^+ \cdot \varphi(x)$, $\epsilon^2 = \frac{8}{k}\log(8n/\delta)$.

Then $k \geq 15\log(8n/\delta)$ leads to $\epsilon < \frac{3}{4}$ which implies $k \geq \frac{\log(8n/\delta)}{\epsilon^2/4-\epsilon^3/6}$. Then for some event $\mathcal{E}$ with probability of at least $1-\delta/2 = 1-\delta'$ for all $j \in 1,\ldots,n$

$$|Z(x_j)| \leq \epsilon||\alpha^+||||\varphi(x_j)|| \leq \epsilon||\alpha^+|| \sup_{x\in\mathcal{X}} ||\varphi(x)|| = C. \quad (75)$$

Then on this event $\mathcal{E}$

$$||g^+ - f^+||_P^2 = \mathbb{E}_{X\sim P_\mathcal{X}}[Z(X)^2] \quad (76)$$

$$\leq \frac{1}{n}\sum_{j=1}^{n}|Z(x_j)|^2 + C^2\sqrt{\frac{\log(2/\delta')}{2n}} \quad (77)$$

$$\leq \epsilon^2||\alpha^+||^2\Big(\frac{1}{n}\sum_{j=1}^{n}||\varphi(x_j)||^2 + \sup_{x\in\mathcal{X}}||\varphi(x)||^2\sqrt{\frac{\log(2/\delta')}{2n}}\Big) \quad (78)$$

$$\leq \epsilon^2||\alpha^+||^2\Big(\mathbb{E}[||\varphi(X)||^2] + 2\sup_{x\in\mathcal{X}}||\varphi(x)||^2\sqrt{\frac{\log(2/\delta')}{2n}}\Big) \quad (79)$$

where in the above the Chernoff-Hoeffding inequality was applied twice. Combining this result with Equation 74 yields the required bound.

I. Excess Risk Bound

The authors of this paper are then able to show that for a large enough n the expected excess risk of a truncated version of the low dimension regression function, $\hat{g}_L$ is

$$\mathbb{E}[||\hat{g}_L - f^*||_P^2] = O\Big(||\alpha^+||\sqrt{\mathbb{E}[||\varphi(X)||^2}\frac{\log n/\delta}{\sqrt{n}} + \inf_{f\in\mathcal{F}_d}||f-f^*||_P^2\Big). \quad (80)$$

And note in Equation 80 that the first term should tend towards 0 for large n. Also for large n the bound should approximately be

$$\mathbb{E}[||\hat{g}_L - f^*||_P^2] \leq 8\inf_{f\in\mathcal{F}_d}||f-f^*||_P^2. \quad (81)$$

This bound, given the high capacity of $\mathcal{F}_d$, provides a good bound for the excess risk of the low dimensional regression function and suggests that it will be able to approximate the ideal regression function f^*.

VI. Example of Forecasting Using Random Projections

An example of this type of regression in practice was conducted by Schneider and Gupta [7]. Here the authors attempted to model how well 231 tablet computers sold on Amazon during a 24-week period in 2012. Specifically, the authors model the sales rank (i.e. each tablet i has $\text{rank}_i \in [1, 231]$) of each type of tablet via a regression model. Several of the features included in the model are standard specifications of computers such as battery life, RAM, and hard drive capacity.

What necessitated the use of random projections in this study was the authors' use of customer review data in their model. The authors used a 'bag-of-words' approach where each review for a tablet was turned into a vector with entries representing word counts of certain words which appeared in the review (note if a word did not appear in a review it received a value of 0). The authors preprocessed this review data by filtering out words with little informational content, and then weighted each review by how many people found it helpful (a metric that is on Amazon for each review). After preprocessing the data, each week of sales for each tablet had a 20,068 feature vector representing word counts from reviews. This turned the problem into a $n = 231 \times 24 = 5544,\ d = 20,068 + 14 = 20,082$ problem, necessitating some form of dimensionality reduction.

The authors fit three models: a baseline model (not using words as covariates), a full model with $k = 300$ (i.e. this model has projected the $20,068$ word covariates down to 300 dimensions) modelling existing products, and a full model with $k = 50$ modelling new products (i.e. attempting to model a tablet's future sales when that tablet's past sales are not included in the training data). A substantial improvement in R^2, as well a decrease in out-of-sample mean absolute percentage error (MAPE), was reported for the full models as can be seen in Table I. The superiority of the full model over the baseline can be seen in Figure 1 where the baseline model produces many wildly inaccurate sales forecast compared to the full model.

TABLE I
RELATIVE MODEL PERFORMANCE

	k	R^2
Baseline	—	0.342
Full - Existing Products	300	0.862
Full - New Products	50	0.619

	k	MAPE
Baseline - Existing	—	163.5%
Full - Existing Products	300	37.2%
Baseline - New Products	—	288.6%
Full - New Products	50	243.9%

This demonstrates that the reviews' informational content is still preserved when it is projected down into a lower dimension, and is useful for forecasting. However, one limitation is seen in the ability to model new products. While the model incorporating word counts does lead to a decrease in out-of-sample MAPE for both new and existing products, the improvement is much smaller for new products. That being said, the fact there is still an improvement points to the full model, which includes 300 extra features, not overfitting on the reviews from the training set.

The results of this study show that even with a fairly naive approach (i.e. a bag-of-words approach) to incorporating review data into a regression model, a random projection based approach is able to improve the model's performance. It is reasonable to suggest that an approach which incorporates

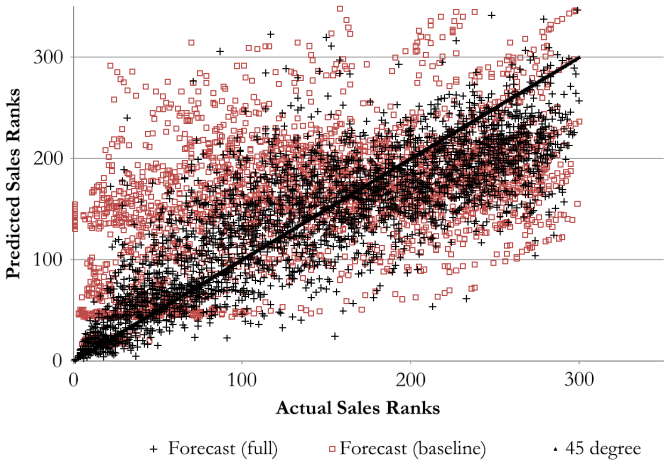

Fig. 1. Predicted vs. Actual Sales for Baseline and Full Model

review data in a more refined way (such as only including a subset of words that appeared in the reviews, or also including phrases) the dimension of the data could have been reduced meaningfully in the preprocessing step and led to an even more accurate model.

VII. Dimension Reduction for Data Visualization - UN Human Development Indicators

One of the benefits of random projection methods is they enable projecting data from high dimensions to one, two, or three dimensions which makes it much easier to visualize a dataset.

To illustrate this I have applied a random projection to a selection of Human Development Indicators from the United Nations [8]. The indicators I selected, which represent how developed a country is, include: dependency ratio, proportion urban population, mean years of school completed, CO^2 per capita, life expectancy, proportion internet users, proportion with basic drinking water, and unemployment rate. After transforming and normalizing the data I was left with these 8 indicators for 134 countries.

I then constructed a projection matrix using normal random variables and the dataset was projected to two dimensions. To allow for a cleaner visualization, in the accompanying chart I only included the top 10, middle 10, and bottom 10 countries based on their UN Human Development Index (an index which the UN produces to rank countries based on their level of development).

From Figure 2 it is clear that the random projection is able to preserve some semblance of the dataset's structure. Countries that are similar are grouped closely together, as well there appears to be an ordering where an increase in both Projected Feature 1 and Projected Feature 2 corresponds to an increase in how developed a country is.

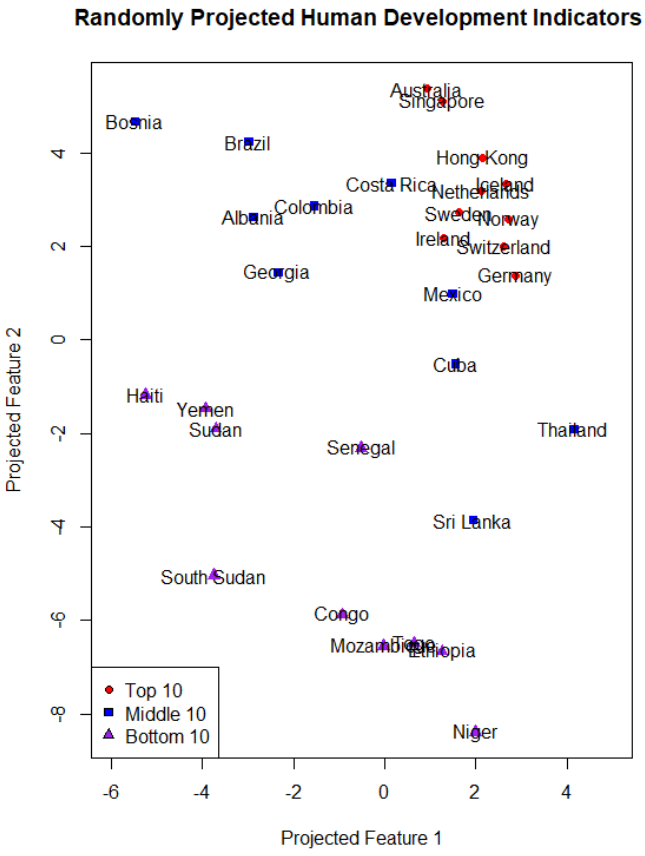

Fig. 2. Random projections applied to UN Human Development Indicators

VIII. Conclusion

It is clear that the JL Lemma serves as an incredibly powerful, and theoretically sound result which can be applied to dimension reduction problems. As discussed, random projections can be conducted in efficient 'database friendly' ways enabling their use on extremely large datasets. As well, random projections fit nicely into the regression framework and can lead to regression models which incorporate high dimensional data and generalize well. With the size and dimension of datasets continues to grow it is clear random projections will continue to play an important role in high dimensional data analysis.

References

[1] Johnson, William B., and Joram Lindenstrauss. "Extensions of Lipschitz mappings into a Hilbert space." *Contemporary Mathematics* 26.189-206 (1984): 1.

[2] Kakade, Sham and Shakhnarovich, Greg. "Random Projections." *University of Chicago CMSC35900 - Spring 2009*. https://ttic.uchicago.edu/~gregory/courses/LargeScaleLearning/lectures/jl.pdf. 2020/12/07.

[3] Mahoney, Michael. "The Johnson-Lindenstrauss Lemma." *Stanford University CS369 September 2009*. https://cs.stanford.edu/people mmahoney/cs369m/Lectures/lecture1.pdf

[4] Dimitris Achlioptas. 2001. Database-Friendly Random Projections. In *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS '01)*. Association for Computing Machinery, New York, NY, USA, 274–281. DOI:https://doi.org/10.1145/375551.375608.

[5] Li, Ping, Trevor J. Hastie, and Kenneth W. Church. “Very sparse random projections.” *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2006.
[6] Maillard, Odalric, and Rémi Munos. “Compressed Least-Squares Regression.” *Advances in Neural Information Processing Systems*. 2009.
[7] Schneider, Matthew J., and Sachin Gupta. “Forecasting Sales of New and Existing Products Using Consumer Reviews: A Random Projections Approach.” *International Journal of Forecasting* 32.2 (2016): 243-256.
[8] Human Development Data. United Nations Development Programme, http://hdr.undp.org/en/data. 2020/12/07.