# *Random Projections and Johnson-Lindenstrauss Lemma*

Uri Shaham

January 20, 2025

## 1 Introduction: Linear Projections

Assume we have a datapoint $x \in \mathbb{R}^d$, that we want to project onto a $p$-dimensional subspace of $\mathbb{R}^d$ spanned by vectors $\{u_1, \ldots, u_p\}$, with $p \ll d$. Let $U = [u_1, \ldots, u_p] \in \mathbb{R}^{d \times p}$. Let $\beta$ represent coefficients of the linear combination of the $u_i$'s, so the data reconstruction is $\hat{x} := U\beta$. Each such projection will have a residual $r = x - \hat{x}$, which will be smallest when $r \perp \text{span}\{u_1, \ldots, u_p\}$. Hence

$$U^T (x - U\beta) = 0 \Rightarrow \beta = \left(U^T U\right)^{-1} U^T x.$$

Note that this is also the formula for the least squares coefficients. Then $\hat{x} = U\beta = U\left(U^T U\right)^{-1} U^T x$. Note that if the vectors $\{u_1, \ldots, u_p\}$ are orthonormal (which makes $U$ an orthogonal matrix), then the formula simplifies to $\hat{x} = U\beta = UU^T x$, which is the same as reconstruction by PCA, for example.

### 1.1 Random Linear Projections

In PCA, for example, the matrix $U$ so that the vectors $\{u_1, \ldots, u_p\}$ are directions with maximal variance. However, we could also use a random $U$, i.e., not learn it at all. For example, by sampling its entries iid from a standard Gaussian. Surprisingly, random $U$ has good properties, in terms of distance preservation, despite the fact that is is totally independent of the data. The JL lemma, described next justifies this.

## 2 The Johnson Lindenstrauss Lemma

We first state a prove that random projection preserves norms:

**Lemma 2.1** (Norm preservation using RP)**.** *Let $x \in \mathbb{R}^d$ and let $A \in \mathbb{R}^{p \times d}$ random matrix with entries sampled iid from a $\mathcal{N}(0,1)$ distribution. Let $\epsilon \in (0, \frac{1}{2})$. Then*

$$\Pr\left((1-\epsilon)\|x\|^2 \le \left\|\frac{1}{\sqrt{p}}Ax\right\|^2 \le (1+\epsilon)\|x\|^2\right) \ge 1 - 2e^{-\frac{\left(\epsilon^2 - \epsilon^3\right)p}{4}}.$$

*Proof.* We first show that $\mathbb{E}\left[\left\|\frac{1}{\sqrt{p}}Ax\right\|^2\right] = \mathbb{E}\left[\|x\|^2\right]$. First, note that $\mathbb{E}\left[\left\|\frac{1}{\sqrt{p}}Ax\right\|^2\right] = \frac{1}{p}\,\mathbb{E}\left[\|Ax\|^2\right]$.

Next, we compute the expectation of the $i$'th entry $\mathbb{E}[[Ax]_i^2]$:

$$\begin{aligned}
\mathbb{E}[[Ax]_i^2] &= \mathbb{E}\left[\left(\sum_{j=1}^{d} A_{ij}x_j\right)^2\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{d}\sum_{j'=1}^{d} A_{ij}A_{ij'}x_jx_{j'}\right] \\
&= \sum_{j=1}^{d}\sum_{j'=1}^{d} x_jx_{j'}\,\mathbb{E}\left[A_{ij}A_{ij'}\right] \\
&= \sum_{j=1}^{d} x_j^2\,\mathbb{E}\left[A_{ij}^2\right] \\
&= \sum_{j=1}^{d} x_j^2 \\
&= \|x\|^2.
\end{aligned}$$

Therefore

$$\frac{1}{p}\,\mathbb{E}\left[\|Ax\|^2\right] = \|x\|^2.$$

Note that $[Ax]_i = \sum_{j=1}^{d} x_jA_{ij}$ is a normal random variable with zero mean and, by the above, $\|x\|^2$ variance. Hence $\tilde{Z}_i := \frac{[Ax]_i}{\|x\|}$ is a standard normal random variable, with $\tilde{Z}_i$ and $\tilde{Z}_k$ independent for $i \neq k$. Thus, we can bound the probability of faliure for one side:

$$\begin{aligned}
\Pr\left(\left\|\frac{1}{\sqrt{p}}Ax\right\|^2 \leq (1-\epsilon)\|x\|^2\right) &= \Pr\left(\sum_{i=1}^{p}\tilde{Z}_i^2 \leq (1-\epsilon)p\right) \\
&= \Pr\left(\chi_p^2 \leq (1-\epsilon)p\right) \\
&\leq \exp\left(-\frac{p}{4}\left(\epsilon^2-\epsilon^3\right)\right),
\end{aligned}$$

where the last transition is obtained using standard $\chi^2$ concentration bounds, which are stated in Lemma 2.2 and are not proved here. A similar argument will show that $\Pr\left(\left\|\frac{1}{\sqrt{p}}Ax\right\|^2 \geq (1+\epsilon)\|x\|^2\right) \leq \exp\left(-\frac{p}{4}\left(\epsilon^2-\epsilon^3\right)\right)$, which together prove the statement. □

**Lemma 2.2** ($\chi^2$ concentration bounds)**.**

$$\Pr\left(\chi_p^2 \leq (1-\epsilon)p\right) \leq \exp\left(-\frac{p}{4}\left(\epsilon^2-\epsilon^3\right)\right).$$

$$\Pr\left(\chi_p^2 \geq (1+\epsilon)p\right) \leq \exp\left(-\frac{p}{4}\left(\epsilon^2-\epsilon^3\right)\right).$$

We can now state and prove the Johnson-Lindenstrauss lemma.

**Lemma 2.3** (JL). *Let $\epsilon \in \left(0, \frac{1}{2}\right)$ and $Q \subset \mathbb{R}^d$ be a set of $n$ points,and let $p \geq \frac{12 \log n}{\epsilon^2}$. Then there exists a mapping $f : \mathbb{R}^d \to \mathbb{R}^p$ such that for all $v, u \in Q$,*

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

The proof is constructive (i.e., constructs $f$ and works by the probabilistic method[1], i.e., we prove that the probability that the desired $f$ exists is strictly greater than 0, hence it must exist. It utilizes the union bound, which says that for a set of events $\{A_1, A_2, \ldots\}$, $\Pr\left(\cup_i A_i\right) \leq \sum_i \Pr\left(A_i\right)$.

*Proof.* Let $f : x \mapsto \frac{1}{\sqrt{p}} Ax$, where $A \in \mathbb{R}^{p \times d}$ is a random matrix with iid $\mathcal{N}(0,1)$ entries. Then the probability that the statement in the lemma fails is

$$\begin{aligned}
&\Pr\left(\exists u, v \in Q : (1-\epsilon)\|v-u\|^2 > \|f(v)-f(u)\|^2 \text{ or } \|f(v)-f(u)\|^2 > (1+\epsilon)\|v-u\|^2\right) \\
&\leq \sum_{u,v \in Q} \Pr\left((1-\epsilon)\|v-u\|^2 > \|f(v)-f(u)\|^2\right) + \Pr\left(\|f(v)-f(u)\|^2 > (1+\epsilon)\|v-u\|^2\right) \\
&\leq 2n^2 \exp\left(-\frac{p}{4}\left(\epsilon^2 - \epsilon^3\right)\right), \qquad (1)
\end{aligned}$$

where the last step is obtained by the norm preservation lemma, applied to the vector $u - v$, and using the fact that the map $f$ is linear. Finally, as $p \geq \frac{12 \log n}{\epsilon^2}$ we have

$$2n^2 \exp\left(-\frac{p}{4}\left(\epsilon^2-\epsilon^3\right)\right) \leq 2n^2 \exp\left(-\frac{\frac{12\log n}{\epsilon^2}}{4}\left(\epsilon^2-\epsilon^3\right)\right) \leq 2n^2 \exp(-3\log n) < \frac{2}{n},$$

which is strictly less than 1 for 3 data points or more. Hence such a map must exist. ☐

A corollary of the norm preservation lemma shows that random projections preserve inner products as well.

**Corollary 2.4.** *Let $u, v \in \mathbb{R}^d$, with $\|u\|, \|v\| \leq 1$, and let $f : x \mapsto \frac{1}{\sqrt{p}} Ax$ be the JL transform as above. Then*

$$\Pr\left(|\langle u, v\rangle - \langle f(u), f(v)\rangle| > \epsilon\right) \leq 4 \exp\left(-\frac{p}{4}\left(\epsilon^2 - \epsilon^3\right)\right).$$

*Proof.* Applying the norm preservation lemma to the vectors $u+v, u-v$ we have that with probability at least $1 - 2\exp\left(-\frac{p}{4}\left(\epsilon^2-\epsilon^3\right)\right)$,

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

so

$$\begin{aligned}
4\langle f(u), f(v)\rangle &= \|f(u+v)\|^2 - \|f(u-v)\|^2 \\
&\geq (1-\epsilon)\|u+v\|^2 - (1+\epsilon)\|u-v\|^2 \\
&= 4\langle u, v\rangle - 2\epsilon(\|u\|^2 + \|v\|^2) \\
&\geq 4\langle u, v\rangle - 4\epsilon,
\end{aligned}$$

so $\langle f(u), f(v)\rangle \geq \langle u, v\rangle - \epsilon$. Similarly, we can get $\langle f(u), f(v)\rangle \leq \langle u, v\rangle + \epsilon$, and both events occur with probability at least $1 - 2\exp\left(-\frac{p}{4}\left(\epsilon^2-\epsilon^3\right)\right)$. Thus, by union bound, the probability of a failure is bounded by $4\exp\left(-\frac{p}{4}\left(\epsilon^2-\epsilon^3\right)\right)$. ☐

[1] `https://en.wikipedia.org/wiki/Probabilistic_method`

## 3 Application: Approximate Nearest Neighbor Search

Given a set of $n$ data points $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$, and a query point $y \in \mathbb{R}^d$, the goal of nearest neighbor search is to find $x_i$ which minimizes the distance $\|x_i - y\|$. A naive implementation of NN search has time complexity $O(nd)$, simply by computing all distances. However, in practice we often don't really need the exact nearest neighbors, and approximate neighbors suffice.

**Definition 3.1** ($\epsilon$-approximate nearest neighbor). *Given a query point $y$, $\epsilon$-approximate nearest neighbor search returns a point $x \in \mathcal{X}$ such that $\|x - y\| \le (1+\epsilon) \min_i \|x_i - y\|$.*

In practice, the approximate nearest neighbor is approached via one more reduction, to a near neighbor search.

**Definition 3.2** ($(\epsilon, r)$-approximate near neighbor search). *Given a query point $y$, and a nonnegative number $r$,$(\epsilon, r)$-approximate near neighbor search works as follows:*

- *If there exists $x \in \mathcal{X}$ with $\|x - y\| \le r$, it returns "Yes" and an index $i$ of a point such that $\|x_i - y\| \le (1+\epsilon)r$.*
- *If there does not exist $x \in \mathcal{X}$ with $\|x - y\| \le r$, it returns "No".*

To solve $\epsilon$-approximate nearest neighbor search using $(\epsilon, r)$-approximate near neighbor search, we can scale the data so that $\max_i \|x_i\| = \frac{1}{2}$, so the diameter (the distance between the two farthest points) is at most 1. We start from $\delta, k$ such that $\frac{1}{(1+\delta)^k}$ is sufficiently small, and run a sequence of $(\delta, r)$-approximate near neighbor searches with $r = \frac{1}{(1+\delta)^k}, \frac{1}{(1+\delta)^{k-1}}, \dots, 1$, and return $i$ corresponding to the minimum $r$ for which the answer is "Yes". Then we know that $\|x_i - y\| \le (1+\delta)r$. In addition, we know that $\min_i \|x_i - y\| > \frac{r}{1+\delta}$, hence altogether

$$\|x_i - y\| \le (1+\delta)r \le (1+\delta)^2 \min_i \|x_i - y\|.$$

That means we have solved $\epsilon$-approximate nearest neighbor search with $\epsilon = 2\delta + \delta^2$, and $k+1$ applications of $\epsilon$-approximate nearest neighbor search.

### 3.1 Solving $(\epsilon, r)$-approximate near neighbor search

**Preprocessing** We partition the space to $d$-dimensional cubes with side length $\frac{\epsilon r}{\sqrt{d}}$. The diameter of each cube is $\epsilon r$. Then for each point $x_i$ and cube $C$ such that intersects the $r$-ball $B(x_i, r)$ around $x_i$, we insert the (key, value) pair $(x_i, C)$ to a dictionary.
**Queries** Given a query point $y$, we find the cube $C$ which contains $y$. We then look for $C$ in the dictionary.

- If $C$ does not exist, then for each $x_i$, $\|x_i - y\| > r$, so we say "No".
- If $C$ is in the dictionary, we get an arbitrary point $x_i$ such that $B(x_i, r)$ intersects $C$. Then $\|y - x_i\| \le \epsilon r + r = (1+\epsilon)r$ (the distance is bounded by $r$ plus the diameter of the cube). Thus we say "Yes" and return $x_i$.

**Space analysis** The volume of $d$-dimensional ball of radius $r$ is approximately $2^{O(d)} r^n / d^{\frac{d}{2}}$. The volume of every cube is $(\epsilon r \sqrt{d})^d$. Thus each ball is intersected by approximately $\frac{2^{O(d)} r^n / d^{\frac{d}{2}}}{(\epsilon r \sqrt{d})^d} = O(1/\epsilon)^d$ cubes. Therefore the size of the dictionary is exponential in the dimension.
**Time analysis** based on the above, the time to build the dictionary is also $O(1/\epsilon)^d$. Finding the cube $C$ that contains $y$ takes $O(d)$ operations (we need to go over all coordinates), and then looking for $C$ in the dictionary is $O(1)$.

### 3.2 Improving performance using JL

By the JL lemma, we know that distances are approximately preserved under random projection to $O(\log n/\epsilon^2)$ dimensions, which is $O(\log n)$ assuming $\epsilon$ is constant. The time to apply the JL transform to all $n$ points is therefore $O(dn \log n)$. The dictionary space and time complexities are $(1/\epsilon)^{O(\log n)}$, which is linear. Query time is $d \log n$ to apply the JL transform to $y$, and $O(\log n)$ to find the cube of $y$.