

Lecture Notes on

The Johnson–Lindenstrauss Lemma

Patrick Shen
pts2125@columbia.edu

Disclaimer: These notes are based on [Alexandr Andoni](#) and [Edo Liberty](#)'s lectures on [Algorithms for LLMs](#) taught at Columbia University Fall 2026. The following typed notes were transcribed by GPT-6 Astra from my handwritten notes.

1 Statement

If we have $n \geq 2$ vectors $x_1, \dots, x_n \in \mathbb{R}^d$, we can approximate their pairwise distances $\|x_i - x_j\|$ with vectors $y_1, \dots, y_n \in \mathbb{R}^k$, where k can be much smaller than d . In fact, it can be shown that we can choose

$$k = O\left(\frac{\log n}{\varepsilon^2}\right),$$

where $0 < \varepsilon < 1$ is the acceptable relative error in the corresponding distances. That is, for every pair i, j ,

$$(1 - \varepsilon)\|x_i - x_j\| \leq \|y_i - y_j\| \leq (1 + \varepsilon)\|x_i - x_j\|.$$

Here and throughout, $\|\cdot\|$ is the Euclidean norm, and $\log$ denotes the natural logarithm.

Note that the bound grows only logarithmically with the number of vectors n , and does not depend on the original dimension d . This means that, even for very large d , we can approximate pairwise distances well using a relatively small number of coordinates.

Furthermore, preserving these distances, together with the vectors' norms, also lets us approximate pairwise inner products. We will come back to this after the proof.

2 Why do we care?

Preserving pairwise inner products allows us to approximate high-dimensional inner products more quickly. This speeds up things like nearest-vector search, matrix multiplication, etc.

3 Proof

Proof outline. Ultimately, we want to show that there is some way of projecting our vectors $x_1, \dots, x_n \in \mathbb{R}^d$ to $y_1, \dots, y_n \in \mathbb{R}^k$ so that pairwise distances are approximately preserved. We do this by explicitly constructing a random projection matrix and showing that it has this property. We rely on some useful properties of random Gaussian projections to construct such a mapping.

3.1 Useful fact #1

Consider a vector $w \in \mathbb{R}^d$ whose components are independent standard normal random variables:

$$w = \begin{pmatrix} w_1 \\ \vdots \\ w_d \end{pmatrix}, \quad w_i \stackrel{\text{iid}}{\sim} N(0, 1).$$

Take a fixed unit vector $u \in \mathbb{R}^d$, so that $\|u\| = 1$, and consider

$$w^\top u = (w_1 \quad \cdots \quad w_d) \begin{pmatrix} u_1 \\ \vdots \\ u_d \end{pmatrix} = \sum_{i=1}^d w_i u_i.$$

Let $z_i := w_i u_i$. Notice that, since $w_i \sim N(0, 1)$, multiplication by u_i scales the variance quadratically:

$$z_i = w_i u_i \sim N(0, u_i^2).$$

Since the z_i are independent, their sum is also Gaussian, with variance equal to the sum of their variances. Thus

$$\begin{aligned} w^\top u &= \sum_{i=1}^d w_i u_i = \sum_{i=1}^d z_i \\ &\sim N\left(0, \sum_{i=1}^d u_i^2\right) = N(0, \|u\|^2) = N(0, 1). \end{aligned}$$

If we view u as a unit direction that w is projected onto, the signed component of the projected vector can be thought of as being sampled from a standard Gaussian distribution. This should feel intuitive: the standard multivariate Gaussian distribution is rotationally symmetric, so it has no preferred direction.

Now, generalizing from our unit vector u to an arbitrary nonzero vector x , we have

$$\begin{aligned} w^\top x &= \|x\| w^\top \frac{x}{\|x\|} \\ &= \|x\| w^\top \hat{x}, \quad \hat{x} := \frac{x}{\|x\|}, \quad \|\hat{x}\| = 1. \end{aligned}$$

Thus

$$w^\top x \sim N(0, \|x\|^2).$$

For $x = 0$, the inner product is simply zero.

All that we have shown so far is that taking the inner product with a random vector constructed from standard Gaussians is the same, in distribution, as sampling from the univariate normal distribution $N(0, \|x\|^2)$.

Now, what has this got to do with dimensionality reduction?

3.2 Dimensionality reduction

Consider the mapping $\Pi : \mathbb{R}^d \rightarrow \mathbb{R}^k$, where

$$\Pi := \frac{1}{\sqrt{k}} R = \frac{1}{\sqrt{k}} \begin{bmatrix} \text{---} & w_1^\top & \text{---} \\ \text{---} & w_2^\top & \text{---} \\ & \vdots & \\ \text{---} & w_k^\top & \text{---} \end{bmatrix} \in \mathbb{R}^{k \times d}.$$

Here the w_i are independent Gaussian vectors in $\mathbb{R}^d$, each with independent $N(0, 1)$ components. We will show that this constructed matrix is a suitable candidate for our original problem.

Note that we think of Π as a matrix projecting a vector in $\mathbb{R}^d$ down to $\mathbb{R}^k$. Now, we will show an interesting property of an arbitrary vector $x \in \mathbb{R}^d$ and its norm after this projection. For $x \neq 0$, consider

$$\begin{aligned} \Pi x &= \|x\| \Pi \hat{x} && \text{(to work with a unit vector)} \\ &= \frac{\|x\|}{\sqrt{k}} R \hat{x} \\ &= \frac{\|x\|}{\sqrt{k}} \begin{bmatrix} \text{---} & w_1^\top & \text{---} \\ & \vdots & \\ \text{---} & w_k^\top & \text{---} \end{bmatrix} \begin{pmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_d \end{pmatrix} \\ &= \frac{\|x\|}{\sqrt{k}} \begin{pmatrix} z_1 \\ \vdots \\ z_k \end{pmatrix}, && z_i := w_i^\top \hat{x} \stackrel{\text{iid}}{\sim} N(0, 1). \end{aligned}$$

The last line uses our previous fact that $w_i^\top \hat{x} \sim N(0, 1)$, together with independence of the rows.

Now consider the squared norm:

$$\begin{aligned} \|\Pi x\|^2 &= \left\| \frac{\|x\|}{\sqrt{k}} \begin{pmatrix} z_1 \\ \vdots \\ z_k \end{pmatrix} \right\|^2 \\ &= \frac{\|x\|^2}{k} \sum_{i=1}^k z_i^2. \end{aligned}$$

Recognize that $\sum_{i=1}^k z_i^2$ follows a chi-square distribution with k degrees of freedom, since it is a sum of the squares of k independent standard normal random variables. Thus

$$\mathbb{E} \left[\sum_{i=1}^k z_i^2 \right] = k, \quad \mathbb{E} \left[\frac{1}{k} \sum_{i=1}^k z_i^2 \right] = 1,$$

and

$$\mathbb{E} [\|\Pi x\|^2] = \mathbb{E} \left[\frac{\|x\|^2}{k} \sum_{i=1}^k z_i^2 \right] = \|x\|^2.$$

This means that the expected squared length of the projected vector is the original squared length $\|x\|^2$. While this tells us something about the norm of x after it is projected, we can say more about how close we are to this expected answer as we vary k . Fundamentally, we are working with a random quantity, namely our matrix R , and we want to know how far off we are from the expected squared length, and how this deviation depends on k .

3.3 How close are we to the expected length?

To answer this, we use a concentration bound for the chi-square distribution. For $0 < \varepsilon < 1$,

$$\Pr \left[\left| \frac{1}{k} \sum_{i=1}^k z_i^2 - 1 \right| > \varepsilon \right] \leq 2e^{-k\varepsilon^2/8}. \quad (1)$$

This is a finite-sample tail bound; it can be obtained by a Chernoff bound using the chi-square moment-generating function.

In English, we are interested in the probability that the squared length of a unit vector $\hat{x} \in \mathbb{R}^d$, projected down to $\mathbb{R}^k$ via Π , differs from its expected squared length of 1 by more than some margin ε . When this event occurs, we think of it as a “failure.”

Using the known result above, this probability is upper bounded by $2e^{-k\varepsilon^2/8}$. Looking at this term, we see that as k grows, the failure probability becomes *exponentially small*.

Because of this exponential behavior, we can pick a “good” k for which the norm is approximately preserved. More rigorously, let $0 < \delta < 1$ be our desired upper bound on the failure probability. We want

$$\begin{aligned} 2e^{-k\varepsilon^2/8} &\leq \delta, \\ e^{-k\varepsilon^2/8} &\leq \frac{\delta}{2}, \\ -\frac{k\varepsilon^2}{8} &\leq \log\left(\frac{\delta}{2}\right), \\ k &\geq \frac{8 \log(2/\delta)}{\varepsilon^2}. \end{aligned}$$

So choosing k at least this large is sufficient. In asymptotic terms, a sufficiently large constant multiple of $\log(2/\delta)/\varepsilon^2$ works; we often write this requirement as

$$k = \Omega\left(\frac{\log(1/\delta)}{\varepsilon^2}\right).$$

With this choice, any fixed unit vector $\hat{x} \in \mathbb{R}^d$ has its squared norm preserved within $1 \pm \varepsilon$, with probability at least $1 - \delta$.

3.4 From one norm to pairwise distances

Note that this statement was made for a unit vector $\hat{x} \in \mathbb{R}^d$. To extend it to an arbitrary nonzero vector, first observe that preserving the squared norm also preserves the norm: on the success event,

$$1 - \varepsilon \leq \sqrt{1 - \varepsilon} \leq \|\Pi\hat{x}\| \leq \sqrt{1 + \varepsilon} \leq 1 + \varepsilon.$$

Thus $\|\Pi\hat{x}\| \approx 1$, and

$$\begin{aligned} \|\Pi x\| &= \left\| \|x\| \Pi \frac{x}{\|x\|} \right\| \\ &= \|x\| \left\| \Pi \frac{x}{\|x\|} \right\| \\ &\approx \|x\|. \end{aligned}$$

More precisely,

$$(1 - \varepsilon)\|x\| \leq \|\Pi x\| \leq (1 + \varepsilon)\|x\|,$$

with probability at least $1 - \delta$ for each fixed x . Again, $x = 0$ is immediate.

While this shows that a vector's norm is approximately preserved through Π , we can easily turn it into a statement about preserving pairwise distances between $x_i, x_j \in \mathbb{R}^d$. Consider the difference vector

$$x := x_i - x_j.$$

By linearity,

$$\|\Pi(x_i - x_j)\| = \|\Pi x_i - \Pi x_j\|.$$

Define

$$y_i := \Pi x_i \in \mathbb{R}^k, \quad y_j := \Pi x_j \in \mathbb{R}^k.$$

Then

$$\|y_i - y_j\| \approx \|x_i - x_j\|.$$

While this shows that the distance between a fixed pair of vectors in $\mathbb{R}^d$ is approximately preserved, our original problem is about all pairs of n vectors. We will now show how to pick our target dimension k so that this property holds simultaneously for all $\binom{n}{2}$ pairs.

3.5 Preserving all pairs

Let f_{ij} be the event of a “failed” projection for the pair $x_i, x_j \in \mathbb{R}^d$: the projected squared distance deviates from its expectation by more than the allowed relative margin ε . Explicitly,

$$f_{ij} := \{ \left| \|\Pi(x_i - x_j)\|^2 - \|x_i - x_j\|^2 \right| > \varepsilon \|x_i - x_j\|^2 \}.$$

If the vectors are identical, this event is empty. Otherwise, the concentration bound gives $\Pr[f_{ij}] \leq \delta$ for our choice of k .

There are $\binom{n}{2}$ pairs of vectors, but we will use the fact that $\binom{n}{2} < n^2$ to make the bound easier to see. We are interested in the probability that all pairs have their distances approximately preserved. We can bound its complement, the event that *some* pair fails, using the union bound:

$$\Pr \left[\bigcup_{i < j} f_{ij} \right] \leq \sum_{i < j} \Pr[f_{ij}] \leq \binom{n}{2} \delta \leq n^2 \delta.$$

Essentially,

$$\Pr[\text{some failure}] \leq n^2 \cdot (\text{individual failure bound}).$$

The pairwise failure events do not need to be independent for this step.

Now, let's say that we are comfortable upper bounding the total failure probability by some constant $0 < \delta_F < 1$. It is enough to have

$$n^2 \delta \leq \delta_F, \quad \text{or equivalently} \quad \delta \leq \frac{\delta_F}{n^2}.$$

Again, this is basically saying that each individual failure probability should be at most our total failure budget divided by n^2 .

Plugging $\delta = \delta_F/n^2$ into our earlier bound, it suffices to take

$$\begin{aligned} k &\geq \frac{8 \log(2/\delta)}{\varepsilon^2} \\ &= \frac{8 \log(2n^2/\delta_F)}{\varepsilon^2} \\ &= \frac{16 \log\left(n\sqrt{2/\delta_F}\right)}{\varepsilon^2}. \end{aligned}$$

For fixed δ_F , this gives the requirement

$$k = \Omega\left(\frac{\log n}{\varepsilon^2}\right),$$

meaning that k must be at least a sufficiently large constant multiple of $\log n/\varepsilon^2$ for this argument to guarantee success.

However, since we are free to choose k , we can simply choose it to be the smallest integer above this threshold. Thus there exists a suitable projection with

$$k = O\left(\frac{\log n}{\varepsilon^2}\right).$$

With probability at least $1 - \delta_F$, all pairwise squared distances are approximately preserved, and hence so are all pairwise distances. Since this probability is positive, a suitable projection exists.

This means that, even as n grows large, we can find a relatively small k that approximately preserves distances. The margin ε controls how much the projected distances may deviate, and δ_F controls the probability that any pair fails. This is the statement we wanted to show.

4 Preserved norms imply preserved inner products

The connection is the identity

$$\langle a, b \rangle = \frac{1}{2} (\|a\|^2 + \|b\|^2 - \|a - b\|^2).$$

So if we preserve these three squared norms, we also approximate the inner product. To ensure the individual norms are preserved as well as pairwise distances, we can include the origin in our set of points.

More precisely, if each of these squared norms is preserved within a relative error ε , the identity gives the additive bound

$$|\langle \Pi a, \Pi b \rangle - \langle a, b \rangle| \leq \frac{\varepsilon}{2} (\|a\|^2 + \|b\|^2 + \|a - b\|^2).$$

This is an additive approximation to the inner product; an inner product near zero need not have a small relative error.

5 Applications to matrix multiplication

Matrix multiplication can be viewed as a bunch of inner-product computations. Write

$$A = \begin{bmatrix} \text{---} & a_1^\top & \text{---} \\ & \vdots & \\ \text{---} & a_n^\top & \text{---} \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad B = \begin{bmatrix} | & & | \\ b_1 & \cdots & b_m \\ | & & | \end{bmatrix} \in \mathbb{R}^{d \times m}.$$

Then

$$AB = \begin{pmatrix} \langle a_1, b_1 \rangle & \cdots & \langle a_1, b_m \rangle \\ \vdots & \ddots & \vdots \\ \langle a_n, b_1 \rangle & \cdots & \langle a_n, b_m \rangle \end{pmatrix} \in \mathbb{R}^{n \times m}.$$

Computing all of these inner products directly takes $O(ndm)$ time.

Now, since $a_i, b_j \in \mathbb{R}^d$, we can use Π to approximate $\langle a_i, b_j \rangle$ with $\langle a_i^*, b_j^* \rangle$, where

$$a_i^* := \Pi a_i, \quad b_j^* := \Pi b_j \in \mathbb{R}^k, \quad \Pi \in \mathbb{R}^{k \times d}.$$

We choose one projection for the combined set of row vectors, column vectors, and the origin, so that the preceding inner-product bound applies to all entries. The useful regime is $k \ll d$.

We can therefore approximate AB with

$$\underbrace{(\Pi A^\top)^\top}_{n \times k} \underbrace{(\Pi B)}_{k \times m} = A \Pi^\top \Pi B.$$

Using a dense Gaussian projection, forming the two projected matrices and multiplying them takes

$$\begin{aligned} O(kdn + kdm + knm) &= O(kd(n + m) + knm) \\ &= O(k[d(n + m) + nm]). \end{aligned}$$

When k is small enough relative to the matrix dimensions, this can be faster than the direct computation.

6 Fast JL

Our construction of the projection matrix was

$$\Pi = \frac{1}{\sqrt{k}} R \in \mathbb{R}^{k \times d},$$

where every entry of R is sampled independently from a standard normal distribution. But we can also construct other projection matrices using Hadamard matrices and random signs.

Consider the mixing step HD , where

$$D = \text{diag}(\sigma_1, \dots, \sigma_d), \quad \Pr[\sigma_i = +1] = \Pr[\sigma_i = -1] = \frac{1}{2},$$

with independent signs. Thus D is a diagonal matrix with a random choice of $+1$ or -1 on each diagonal entry.

The Hadamard matrix is constructed recursively as

$$H_d = \begin{pmatrix} H_{d/2} & H_{d/2} \\ H_{d/2} & -H_{d/2} \end{pmatrix}, \quad H_1 = (1).$$

Here d is a power of two; otherwise we can pad vectors with zeros to the next power of two. This recursion gives the unnormalized matrix, so $H_d^\top H_d = dI$ and $H_d/\sqrt{d}$ is orthogonal.

The product HD mixes the coordinates but still maps $\mathbb{R}^d$ to $\mathbb{R}^d$. To reduce the dimension, we also need a sampling or projection step. For example, with suitable random row sampling $S \in \mathbb{R}^{k \times d}$, a subsampled Hadamard map has the form

$$\Pi_{\text{fast}} = \sqrt{\frac{d}{k}} S \frac{H_d}{\sqrt{d}} D = \frac{1}{\sqrt{k}} S H_d D.$$

The dimension needed for this structured construction requires its own concentration analysis; it does not follow directly from the Gaussian proof above.

Using the fast Walsh–Hadamard transform, we can apply the Hadamard mixing step to n vectors in

$$O(nd \log d)$$

time.

More sophisticated constructions are known as *Fast Johnson–Lindenstrauss Transforms* (FJLTs). The classical FJLT of Ailon and Chazelle combines randomized Fourier or Hadamard preconditioning with a sparse projection in order to apply a Johnson–Lindenstrauss embedding faster than a dense random projection. The exact running time depends on the particular construction and parameter regime.

7 There are an exponential number of "almost orthogonal" vectors in dimensional reduction

The Johnson–Lindenstrauss lemma has a useful geometric consequence. Although $\mathbb{R}^k$ contains at most k mutually orthogonal nonzero vectors, it can contain exponentially many vectors that are approximately orthogonal.

For every $0 < \varepsilon < 1$, there exists a constant $c > 0$ such that $\mathbb{R}^k$ contains at least

$$\exp(c\varepsilon^2 k)$$

unit vectors $u_1, \dots, u_N$ satisfying

$$|\langle u_i, u_j \rangle| \leq \varepsilon \quad \text{for all } i \neq j.$$

Proof. Let

$$N := \lfloor \exp(c\varepsilon^2 k) \rfloor,$$

where $c > 0$ is a sufficiently small absolute constant.

Consider the N standard basis vectors

$$e_1, \dots, e_N \in \mathbb{R}^N.$$

These vectors are exactly orthogonal, and for $i \neq j$,

$$\|e_i\| = 1, \quad \langle e_i, e_j \rangle = 0, \quad \|e_i - e_j\| = \sqrt{2}.$$

Apply the Johnson–Lindenstrauss lemma to

$$0, e_1, \dots, e_N$$

with distortion parameter chosen to be a sufficiently small constant multiple of ε .

JL tells us that these points can be embedded into dimension

$$O\left(\frac{\log N}{\varepsilon^2}\right).$$

But by our choice of N ,

$$\log N = O(\varepsilon^2 k),$$

so

$$O\left(\frac{\log N}{\varepsilon^2}\right) = O(k).$$

Choosing c sufficiently small ensures that the required target dimension is at most k . Hence there exists a linear map

$$\Pi : \mathbb{R}^N \rightarrow \mathbb{R}^k$$

that approximately preserves the norms and pairwise distances of all these vectors.

Let

$$y_i := \Pi e_i.$$

Since the origin was included, each $\|y_i\|$ is close to 1, and since

$$\|e_i - e_j\| = \sqrt{2},$$

each

$$\|y_i - y_j\|$$

is close to $\sqrt{2}$.

Using the identity

$$\langle y_i, y_j \rangle = \frac{1}{2} (\|y_i\|^2 + \|y_j\|^2 - \|y_i - y_j\|^2),$$

it follows that

$$|\langle y_i, y_j \rangle| = O(\varepsilon).$$

After normalizing each y_i to a unit vector

$$u_i := \frac{y_i}{\|y_i\|},$$

the inner products change only by another constant factor. By choosing the JL distortion to be a sufficiently small constant multiple of ε , we obtain

$$|\langle u_i, u_j \rangle| \leq \varepsilon.$$

Thus we have constructed

$$N = \exp(\Omega(\varepsilon^2 k))$$

pairwise approximately orthogonal unit vectors in $\mathbb{R}^k$.