1 Fast Johnson-Lindenstrauss embeddings

1.1 Recap: J-L lemma

Lemma 1. For any \( \varepsilon \in (0, 1) \), \( \{x_1, \ldots, x_n\} \subset \mathbb{R}^d \), \( k \in \mathbb{N} \) such that \( k \geq C \log(n)/\varepsilon^2 \), there is a matrix \( A \in \mathbb{R}^{k \times d} \) such that

\[
(1 - \varepsilon) \|x_i - x_j\|_2^2 \leq \|Ax_i - Ax_j\|_2^2 \leq (1 + \varepsilon) \|x_i - x_j\|_2^2.
\]

Moreover, such a matrix can be found in randomized polynomial time.

**Crux of proof:** Specify a distribution for random matrix \( A \in \mathbb{R}^{k \times d} \) such that for all \( u \in \mathbb{S}^{d-1} \),

\[
P\left( \|Au\|_2^2 - 1 \geq \varepsilon \right) \leq \exp\left( -\Omega(k\varepsilon^2) \right).
\]

If \( A = (A_{i,j}) \) consists of \( A_{i,j} \sim \mathcal{N}(0, 1/k) \), then \( A \) is dense (all entries are non-zero) with probability 1. Therefore straightforward matrix multiplication with \( x \in \mathbb{R}^d \) takes \( O(k\|x\|_0) \) operations, where \( \|x\|_0 \) is the number of non-zero entries in \( x \).

Today: If \( x \) is dense (i.e., \( \|x\|_0 = \Omega(d) \)), then it is possible to get a running time of roughly \( \tilde{O}(k + d) \) (ignoring logarithmic factors). This is a result due to Ailon and Chazelle (2006).

**Homework:** If \( x \) is sparse, a different technique roughly leads to a running time of roughly \( \tilde{O}(\sqrt{k}\|x\|_0) \).

1.2 Bernoulli-Gaussian random matrix

We’ll sparsify the dense \( \mathcal{N}(0, 1) \) construction by randomly setting some entries to zero. Fix some \( q \in (0, 1) \), and define i.i.d. random variables

\[
B_{i,j} \sim \text{Bernoulli}(q) \quad Z_{i,j} \sim \mathcal{N}(0, 1)
\]

for \( i \in [k], j \in [d] \). Define the matrix \( S = (S_{i,j}) \in \mathbb{R}^{k \times d} \) by

\[
S_{i,j} := \frac{1}{\sqrt{q}} B_{i,j} Z_{i,j}
\]

and set

\[
A := \frac{1}{\sqrt{k}} S.
\]

On average, this matrix has \( qkd \) non-zero entries. So if \( q = O(1/d) \) or so, then the running time for computing \( Ax \) is \( O(k) \). This is a considerable improvement when \( x \) is dense.

Does this distribution for \( A \) satisfy (1)? Let the random vector \( Y = (B_1 Z_1, \ldots, B_d Z_d)/\sqrt{q} \) be a copy of one of the rows of \( S \). We saw last time that it suffices to require
1. $\mathbb{E}(Y) = 0$

2. $\mathbb{E}(YY^\top) = I_d$, and

3. $Y$ is $O(1)$-subgaussian.

The first one is clearly satisfied: the $B_i$ and $Z_i$ are independent, so every entry of $S$ has mean zero (since $\mathbb{E}(Z_i) = 0$). The second requirement is satisfied because of the $1/\sqrt{q}$ factor: for any unit vector $u \in \mathbb{S}^{d-1}$,

$$
\text{var}(u^\top Y) = \sum_{i=1}^{d} \mathbb{E} \left( \frac{B_i Z_i u_i}{\sqrt{q}} \right)^2 = \sum_{i=1}^{d} \frac{\mathbb{E}(B_i^2)\mathbb{E}(Z_i^2)u_i^2}{q} = \sum_{i=1}^{d} u_i^2 = 1.
$$

Unfortunately, the last requirement is not true unless $q = \Omega(1)$. To see this, we use a moment argument via the following claim (which is easy to prove).

**Claim 1.** Every zero-mean $v$-subgaussian random variable $Y$ satisfies $v \geq (1/4)\sqrt{\mathbb{E}Y^4}$.

Set $u := (1, 0, \ldots, 0)$; and now compute the fourth moment of $u^\top Y$:

$$
\mathbb{E}(u^\top Y)^4 = \mathbb{E} \left( \frac{1}{\sqrt{q}} \sum_{i=1}^{d} B_i Z_i u_i \right)^4 = \frac{1}{q^2} \mathbb{E}(B_1) \mathbb{E}(Z_1^4) = \frac{3}{q}.
$$

Therefore $Y$ can only be $O(1)$-subgaussian if $q = \Omega(1)$.

The fourth moment of $u^\top Y$ can also be thought of as the variance of $(u^\top Y)^2$ (up to an additive constant). Since $\|Au\|_2^2$ is the empirical average of independent random variables each with variance $\Omega(1/q)$, we again see the requirement for $q = \Omega(1)$ because otherwise the empirical average would not converge to its expectation with high probability.

### 1.3 Sparse vs. dense vectors

Why did the Bernoulli-Gaussian construction fail?

- Suppose $u$ only has a few (say, $O(1)$) non-zero entries. If $q = o(1)$, then

$$
\mathbb{P}(u^\top Y = 0) \geq (1 - o(1))^{O(1)} = \Omega(1).
$$

The lower bound is gotten from the probability that $Y$ is zero in the exactly the same locations as where $u$ is non-zero.

This suggests that *sparse* $u$ are problematic.

- What about the opposite of a sparse vector? A dense vector (*i.e.*, almost all entries are non-zero)? Not quite, because one can have very tiny values in most entries, and a few very large entries.

What we really want is to prevent the “energy” of $u$ from being concentrated in a few coordinates. We can do this by requiring every entry in $u$ to be small: a bound on $\|u\|_\infty$. 

2
• Note that a uniform random unit vector $U$ satisfies

$$\|U\|_\infty = O\left(\sqrt{\frac{\log(d/\delta)}{d}}\right)$$

with probability at least $1 - \delta$. So in this sense, “typical” unit vectors $u$ are dense in the sense of having $\|u\|_\infty = O(\sqrt{\log(d)/d})$.

• How does the construction do for unit vectors $u$ with small $\|u\|_\infty$? Let us do the moment calculation again:

$$\mathbb{E}(u^\top Y)^4 = \frac{1}{q^2} \mathbb{E}\left(\sum_{i=1}^d B_i Z_i u_i \right)^4$$

$$= \frac{1}{q^2} \left( \sum_{i=1}^d \mathbb{E}(B_i Z_i u_i)^4 + 3 \sum_{i \neq j} \mathbb{E}(B_i Z_i u_i)^2 (B_j Z_j u_j)^2 \right)$$

$$= \frac{1}{q^2} \left( \sum_{i=1}^d 3q u_i^4 + 3 \sum_{i \neq j} q^2 u_i^2 u_j^2 \right) = 3 \left( \frac{1}{q} - 1 \right) \|u\|_4^4 + 3.$$

By Hölder’s inequality,

$$\|u\|_4^4 \leq \|u\|_2^2 \|u\|_\infty = \|u\|_\infty^2.$$

Therefore, $q$ can be as small as $\|u\|_\infty^2$ and still have the fourth moment be bounded. If $\|u\|_\infty = O(\sqrt{\log(d)/d})$, then this means $q$ can be as small as $O(\log(d)/d)$, which would lead to a computational complexity of $O(k \log(d))$ for multiplying $A$ by a vector $x$.

### 1.4 Analysis for dense vectors

Recall that $A = \frac{1}{\sqrt{k}} S$ where $S_{i,j} = \frac{1}{\sqrt{q}} B_{i,j} Z_{i,j}$. Here is a slick analysis based on conditioning on the $B_{i,j}$’s. For now, assume the $B_{i,j}$’s are fixed. Then

$$\|Au\|_2^2 = \sum_{i=1}^k \left( \frac{1}{\sqrt{kq}} \sum_{j=1}^d B_{i,j} Z_{i,j} u_j \right)^2.$$

Each term in the outer sum is the square of a linear combination of independent standard normal random variables:

$$X_i := \frac{1}{\sqrt{kq}} \sum_{j=1}^d B_{i,j} u_j Z_{ij} \sim \mathcal{N}\left(0, \frac{1}{kq} \sum_{j=1}^d B_{i,j}^2 u_j^2 \right)$$

and

$$\|Au\|_2^2 = \sum_{i=1}^k X_i^2.$$

Let $Q_1, Q_2, \ldots, Q_k \overset{i.i.d.}{\sim} \chi^2(1)$, and define

$$v_i := \frac{1}{kq} \sum_{j=1}^d B_{i,j}^2 u_j^2, \quad \forall i \in [k].$$
Then $\frac{1}{k} \sum_{i=1}^{k} X_i^2$ (and hence $\|Au\|_2^2$) has the same distribution as
\[ \sum_{i=1}^{k} v_i Q_i. \]
We now apply the subgamma Chernoff bound. Recall that $X$ is $(v, c)$-subgamma if for all $\lambda \in (0, 1/c)$,
\[ \psi_{X - EX}(\lambda) \leq \frac{v \lambda^2}{2(1 - c\lambda)}. \]

**Lemma 2** (Subgamma Chernoff bound). Let $X_1, X_2, \ldots, X_n$ be independent random variables, and $X_i$ be $(v_i, c_i)$-subgamma for each $i \in [n]$. Then, for $v := \sum_{i=1}^{n} v_i$, $c := \max_{i \in [n]} c_i$, and $Y := \sum_{i=1}^{n} (X_i - EX_i)$,
\[ P(Y \geq t) \leq \exp\left( -\frac{v c^2}{2} \cdot h_1\left( \frac{ct}{v} \right) \right) \leq \exp\left( -\frac{t^2}{2(v + ct)} \right). \]
In other words, with probability at least $1 - \delta$,
\[ Y \leq \sqrt{2v \ln(1/\delta)} + c \ln(1/\delta). \]

We saw previously that $Q_i$ is $(2, 2)$-subgamma, and therefore $\sum_{i=1}^{k} v_i Q_i$ is
\[ \left( 2 \sum_{i=1}^{k} v_i^2, 2 \max_{i \in [k]} v_i \right) - \text{subgamma}. \]
The same holds for $-\sum_{i=1}^{k} v_i Q_i$. Using the subgamma Chernoff bound with a union bound, with probability at least $1 - \delta$,
\[ \left| \sum_{i=1}^{k} v_i (Q_i - 1) \right| \leq 2 \sqrt{\sum_{i=1}^{k} v_i^2 \log(2/\delta) + 2 \max_{i \in [k]} v_i \log(2/\delta)}. \]
Note that the summation inside the square-root can be bounded as
\[ \sum_{i=1}^{k} v_i^2 \leq \left( \sum_{i=1}^{k} v_i \right) \left( \max_{i \in [k]} v_i \right). \]
Therefore, the bound we have is
\[ \left| \sum_{i=1}^{k} v_i (Q_i - 1) \right| \leq 2 \sqrt{\left( \sum_{i=1}^{k} v_i \right) \left( \max_{i \in [k]} v_i \right) \log(2/\delta) + 2 \max_{i \in [k]} v_i \log(2/\delta)}. \]

We need to bound $\sum_{i=1}^{k} v_i$ and $\max_{i \in [k]} v_i$. We will show that
\[ 1 - O(\varepsilon) \leq \sum_{i=1}^{k} v_i \leq 1 + O(\varepsilon) \quad \text{and} \quad \max_{i \in [k]} v_i \log(2/\delta) \leq O(\varepsilon^2). \]
Recall that the $v_i$ depend on the $B_i$, which we were holding fix. We will bound them probabilistically using Bernstein’s inequality.
Lemma 3 (Bernstein’s inequality, bounded case). Let $X_1, X_2, \ldots, X_n$ be independent random variables with $X_i - \mathbb{E}X_i \leq b$ for all $i \in [n]$, $v := \sum_{i=1}^{n} \text{var}(X_i)$, and $S_n := \sum_{i=1}^{n} X_i$. Then

$$\mathbb{P}(S_n \geq \mathbb{E}S_n + t) \leq \exp\left(-\frac{v}{(b/3)^2} \cdot h_1\left(\frac{(b/3)t}{v}\right)\right).$$

In other words, with probability at least $1 - \delta$,

$$S_n \leq \mathbb{E}S_n + \sqrt{2v \ln(1/\delta)} + \frac{b \ln(1/\delta)}{3}.\quad \text{(Bernstein’s inequality, bounded case)}$$

Recall

$$v_i := \sum_{j=1}^{d} \frac{u_j^2}{kq} B_{i,j}, \quad \forall i \in [k].$$

We know

$$\mathbb{E}\left(\frac{u_j^2}{kq} B_{i,j}\right) = \frac{u_j^2}{k}, \quad \text{var}\left(\frac{u_j^2}{kq} B_{i,j}\right) = \frac{u_j^4}{k^2q} q(1-q) \leq \frac{u_j^2 \|u\|_\infty^2}{k^2q},$$

and

$$\left|\frac{u_j^2}{kq} B_{i,j} - \mathbb{E}\left(\frac{u_j^2}{kq} B_{i,j}\right)\right| \leq \frac{\|u\|_\infty^2}{kq}.$$

This implies

$$\text{var}(v_i) = \sum_{j=1}^{d} \text{var}\left(\frac{u_j^2}{kq} B_{i,j}\right) \leq \frac{\|u\|_\infty^2}{k^2q},$$

$$\text{var}\left(\sum_{i=1}^{k} v_i\right) = \sum_{i=1}^{d} \text{var}(v_i) \leq \frac{\|u\|_\infty^2}{kq}.$$

Therefore, with probability at least $1 - \delta$,

$$\max_{i \in [k]} v_i \leq 1 + \frac{\sqrt{2(\|u\|_\infty^2/q) \ln(3k/\delta)} + (\|u\|_\infty^2/q) \ln(3k/\delta)/3}{k},$$

and

$$\left|\sum_{i=1}^{k} v_i - 1\right| \leq \frac{\sqrt{2(\|u\|_\infty^2/q) \ln(3k/\delta)} + (\|u\|_\infty^2/q) \ln(3k/\delta)}{3k}.$$

Let’s assume $k = \Theta(\log(1/\delta)/\varepsilon^2)$. Then it can be checked that $q$ should satisfy

$$q \geq O(\|u\|_\infty^2 \ln(3k/\delta))$$

to guarantee that with probability at least $1 - \delta$,

$$1 - O(\varepsilon) \leq \sum_{i=1}^{k} v_i \leq 1 + O(\varepsilon) \quad \text{and} \quad \max_{i \in [k]} v_i \log(2/\delta) \leq O(\varepsilon^2).$$

This in turn implies that with probability at least $1 - O(\delta)$,

$$\|Au\|_2^2 - 1 \leq O(\varepsilon).$$

So if $\|u\|_{\infty} \leq O(\sqrt{\log(d)/d})$, we obtain a running time of $O(k \log(k) \log(d))$, which is usually much better than $O(kd)$. 

So if $\|u\|_{\infty} \leq O(\sqrt{\log(d)/d})$, we obtain a running time of $O(k \log(k) \log(d))$, which is usually much better than $O(kd)$. 

So if $\|u\|_{\infty} \leq O(\sqrt{\log(d)/d})$, we obtain a running time of $O(k \log(k) \log(d))$, which is usually much better than $O(kd)$. 

So if $\|u\|_{\infty} \leq O(\sqrt{\log(d)/d})$, we obtain a running time of $O(k \log(k) \log(d))$, which is usually much better than $O(kd)$. 

So if $\|u\|_{\infty} \leq O(\sqrt{\log(d)/d})$, we obtain a running time of $O(k \log(k) \log(d))$, which is usually much better than $O(kd)$. 

So if $\|u\|_{\infty} \leq O(\sqrt{\log(d)/d})$, we obtain a running time of $O(k \log(k) \log(d))$, which is usually much better than $O(kd)$.
1.5 Densification

What is not so nice about the Bernoulli-Gaussian matrix is that it fails pretty miserably for sparse vectors. Fortunately, there is a simple trick for quickly turning sparse vectors into dense vectors: the Fourier transform. Technically, we will use the real-valued counterpart called the Hadamard transform, although if we needed to handle complex numbers, we could use the Fourier transform.

Assume \( d \) is a power of two. The \( d \times d \) Hadamard matrix \( H_d \in \{ \pm 1 \}^{d \times d} \) is defined recursively as follows:

\[
H_2 := \begin{bmatrix} +1 & +1 \\ +1 & -1 \end{bmatrix}, \quad H_d := \begin{bmatrix} H_{d/2} & H_{d/2} \\ H_{d/2} & -H_{d/2} \end{bmatrix}.
\]

(Okay, we could’ve started from one if we really wanted.) Two very important things to know about \( H_d \) are as follows.

1. \( \frac{1}{\sqrt{d}} H_d \) is an orthogonal matrix, so \( \|H_d u/\sqrt{d}\|_2 = \|u\|_2 \).

2. For any vector \( u \in \mathbb{R}^d \), \( H_d u \) can be computed in \( O(d \log(d)) \) time, using divide-and-conquer, just like the FFT algorithm.

Here is another important fact: the “time-frequency” uncertainty principle. Let \( \hat{u} := H_d u/\sqrt{d} \) be the Hadamard transform of \( u \). Then \( \|u\|_0 \|\hat{u}\|_0 \geq d \). This means that if \( u \) is sparse, then \( \hat{u} \) must be dense.

So one idea to fix the Bernoulli-Gaussian matrix to handle sparse vectors is to compose it with the (scaled) Hadamard matrix. However, there are two issues with this:

1. We need not just that \( \hat{u} \) has large number of non-zero entries, but for the energy to be spread out: we need \( \|\hat{u}\|_\infty \) to be small. This is alleviated by the following fact: \( \|\hat{u}\|_\infty \leq \|u\|_1 / \sqrt{d} \).

2. There are dense vectors that become sparse after applying the Hadamard transform: just pick any sparse vector \( \hat{u} \), and set \( u := H_d \hat{u}/\sqrt{d} \).

Here is a fix that works: put a random sign on every column of \( H_d \). Let \( X \in \mathbb{R}^d \) be a Rademacher random vector, and instead of using \( H_d/\sqrt{d} \), use \( R/\sqrt{d} \), where

\[
R := H_d \text{diag}(X).
\]

Observe that:

1. \( \frac{1}{\sqrt{d}} R \) is an orthogonal matrix, so \( \|H_d u/\sqrt{d}\|_2 = \|u\|_2 \).

2. For any vector \( u \in \mathbb{R}^d \), \( R_d u \) can be computed in \( O(d \log(d)) \) time: first apply \( \text{diag}(X) \) to \( u \) in \( O(d) \) time, then apply \( H_d \) in \( O(d \log(d)) \) time.

Now we just check that \( R/\sqrt{d} \) has the desired densification property. Let \( v = (v_1, v_2, \ldots, v_d) := R u/\sqrt{d} \) for \( u \in S^{d-1} \). Then \( v_i = u^\top r_i/\sqrt{d} \), where \( r_i \) is the \( i \)-th row of \( R \). What is the distribution of \( r_i \)? It is a Rademacher random vector. Therefore, the subgaussian Chernoff bound and a union bound imply that with probability at least \( 1 - \delta \),

\[
|v_i| \leq \sqrt{\frac{2 \ln(d/\delta)}{d}} \quad \forall i \in [d].
\]
This means, with probability at least $1 - \delta$,

$$\|Ru/\sqrt{d}\|_\infty \leq O\left(\sqrt{\frac{\log(d/\delta)}{d}}\right)$$

which is exactly what we wanted.

Putting everything together, we set

$$A := \frac{1}{\sqrt{kd}}SH_d\text{diag}(X)$$

where $S = (B_{i,j}Z_{i,j}/\sqrt{q})$ is a matrix of scaled i.i.d. Bernoulli-Gaussian random variables where $B_{i,j} \sim \text{Bernoulli}(q)$ and $Z_{i,j} \sim \mathcal{N}(0, 1)$, and $X$ is a Rademacher random vector. Then as long as

$$q \geq O\left(\frac{(log(k/\delta) \log(d/\delta))}{d}\right),$$

the random matrix $A$ satisfies (1). Moreover, $A$ can be applied in time $O((k \log(k) + d) \log(d)) = O(k + d)$. 