1 Preliminaries

Last lecture we introduced the Johnson-Lindenstrauss lemma, a foundational result in dimensionality reduction. We considered a distribution $\mathcal{D}_{m \times d}$ over $m \times d$ matrices which could be sampled as follows: generate a random matrix $G$ with each entry $g_{ij}$ an i.i.d. standard normal variable (i.e. $g_{ij} \sim \mathcal{N}(0, 1)$) and then scale $G$ by $1/\sqrt{m}$. We proved that

$$\text{Theorem 1. If } \Pi \text{ is chosen from } \mathcal{D}_{m \times d} \text{ and } m = O((\log(1/\delta))/\epsilon^2), \text{ then for any vector } x,$$

$$\begin{aligned}
(1 - \epsilon)\|x\|_2^2 &\leq \|\Pi x\|_2^2 \leq (1 + \epsilon)\|x\|_2^2
\end{aligned} \tag{1}
$$

with probability $1 - \delta$.

One common way of applying this lemma in practice is to choose $\delta$ small enough so that (1) holds simultaneously for many vectors $x$ by a union bound. For example, we showed that, if we have $n$ points $v_1, \ldots, v_n \in \mathbb{R}^d$, then as long as we choose $\delta = \delta'/(n^2/2),$

$$\begin{aligned}
(1 - \epsilon)\|v_i - v_j\|_2^2 &\leq \|\Pi v_i - \Pi v_j\|_2^2 \leq (1 + \epsilon)\|v_i - v_j\|_2^2
\end{aligned}
$$

for all pairs $v_i, v_j$ with probability $1 - \delta'$. This is the original form of the Johnson-Lindenstrauss lemma, and is useful in proving that $\Pi v_1, \ldots, \Pi v_n$ can be used in any downstream task that depends on the Euclidean distance between data points (e.g. distance based clustering, near neighbor search, etc.).

2 Beyond the Union Bound

At the end of last lecture, we sought to apply Johnson-Lindenstrauss dimensionality reduction to approximately solving a least square regression problem. Specifically, for some $A \in \mathbb{R}^{d \times s}$ and some $y \in \mathbb{R}^d$, we want to approximately solve:

$$\min_{x \in \mathbb{R}^s} \|Ax - y\|_2^2 \tag{2}$$

by instead solving the “sketched” problem

$$\min_{x \in \mathbb{R}^s} \|\Pi Ax - \Pi y\|_2^2. \tag{3}$$

As long as $\Pi$ is chosen so that $m \leq d$, then $\Pi A$ contains fewer data points than $A$ and (3) can be solved much faster than (2): in $O(ms^2)$ vs. $O(ds^2)$ time.
Let \( \tilde{x}^* \) be the optimal solution for (3). We want to argue that
\[
\| A \tilde{x}^* - y \|_2^2 \leq (1 + \epsilon) \min_{x \in \mathbb{R}^s} \| Ax - y \|_2^2,
\]
and saw that, to do so, it suffices to prove:
\[
\forall x \in \mathbb{R}^s \quad (1 - \epsilon) \| Ax - y \|_2^2 \leq \| \Pi(Ax - y) \|_2^2 \leq (1 + \epsilon) \| Ax - y \|_2^2. \tag{4}
\]

Proving this statement requires establishing a Johnson-Lindenstrauss type bound for an infinity of possible vectors \( Ax - y \), which obviously can’t be tackled with a union bound argument. Today we will see how to prove this result using a different approach.

3 Subspace Embeddings

We will prove a more general statement that implies (4) and is useful in other applications.

**Theorem 2.** Let \( \mathcal{U} \subseteq \mathbb{R}^d \) be an \( s \)-dimensional linear subspace in \( \mathbb{R}^d \). If \( \Pi \in \mathbb{R}^{m \times d} \) is chosen from any distribution \( \mathcal{D} \) satisfying Theorem 1, then with probability \( 1 - \delta \),
\[
(1 - \epsilon) \| v \|_2 \leq \| \Pi v \|_2 \leq (1 + \epsilon) \| v \|_2 \tag{5}
\]
for all \( v \in \mathcal{U} \), as long as \( m = O \left( \frac{s \log(1/\epsilon) + \log(1/\delta)}{\epsilon^2} \right) \).

![Figure 1: Theorem 2 extends Theorem 1 to all points in a linear subspace \( \mathcal{U} \).](image)

How does Theorem 2 imply (4)? We can apply it to the \( s + 1 \) dimensional subspace spanned by \( A \)'s \( s \) columns and \( y \). Every vector \( Ax - y \) lies in this subspace. So, for regression, we will require dimension \( m = O \left( \frac{(s+1)\log(1/\epsilon)}{\epsilon^2} \right) \).

We start with the observation that Theorem 2 holds as long as (5) holds for all points on the unit sphere in \( \mathcal{U} \). This is a consequence of linearity. We denote the sphere \( S_\mathcal{U} \):
\[
S_\mathcal{U} = \{ v \mid v \in \mathcal{U} \text{ and } \| v \|_2 = 1 \}.
\]

Any point \( v \in \mathcal{U} \) can be written as \( cx \) for some scalar \( c \) and some point \( x \in S_\mathcal{U} \). If
\[
(1 - \epsilon) \| x \|_2 \leq \| \Pi x \|_2 \leq (1 + \epsilon) \| x \|_2 \text{ then } c(1 - \epsilon) \| x \|_2 \leq c\| \Pi x \|_2 \leq c(1 + \epsilon) \| x \|_2 \text{ and thus (1 - \epsilon)\| cx \|_2 \leq \| \Pi cx \|_2 \leq (1 + \epsilon)\| cx \|_2.}
\]

\(^1\)It’s possible to obtain a slightly tighter bound of \( O \left( \frac{s \log(1/\delta)}{\epsilon^2} \right) \). It’s a nice challenge to try proving this. Hint: use a constant factor net \( N_{O(1)} \) instead of an \( \epsilon \) net \( N_\epsilon \) as we do below.
4 An argument via $\epsilon$-nets

We will prove Theorem 2 by showing that there exists a large but finite set of points $N_\epsilon \subset S_\mathcal{U}$ such that, if (5) holds for all $v \in N_\epsilon$, then it must hold for all $v \in S_\mathcal{U}$, and by the argument above, for all $v \in \mathcal{U}$. $N_\epsilon$ is called an “$\epsilon$-net”.

**Lemma 3.** For any $\epsilon \leq 1$, there exists a set $N_\epsilon \subset S_\mathcal{U}$ with $|N_\epsilon| = \left(\frac{4}{\epsilon}\right)^d$ such that $\forall v \in S_\mathcal{U}$,

$$\min_{x \in N_\epsilon} \|v - x\| \leq \epsilon.$$

**Construction of the $\epsilon$-net.**

**Proof.** Consider the following greedy procedure for constructing $N_\epsilon$ (which we don’t actually need to implement – it’s just for the proof argument):

- Set $N_\epsilon = \{\}$
- While such a point exists, choose an arbitrary point $v \in S_\mathcal{U}$ where $\nexists x \in N_\epsilon$ with $\|v - x\| \leq \epsilon$. Set $N_\epsilon = N_\epsilon \cup \{v\}$.

After running this procedure, we have $N_\epsilon = \{x_1, \ldots, x_{|N_\epsilon|}\}$ points that satisfy the condition $\min_{x \in N_\epsilon} \|v - x\| \leq \epsilon$ for all $v \in S_\mathcal{U}$. So we just need to bound $|N_\epsilon|$.

To do so, we note that, for all $i, j$, $\|x_i - x_j\| \geq \epsilon$. If not, then either $x_i$ or $x_j$ would not have been added to $N_\epsilon$ by our greedy procedure. Accordingly, if we place balls of radius $\epsilon/2$ around each $x_i$:

$$B(x_1, \epsilon/2), \ldots, B(x_{|N_\epsilon|}, \epsilon/2)$$

then for all $i, j$, $B(x_i, \epsilon/2)$ does not intersect $B(x_j, \epsilon/2)$.

The volume of a $d$ dimensional ball of radius $r$ is $cr^d$ for some value $c$ that does not depend on $r$. So the total volume of $B(x_1, \epsilon/2) \cup \ldots \cup B(x_{|N_\epsilon|}, \epsilon/2)$ is $|N_\epsilon| \cdot c \left(\frac{\epsilon}{2}\right)^d$. At the same time, $B(x_1, \epsilon/2), \ldots, B(x_{|N_\epsilon|}, \epsilon/2)$ are contained inside a ball of radius $1 + \epsilon/2$, which has volume $< \epsilon 2^d$. So we have:

$$|N_\epsilon| \cdot c \left(\frac{\epsilon}{2}\right)^d < 2^d \quad \text{which implies} \quad |N_\epsilon| \leq \left(\frac{4}{\epsilon}\right)^d.$$
Extension to all vectors.

We are now ready to prove Theorem 2.

**Proof.** Choose \( m = O \left( \frac{\log(\sqrt{N}/\delta)}{\epsilon^2} \right) \) so that (5) holds for all \( x \in N_\epsilon \).

Now consider any \( v \in S_d \). It’s not hard to see that, for some \( x_0, x_1, x_2 \ldots \in N_\epsilon \), \( v \) can be written:

\[
v = x_0 + c_1x_1 + c_2x_2 + \ldots
\]

for constants \( c_1, c_2, \ldots \) where \( |c_i| \leq \epsilon^i \). Applying triangle inequality, we have

\[
\|\Pi v\|_2 = \|\Pi x_0 + c_1\Pi x_1 + c_2\Pi x_2\|_2 \\
\leq \|\Pi x_0\| + \epsilon\|\Pi x_1\| + \epsilon^2\|\Pi x_2\| + \ldots \\
\leq (1 + \epsilon) + \epsilon(1 + \epsilon) + \epsilon^2(1 + \epsilon) + \ldots \\
\leq 1 + O(\epsilon).
\]

Similarly,

\[
\|\Pi v\|_2 = \|\Pi x_0 + c_1\Pi x_1 + c_2\Pi x_2\|_2 \\
\geq \|\Pi x_0\| - \epsilon\|\Pi x_1\| - \epsilon^2\|\Pi x_2\| - \ldots \\
\leq (1 - \epsilon) - \epsilon(1 + \epsilon) - \epsilon^2(1 + \epsilon) - \ldots \\
\leq 1 - O(\epsilon).
\]

So we have proven

\[
1 - O(\epsilon) \leq \|\Pi v\|_2 \leq 1 + O(\epsilon)
\]

for all \( v \) in \( S_d \). As discussed early, this is sufficient to prove the theorem. \( \square \)

## 5 Faster Johnson-Lindenstrauss dimensionality reduction

Theorem 2 shows that, if we solve our regression problem using \( \Pi A \) and \( \Pi y \) in place of \( A \) and \( y \), we can reduce our running time from \( O(ds^2) \) to approximately \( O(s^3) \), at least if we are willing to settle for an approximate solution.

But that’s not counting the cost to compute \( \Pi A \) and \( \Pi y \). Naively, that cost is \( O(ds^2) \)! I.e., the cost to multiple \( A \in \mathbb{R}^{d \times s} \) by our sketching matrix \( \Pi \in \mathbb{R}^{s \times d} \). If we want to actually speed up least squares regression, we need to do better than that.

The following remarkable result of Ailon and Chazelle [1] shows how to do much better:

**Theorem 4.** For all \( m, d \), there exists a set of \( m \times d \) matrices \( F \) such that, for all \( x \) and all \( \Pi \in F \), \( \Pi x \) can be computed in \( O(d \log d) \) time. Moreover, if \( m = O \left( \frac{\log(d/\delta)^2 \log(1/\delta)}{\epsilon^2} \right) \) and \( \Pi \) is drawn uniformly at random from \( F \), then for any \( x \),

\[
(1 - \epsilon)\|x\|_2^2 \leq \|\Pi x\|_2^2 \leq (1 + \epsilon)\|x\|_2^2
\]

with probability \( 1 - \delta \).
What’s the consequence for regression? Using the same $\epsilon$-net argument that we used for random Gaussian matrices, we will need to sketch to dimension $m = O(s \log^2 d/\epsilon^2)$ to get an approximate solution with error $\epsilon$. We can compute $\Pi A$ and $\Pi y$ in $O(md \log d)$ time. We can thus obtain an approximate solution in total time $O(sd \log^2 d + s^3 \log^2 d)$ time.

This is a pretty remarkable runtime – the first term is only a polylog factor larger than how long it takes to simply read all of the entries in $A$!

**Construction**

We will describe a distribution over matrices that achieves Theorem 4 by describing an algorithm for selecting a matrix from the distribution randomly. Ailon and Chazelle’s construction relies on what’s known as the “Fast Hadamard Transform”, $H_k$, which is a square matrix of size $d = 2^k$ for some integer $k$.

\[
H_1 = 1 \quad H_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad H_3 = \frac{1}{\sqrt{4}} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \quad \cdots \quad H_k = \frac{1}{\sqrt{2^k}} \begin{bmatrix} H_{k-1} & H_{k-1} \\ H_{k-1} & -H_{k-1} \end{bmatrix}
\]

Assuming for now that $d$ is a power of 2 (if it’s not, you can pad with zeros until it is) our construction for $\Pi \in \mathbb{R}^{m \times d}$ is:

- Chose a $d \times d$ diagonal matrix $D$ by selecting each diagonal entry independently to be $\pm 1$, each with probability $1/2$.

- Chose a random $m \times d$ sampling matrix $S$, which contains a single entry of $\sqrt{d/m}$ in each row in position $i$, where $i$ is chosen uniformly at random from $1, \ldots, d$.

- Set $\Pi = SHD$.

$SHD$ is called a “subsampled randomized Hadamard transform”. To understand the performance of $SHD$, notice that every $H_k$ has two important properties:

1. $H_k x$ can be be computed in $O(d \log d)$ time (using a divide-and-conquer algorithm).

2. $H_k$ is orthonormal: i.e. $H_k^T H_k = I$ and thus $\|H_k x\|_2 = \|x\|_2$ for all $x$.

Using property 1, we see that it’s possible to compute $\Pi x = SHD x$ in $O(d \log d)$ time. We will use property 2 shortly.

**Intuition**

$\Pi$ can be applied quickly to vectors, but why should we expect it to preserve norms with high probability?

Consider what would happen if we instead tried to approximate $\|x\|_2$ by $\|Sx\|_2$ – i.e. we sketch $x$ by simply sub-sampling its coordinates. $E \|Sx\|_2 = \|x\|_2$, so the estimate is correct in expectation, but it does not concentrate well for all $x$. If $x$ is very sparse (imagine it is...
only non-zero in one location) then with good probability we will simply get an estimate of \(\|Sx\|_2 = 0\).

Ailon and Chazelle’s main observation was that \(H\) can avoid this bad case by “spreading out” sparse vectors, without changing their norm (since it’s orthonormal). In the most extreme case, if \(x\) only has a single non-zero entry, all entries in \(Hx\) will have the same absolute value, \(\|SHx\|_2\) exactly equals \(\|x\|_2\).

This effect holds more generally. In fact, the original paper was inspired by the uncertainty principal in physics. There are many different ways to state the uncertainty principal, but one is that “no function can be locally concentrated in both the time and frequency domain”. \(H\) is a discrete version of the Fourier transform, so multiplying \(x\) by \(H\) converts it to a sort of “frequency domain”. If \(x\) is locally concentrated (i.e., sparse or approximately sparse) than \(Hx\) won’t be.

Why introduce the random diagonal matrix \(D\)? If we simply used \(Hx\) then \(\Pi\) wouldn’t be randomized. It would be trivial to cook up some \(x\) so that, e.g. \(Hx = [1; 0; 0; \ldots, 0]\), in which case \(\|SHx\|_2\) would fail to estimate \(\|x\|_2\) with high probability. The diagonal matrix prevents such a case for observing – \(D\) randomly flips every entry of \(x\), making it extremely unlikely that such bad cases occur.

The final effect is that \(SHD\) serves as a very effective “pseudorandom” sign matrix, even though it can be multiplied by a vector in \(O(d \log d)\) time and only takes \(O(d)\) random bits to specify.

\[
| [HDx]_i | \leq \frac{\sqrt{\log(d/\delta)}}{\sqrt{d}} \|x\|_2
\]

with probability \(1 - \delta\).

**Analysis**

Making the intuition above formal is surprisingly simple. We first prove:

**Lemma 5.** If \(\Pi = SHD\) is chosen as described and \(m = \log(d/\delta)\) then, for all \(i \in 1, \ldots d\),

\[
[H Dx]_i \leq \frac{\sqrt{\log(d/\delta)}}{\sqrt{d}} \|x\|_2
\]

with probability \(1 - \delta\).
Proof. To prove this lemma, consider any particular row of $HDx$ – i.e. any particular $i$. We will prove the bound for each row and then obtain the result via a union bound. For any one row, $[HDx]_i$ is simply equivalent to multiplying $x$ by a vector with i.i.d. random sign vector (and then scaling by $1/\sqrt{d}$). This allows to apply:

**Lemma 6** (Corollary of Hoeffding Bound\(^2\)). If $\sigma_1, \ldots, \sigma_d$ are each selected independently and uniformly from $\{-1, +1\}$ then:

$$Pr\left[\left|\sum_{i=1}^{d} \sigma_i x_i\right| \geq t\right] \leq 2e^{-\frac{x^2}{2x^2}}.$$  

Alternatively, a similar tail bound can be proven using a moment method and the Khintchine inequality:\(^3\)

$$\left(\mathbb{E}\left[\sum_{i=1}^{d} \sigma_i x_i\right]^p\right)^{1/p} \leq O(\sqrt{p}\|x\|_2).$$

So if we choose $t = O\left(\sqrt{\log(d/\delta)}\|x\|_2\right)$ then $|[HDx]_i| \leq \frac{\log(d/\delta)}{\sqrt{d}}\|x\|_2$ with probability $1 - \delta/d$. Lemma 5 then holds by a union bound. \(\square\)

With Lemma 5 in place, we can condition on the event that each $(|HDx|_i)^2 \leq \log(d/\delta)\|x\|_2^2$. Now consider our estimator $\|SHDx\|_2^2$, which equals

$$\|SHDx\|_2^2 = \frac{d}{m} \sum_{k=1}^{m} [HDx]_{i_k}^2.$$  

Here each $i_k$ is a random index in $1, \ldots, d$. Since $H$ is orthonormal, $\|HDx\|_2^2 = \|x\|_2^2$ and thus

$$\mathbb{E}\|SHDx\|_2^2 = d \cdot \mathbb{E}[|HDx|_k^2] = \mathbb{E}\|HDx\|_2^2 = \|x\|_2^2.$$  

So our estimator is correct in expectation. Additionally, considering (7) and Lemma 5, $\|SHDx\|_2^2$ is an average of $m$ random variables, each bounded in $[0, \log(d/\delta)\|x\|_2^2]$. Theorem 4 then follows either from a Bernstein bound, or a Hoeffding bound. We need to choose $m = O\left(\frac{\log(d/\delta)^2 \log(1/\epsilon)}{\epsilon^2}\right)$.

References


\(^2\)See e.g. Theorem 4 in http://cs229.stanford.edu/extra-notes/hoeffding.pdf for a Hoeffding bound that can be used.

\(^3\)For a proof of this bound see http://people.seas.harvard.edu/~minilek/cs229r/fall15/lec/lec11.pdf.