CS 761: Randomized Algorithms **University of Waterloo: Winter 2025**

Lecture 12

February 27, 2025

Instructor: Sepehr Assadi

Topics of this Lecture

1 Subspace embeding

In the previous note, we saw low-dimensional embeddings of vectors that preserve the ℓ_2-norm. In particular, we proved the following claim:

Claim 1. *Let $\mathbf{S} \in \mathbb{R}^{t\times d}$ be a matrix of independent $\mathcal{N}(0,1)$ variables. For a vector $v \in \mathbb{R}^d$ such that $\|v\| = 1$ and $t = 10\ln(1/\delta)/\varepsilon^2$,*

$$\Pr_{\mathbf{S}}[\|\mathbf{S}v\|/\sqrt{t} \notin [1-\varepsilon, 1+\varepsilon]] \leqslant 2\delta.$$

In this lecture, we will see how to extend this notion to embedding an entire subspaces (with infinitely many vectors). We continue to use our notation of $a \approx_\varepsilon b$ to mean $(1-\varepsilon)b \leqslant a \leqslant (1+\varepsilon)b$.

Definition 2 (ε-subspace embedding)**.** Let U be a d-dimensional subspace of $\mathbb{R}^n$. Then $S \in \mathbb{R}^{t\times n}$ is an ε-subspace embedding of U iff for all $y \in U$, $\|Sy\| \approx_\varepsilon \|y\|$.

Recall that if $A \in \mathbb{R}^{n\times d}$ is a matrix whose columns are a basis for U, then $U = \{Ax \mid x \in \mathbb{R}^d\}$. In particular, if we use an orthonormal basis $u_1, \ldots, u_d$ as the columns of A, and let $S = A^T$, we obtain that for $y = \sum_i \alpha_i u_i \in U$, $\|Sy\|^2 = \sum_i \alpha_i^2 = \|y\|^2$. This is also tight, since if S has fewer than d columns, $\dim(\mathrm{Ker}(S) \cap U) > 0$, which means S maps a non-zero vector within U to 0. However, for algorithmic applications, this solution is not good enough, since we may not always have a handle on U — our aim will be to develop an analogue of Claim 1: an embedding where S is oblivious to U (hence necessarily random).

Theorem 3. *If $\mathbf{S}$ is a $t\times n$ matrix of independent $\mathcal{N}(0,1)$'s where $t = O(d/\varepsilon^2)$, then with probability $1-1/2^d$, $\mathbf{S}$ is an ε-subspace embedding of U.*

As above, let $A \in \mathbb{R}^{d\times n}$ be a matrix whose columns are an orthonormal basis for U, so $U = \{Ax \mid x \in \mathbb{R}^d\}$. As such, the theorem is equivalent to showing that $\|\mathbf{S}Ax\| \approx_\varepsilon \|Ax\|$ for all $x \in \mathbb{R}^d$. In fact, because $\mathbf{S}$ and A are linear, and $\|cx\| = c\|x\|$ for any $c \in \mathbb{R}$, it is enough to show the above for all $x \in \mathbb{R}^d$ such that $\|x\| = 1$.

A short proof using two facts about Gaussian random variables follows. First, for any orthonormal matrix A, the matrix $\mathbf{S}A$ itself is a $t \times d$ matrix of independent $\mathcal{N}(0,1)$'s. So it is enough to show that $\|\mathbf{T}x\| \approx_\varepsilon \|x\|$ for each $x \in \mathbb{R}^d$, and where $\mathbf{T}$ is a $t \times d$ matrix of independent $\mathcal{N}(0,1)$'s. Second, it turns out that the singular values of $\mathbf{T}$ are concentrated around 1. That is, $\sigma_{\max}(\mathbf{T}) \leqslant 1+\varepsilon$, and $\sigma_{\min}(\mathbf{T}) \geqslant 1-\varepsilon$ with high probability. Singular values measure exactly how much $\mathbf{T}$ stretches its input in the ℓ_2 norm, so these two facts are sufficient to get the theorem.

In the rest of this lecture, we show a self-contained proof of Theorem 3 that both showcase new ideas for proving such statements and also has the benefit of working with any JLL-type matrix $\mathbf{S}$ that can preserve the norm of vectors approximately (regardless whether its distribution is Gaussian or not).

1.1 γ-Nets

The main idea we want to use to prove Theorem 3 is a union bound and Claim 1 – but it is of course impossible to union bound over the uncountably many vectors in U. Instead, we will union bound over some (finitely many) vectors, and somehow interpolate between them to get the full theorem. Let $\mathcal{S}^d$ denote the unit sphere $\{x \in \mathbb{R}^d \mid \|x\| = 1\}$ in d dimensions.

Definition 4 (γ-nets)**.** A set $N \subseteq \mathcal{S}^d$ is a γ-net if for all $x \in \mathcal{S}^d$, there is a $y \in N$ such that $\|x-y\| \leqslant \gamma$.

Note that in the literature, these are typically called ε-nets (we avoided using ε to avoid confusion with the distortion parameter ε of our embedding).

The most immediate question is: how do we construct a small, or even finite, γ-net? Greedily! In particular, start with $N = \emptyset$, and while there is a vector $x \in \mathcal{S}^d$ with no $y \in N$ such that $\|x-y\| \leqslant \gamma$, add x to N. To analyse this algorithm, we have the following proposition about pairwise distances in $\mathcal{S}^d$.

Proposition 5. *Let $M \subseteq \mathcal{S}^d$ such that for each $x, y \in M$, $\|x-y\| > \gamma$. Then $|M| \leqslant (4/\gamma)^d$.*

Note that Proposition 5 immediately implies that the greedy algorithm produces a γ-net of size $O(1/\gamma)^d$, since a point is added to N iff its pairwise distance with all previous points is $> \gamma$.

Proof of Proposition 5. Let $\mathcal{B}_r^d(x)$ denote the d-dimensional closed ball of radius r centered on the point x, i.e. $\{y \in \mathbb{R}^d \mid \|x-y\| \leqslant r\}$. Suppose we place a ball of radius $\gamma/2$ around each point in M. Then for any two points $x, y \in M$, since $\|x-y\| > \gamma$, $\mathcal{B}_{\frac{\gamma}{2}}^d(x)$ and $\mathcal{B}_{\frac{\gamma}{2}}^d(y)$ do not intersect. On the other hand, since $M \subseteq \mathcal{S}^d$, we know that for each $x \in M$, $\mathcal{B}_{\frac{\gamma}{2}}^d(x)$ is contained in $\mathcal{B}_{1+\frac{\gamma}{2}}^d(0)$. Hence, the number of points in M is upper bounded by the ratio of these balls' volumes:

$$|M| \leqslant \frac{\mathrm{Vol}(\mathcal{B}_{1+\frac{\gamma}{2}}^d)}{\mathrm{Vol}(\mathcal{B}_{\frac{\gamma}{2}}^d)} = \frac{(1+\gamma/2)^d}{(\gamma/2)^d} \leqslant \left(\frac{4}{\gamma}\right)^d,$$

where the first equality holds because the volume of $\mathcal{B}_r^d$ is equal to $f(d) \cdot r^d$ for some function $f : \mathbb{N} \to \mathbb{R}$, and the second inequality holds because $\gamma \leqslant 2 \implies (1+\gamma/2) \leqslant 2$. (If $\gamma > 2$, a single point in N is enough to cover all of $\mathcal{S}^d$). □

We will now see two different ways to prove Theorem 3.

1.2 A (weaker) One Shot Argument by Using ε-Nets

Let N be an ε-net for $\mathcal{S}^d$ of size $\leqslant (4/\varepsilon)^d$, which exists by Proposition 5. Also, as before, let A be a matrix whose columns are an orthonormal basis for U. Recall that for any $x \in \mathcal{S}^d$, $\|Ax\| = \|x\| = 1$, and it is sufficient to show that $\|\mathbf{S}Ax\| \approx_\varepsilon \|Ax\|$ for $x \in \mathcal{S}^d$ to obtain Theorem 3. For any $x \in N$, we can

apply Claim 1 with $\delta = 1/(4|N|^2)$ to obtain that when $\mathbf{S}$ is a $t \times n$ matrix of independent $\mathcal{N}(0,1)$'s with $t = 20\ln(4|N|)/\varepsilon^2$,

$$\Pr[\|\mathbf{S}Ax\| \not\approx_\varepsilon 1] \leqslant \frac{1}{2|N|^2}.$$

Using a union bound over all the above event for all $x \in N$, we have that $\mathbf{S}Ax \approx_\varepsilon 1$ for all $x \in N$ with probability $\geqslant 1 - 1/|N| \geqslant 1 - 2^{-d}$.

From this point onwards, we will fix S to be a $t \times n$ matrix that satisfies the above (i.e. we will make no further use of randomness). The rest of the argument shows that now that S "works" for the points in the net N, it should "somewhat work" for all the points in $\mathcal{S}^d$ (and by extension $\mathbb{R}^d$) as well.

We bound each direction of the inequalities in the following two claims separately. The proofs are quite similar by symmetry and we only provide both for completeness.

Claim 6. *For any point $x \in \mathcal{S}^d$, $\|SAx\| \leqslant 1 + 3\varepsilon$.*

Proof. Let $x^\star$ be a maximizer of $\|SAx\|$ over $x \in \mathcal{S}^d$ (such an $x^\star$ exists because $\|SAx\|$ is a continuous function, and $\mathcal{S}^d$ is compact).

Let $y \in N$ be the closest point to x in the net, i.e., a minimizer of $\|x^\star - y\|$. Define $z := (x^\star - y)/\|x^\star - y\|$, and note that $z \in \mathcal{S}^d$, and $\|SAz\| \geqslant \|SAx^\star\|$ by our choice of $x^\star$. Using the linearity of S and A, and the triangle inequality, we have:

$$\|SAx^\star\| = \|SA(x^\star - y) + SAy\| \leqslant \|SA(x^\star - y)\| + \|SAy\| \leqslant \|x^\star - y\| \cdot \|SAz\| + (1 + \varepsilon),$$

where the last inequality holds because S preserves the norm of $y \in N$. But since $\|SAz\| \leqslant \|SAx^\star\|$, and $\|x^\star - y\| \leqslant \varepsilon$ because N is an ε-net, we can rearrange the above to see that

$$\|SAx^\star\| \leqslant \frac{1+\varepsilon}{1-\varepsilon} \leqslant 1 + 3\varepsilon,$$

proving the claim. □

Claim 7. *For any point $x \in \mathcal{S}^d$, $\|SAx\| \geqslant 1 - 3\varepsilon$.*

Proof. Let $x^\star$ be a minimizer of $\|SAx\|$ over $x \in \mathcal{S}^d$ (such an $x^\star$ exists because $\|SAx\|$ is a continuous function, and $\mathcal{S}^d$ is compact).

Let $y \in N$ be the closest point to x in the net, i.e., a minimizer of $\|x^\star - y\|$. Define $z := (x^\star - y)/\|x^\star - y\|$, and note that $z \in \mathcal{S}^d$, and $\|SAz\| \leqslant \|SAx^\star\|$ by our choice of $x^\star$. Using the linearity of S and A, and the triangle inequality, we have:

$$\|SAx^\star\| = \|SAy + SA(x^\star - y)\| \geqslant \|SAy\| - \|SA(x^\star - y)\| \geqslant (1 - \varepsilon) - \|x^\star - y\| \cdot \|SAz\|,$$

where the last inequality holds because S preserves the norm of $y \in N$. But since $\|SAz\| \geqslant \|SAx^\star\|$, and $\|x^\star - y\| \leqslant \varepsilon$ because N is an ε-net, we can rearrange the above to see that

$$\|SAx^\star\| \geqslant \frac{1-\varepsilon}{1+\varepsilon} \geqslant 1 - 3\varepsilon,$$

proving the claim. □

Putting it all together, we have shown that if $\mathbf{S}$ is a $t \times n$ matrix of $\mathcal{N}(0,1)$'s, where

$$t = 10\ln(4|N|)/\varepsilon^2 \leqslant 10\left(d\ln\left(\frac{4}{\varepsilon}\right) + O(1)\right)/\varepsilon^2 = O\left(\frac{1}{\varepsilon^2} \cdot d\lg\left(\frac{1}{\varepsilon}\right)\right),$$

then for any d-dimensional subspace $U \subseteq \mathbb{R}^n$, with probability $\geqslant 1 - 2^{-d}$, we have $\|\mathbf{S}x\| \approx_{3\varepsilon} \|x\|$ for all $x \in U$. To get the right distortion and probability bound, we need to use an ε'-net with $\varepsilon' = \varepsilon/3$, to make t larger (say $20\ln(4|N|)/(\varepsilon')^2$). While this is very close to the right answer, we are still off by a factor of $\log(1/\varepsilon)$ from the promised statement in Theorem 3.

1.3 The Optimal Bound: Chaining via $(1/2)$-Nets

We now show a general way of using nets via a simple application of the so-called *chaining* technique.

In a quest to drop the $\log(1/\varepsilon)$ dependence in our previous proof, we start with a much coarser $(1/2)$-net N of $\mathcal{S}^d$. This time, we will pick $\delta = 1/(2|N|^4)$ instead, so $\mathbf{S}$ is a $t \times n$ matrix of $\mathcal{N}(0,1)$'s with $t = 40\ln(4|N|)/\varepsilon^2 = O(d/\varepsilon^2)$. We chose a smaller value of δ relative to $|N|$ so that we can union bound over $|N|^2$ events, but we do not pay a $\log(1/\varepsilon)$ factor in t because N is smaller. By applying Claim 1 to each point, and each pair of points in N, we have:

- For each $x \in N$, $\|\mathbf{S}Ax\| \approx_\varepsilon 1$ with probability $1 - 1/|N|^4$.
- For each pair $x, y \in N$, $\|\mathbf{S}A(x-y)\| \approx_\varepsilon \|A(x-y)\|$ with probability $1 - 1/|N|^4$.

All of the events above hold simultaneously with probability $\geqslant 1 - 1/|N|^2 \geqslant 1 - 2^{-d}$ by the union bound, and from this point we fix S to be a $t \times n$ matrix satisfying the above. For brevity, let $T = SA$. Before considering an arbitrary point in $\mathcal{S}^d$, we first show that T roughly preserves even inner products of pairs of points in N.

Claim 8. *For $x, y \in N$, we have $|\langle Tx\,,\,Ty\rangle - \langle x\,,\,y\rangle| \leqslant 3\varepsilon\,\langle x\,,\,y\rangle + 5\varepsilon$.*

Proof. This is because

$$\begin{aligned}
2\,\langle Tx\,,\,Ty\rangle &= \|T(x-y)\|^2 - \|Tx\|^2 - \|Ty\|^2 \\
&\leqslant \|x-y\|^2 \cdot (1+\varepsilon)^2 - \left(\|x\|^2 + \|y\|^2\right)\cdot(1-\varepsilon)^2 && (T \text{ preserves norms on } N) \\
&\leqslant \|x-y\|^2 \cdot (1+3\varepsilon) - \left(\|x\|^2 + \|y\|^2\right)\cdot(1-2\varepsilon) && (??) \\
&= \left(\|x-y\|^2 - \|x\|^2 - \|y\|^2\right)\cdot(1+3\varepsilon) + \left(\|x\|^2 + \|y\|^2\right)\cdot 5\varepsilon \\
&= 2\,\langle x\,,\,y\rangle \cdot (1+3\varepsilon) + 10\varepsilon. && (1)
\end{aligned}$$

And similarly,

$$\begin{aligned}
2\,\langle Tx\,,\,Ty\rangle &= \|T(x-y)\|^2 - \|Tx\|^2 - \|Ty\|^2 \\
&\geqslant \|x-y\|^2 \cdot (1-\varepsilon)^2 - \left(\|x\|^2 + \|y\|^2\right)\cdot(1+\varepsilon)^2 \\
&\geqslant \|x-y\|^2 \cdot (1-2\varepsilon) - \left(\|x\|^2 + \|y\|^2\right)\cdot(1+3\varepsilon) \\
&&& (\text{as } (1+\varepsilon)^2 \leqslant (1+3\varepsilon) \text{ and } (1-\varepsilon)^2 \geqslant (1-2\varepsilon)) \\
&= \left(\|x-y\|^2 - \|x\|^2 - \|y\|^2\right)\cdot(1-2\varepsilon) - \left(\|x\|^2 + \|y\|^2\right)\cdot(5\varepsilon) \\
&= 2\,\langle x\,,\,y\rangle \cdot (1-2\varepsilon) - 10\varepsilon. && (2)
\end{aligned}$$

Combining (1) and (2), we obtain that

$$\langle x\,,\,y\rangle \cdot (1-2\varepsilon) - 5\varepsilon \leqslant \langle Tx\,,\,Ty\rangle \leqslant \langle x\,,\,y\rangle \cdot (1+3\varepsilon) + 5\varepsilon,$$

proving the claim. □

We are now ready to prove Theorem 3 with the correct bounds. We will start with $z_0 := x$, for an arbitrary point $x \in \mathcal{S}^d$, and for $i \geqslant 1$, define the sequences

$$y_i = \operatorname*{argmin}_{y \in N} \|z_{i-1} - y\|,$$

and

$$z_i = \frac{z_{i-1} - y_i}{\|z_{i-1} - y\|}.$$

Rewriting x using successively more terms of z_i and y_i, we have

$$\begin{aligned} x &= z_0 \\ &= y_1 + \|z_0 - y_1\| \cdot z_1 \\ &= y_1 + \|z_0 - y_1\| \cdot (y_2 + \|z_1 - y_2\| \cdot z_2) \end{aligned}$$

(Expanding $n-1$ times)

$$= \sum_{i=1}^{n} \left(\prod_{j=1}^{i-1} \|z_{j-1} - y_j\| \right) y_i + \left(\prod_{j=1}^{n} \|z_{j-1} - y_j\| \right) \cdot z_n$$

Roughly speaking, our plan is the following:

1. Consider the sequence (of points in $\mathbb{R}^d$) obtained by dropping the z_n term from the n-th expansion above, and show that T preserves the norms of each point in this sequence.
2. Show that this sequence of points approaches x in the limit, and hence T preserves the norm of x.

For $n \geqslant 1$, let

$$y'_n = \left(\prod_{j=1}^{n-1} \|z_{j-1} - y_j\| \right) \cdot y_n$$

and define

$$x_n = x - \left(\prod_{j=1}^{n} \|z_{j-1} - y_j\| \right) \cdot z_n.$$

(Here x_n is the sequence we alluded to in the plan, and y'_n is just defined for brevity.) The upshot is that since $\|z_n\| = 1$ and $\|y_n\| = 1$, we have

$$\|x - x_n\| = \prod_{j=1}^{n} \|z_{j-1} - y_j\| \leqslant \frac{1}{2^n}, \tag{3}$$

and

$$\|y'_n\| = \prod_{j=1}^{n-1} \|z_{j-1} - y_j\| \leqslant \frac{1}{2^{n-1}}, \tag{4}$$

because each y_j is chosen to minimize $\|y_j - z_{j-1}\|$, and N is a $(1/2)$-net. On the other hand, for any $n \geqslant 1$,

$$\begin{aligned} \|Tx_n\|^2 &= \left\langle T\sum_{i=1}^{n} y'_i \,,\, T\sum_{i=1}^{n} y'_i \right\rangle \\ &= \sum_{i=1}^{n} \|Ty'_i\|^2 + \sum_{i=1}^{n} \sum_{j \in [n] \setminus \{i\}} \langle Ty'_i \,,\, Ty'_j \rangle \\ &\leqslant \sum_{i=1}^{n} \|y'_i\|^2 \cdot (1 + 3\varepsilon) + \sum_{i=1}^{n} \sum_{j \in [n] \setminus \{i\}} \langle y'_i \,,\, y'_j \rangle \cdot (1 + 3\varepsilon) + 8\varepsilon \|y'_i\| \|y'_j\| \\ &\qquad \text{(T preserves norms and Claim 8)} \\ &\leqslant \left\| \sum_{i=1}^{n} y'_i \right\|^2 \cdot (1 + 3\varepsilon) + \sum_{1 \leqslant i \neq j \leqslant n} \frac{8\varepsilon}{2^{i-1} \cdot 2^{j-1}} \\ &\leqslant \|x_n\|^2 \cdot (1 + 3\varepsilon) + 64\varepsilon, \end{aligned}$$

where the second last inequality follows because we placed the $\langle y'_i \,,\, y'_j \rangle \cdot (1 + 3\varepsilon)$ terms into the earlier sum, and used Eq (4) to upper bound the norms $\|y'_i\|$ and $\|y'_j\|$.

To finish, recall Eq (3) and use

$$\lim_{n\to\infty} \|x - x_n\| = 0 \implies \lim_{n\to\infty} x_n = x.$$

And because T and $\|\cdot\|^2$ are both continuous functions, this also means $\lim_{n\to\infty} \|x_n\|^2 = \|x\|^2$ and $\lim_{n\to\infty} \|Tx_n\|^2 = \|Tx\|^2$. Hence we have that

$$\|Tx\|^2 = \lim_{n\to\infty} \|Tx_n\|^2 \leqslant \lim_{n\to\infty} \|x_n\|^2 \cdot (1+3\varepsilon) + 64\varepsilon = \|x\|^2 \cdot (1+3\varepsilon) + 64\varepsilon = 1 + 67\varepsilon,$$

which of course implies $\|Tx\| \leqslant 1 + 67\varepsilon$ since $\varepsilon \geqslant 0$. One can repeat the argument above, interjecting lower bounds instead of upper bounds and obtain a similar lower bound on $\|Tx\|$.

Putting it all together, we have shown that if $\mathbf{S}$ is a $t \times n$ matrix of $\mathcal{N}(0,1)$'s, where

$$t = 10\ln(4|N|^2)/\varepsilon^2 \leqslant 10(d\ln(8) + O(1))/\varepsilon^2 = O\left(\frac{1}{\varepsilon^2} \cdot d\right),$$

then for any d-dimensional subspace $U \subseteq \mathbb{R}^n$, with probability $\geqslant 1 - 2^{-d}$, we have $\|\mathbf{S}x\| \approx_{100\varepsilon} \|x\|$ for all $x \in U$. To get the statement of Theorem 3 exactly, we substitute $\varepsilon' = \varepsilon/100$, and increase t to $20\ln(4|N|^2)/\varepsilon'^2$, paying only a constant blow up in the size of $\mathbf{S}$. This concludes the proof of Theorem 3.

2 A Simple Application: ℓ_2-Regression

Suppose we are given a matrix $A \in \mathbb{R}^{n\times d}$ and a vector $b \in \mathbb{R}^n$, and wish to find $\operatorname{argmin}_{x\in\mathbb{R}^d} \|Ax - b\|$. This is a fairly standard problem, and has an easy solution: Take the derivative of $f(x) = \|Ax - b\|$, equate it with 0, and solve for x ($x = (A^\top A)^{-1} A^\top b$). The solution above requires us to have all of A in memory, and we would like to avoid this.

We can use subspace embedding to convert this to a problem in $\mathbb{R}^d$ as follows: Consider the matrix $M \in \mathbb{R}^{n\times(d+1)}$ where the first d columns are A, and the last column is b, and for any $x \in \mathbb{R}^d$, define the vector $y \in \mathbb{R}^{d+1}$ where the first d entries of y are x, and the last entry is -1. Then $\|Ax - b\| = \|My\|$. Using Theorem 3 on the $(d+1)$-dimensional column space of M, we obtain an ε-subspace embedding matrix S of $O(d/\varepsilon^2)$ rows. We have

$$\|SMy\| \approx_\varepsilon \|My\| \implies \|S(Ax - b)\| \approx_\varepsilon \|Ax - b\| \implies \|SAx - Sb\| \approx_\varepsilon \|Ax - b\|.$$

Hence we can simply store SA instead of A (this is useful, if e.g. A is arriving online, column by column), and solve the resulting regression problem to get a good approximation for the original instance.

> **Remark.** If we are actually using the JLL to solve an algorithmic problem, the matrices with independent gaussians are not the most performant choices ($\mathbf{S}$ is likely to be dense, for example). There are many works (e.g. [AC09]) which show JLL-like statements for matrices that are much more efficient to work with.

You can find several more remarkable examples of applications of JLL and subspace embedding to various problems, while compressing the input in [CW09].

References

[AC09] Nir Ailon and Bernard Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009. 6

[CW09] Kenneth L Clarkson and David P Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 205–214, 2009. 6