CS 15-851: Algorithms for Big Data	Spring 2024

Lecture Lecture 02 — Jan. 25

Prof. David Woodruff — *Scribe: Noah Singer*

Notes: Please prove the fact given as a hint in Homework 1, Problem 1.

1 Recap: Subspace embeddings

Last week, we defined an important property for matrices:

Definition (Subspace embeddings)**.** Let $A \in \mathbb{R}^{n\times d}$ and $\epsilon > 0$. A matrix $S \in \mathbb{R}^{k\times n}$ is an *ϵ-subspace embedding for A* if for all $x \in \mathbb{R}^d$,

$$\|SAx\|_2 \in (1 \pm \epsilon)\|Ax\|_2,$$

where $\alpha \in (1 \pm \epsilon)\beta$ denotes $(1 - \epsilon)\beta \leq \alpha \leq (1 + \epsilon)\beta$.

Then, we showed a specific construction of such matrices:

Theorem 1 (Dense Gaussian matrices are subspace embeddings)**.** *Let $\epsilon > 0$, $k = O(d/\epsilon^2)$. Let $S \in \mathbb{R}^{k\times n}$ be a matrix of i.i.d. $\mathcal{N}(0, 1/k)$ random variables. Then for any fixed $A \in \mathbb{R}^{n\times d}$, S is an ϵ-subspace embedding w.h.p. over the choice of S.*

To recap, the proof of this theorem went as follows:

1. Assume WLOG that A has orthonormal columns and $\|x\|_2 = 1$.
2. Show that SA has the distribution of a $k \times d$ matrix of i.i.d. $\mathcal{N}(0, 1/k)$ random variables using sum and independence properties of Gaussians.
3. Show that for a fixed unit vector x, w.h.p. $\|SAx\|_2^2 \in 1 \pm \epsilon$ (in particular, w.p. $1 - 2^{-\Omega(d)}$). Specifically, this step uses a concentration inequality for sums of squared Gaussians ("χ^2 random variables) called the "Johnson-Lindenstrauss theorem".
4. To extend the concentration of norm to *all* x simultaneously, first build a *γ-net* for the subspace: A set of vectors M such that for all unit vectors x, there exists $y \in M$ with $\|y - Ax\|_2 \leq \gamma$. This uses a greedy construction, with a ball-volume-packing argument to show that the net cannot grow too large. Crucially, the size of this net is independent of n (just exponential in d).
5. Apply a union bound to show that w.h.p., the concentration holds for the lengths of all vectors in the net M and their differences.
6. Conditioning on this event, use a "chaining argument" to show that any vector Ax is an infinite linear combination of the net vectors with geometrically decreasing weights. Under this assumption and the concentration, Ax's length must also be approximately preserved.

2 From subspace embedding to linear regression

Recall, our original goal was to show that the solution to a linear regression problem $\min_y \|Ay - b\|_2$ is well-approximated by the solution to the (smaller) sketched problem $\min_y \|SAy - Sb\|_2$. We make the following claim:

Theorem 2 (Linear regression from subspace embeddings)**.** *Let $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^d$. Let A' denote the $n \times (d+1)$ matrix adjoining b (as a column vector) to the right of A. If $S \in \mathbb{R}^{k\times n}$ is a ϵ-subspace embedding for A', then $y^* = \arg\min_y \|SAy - Sb\|_2$ satisfies $\|Ay^* - b\|_2 \le (1+\epsilon)\|Ay - b\|_2$.*

Proof. Let $\widehat{y} = \arg\min_y \|Ay - b\|_2$. We claim that

$$(1-\epsilon)\|Ay^* - b\|_2^2 \le \|SAy^* - Sb\|_2^2 \tag{1}$$

and

$$\|SA\widehat{y} - Sb\|_2^2 \le (1+\epsilon)\|A\widehat{y} - b\|_2^2. \tag{2}$$

Given these, we deduce:

$$\begin{aligned}
\|Ay^* - b\|_2^2 &\le \frac{\|SAy^* - Sb\|_2^2}{1-\epsilon} && (1)\\
&\le \frac{\|SA\widehat{y} - Sb\|_2^2}{1-\epsilon} && (y^* \text{ minimized sketched problem})\\
&\le \frac{(1+\epsilon)\|A\widehat{y} - b\|_2^2}{1-\epsilon} && (2)\\
&\le (1+3\epsilon)\min_y \|Ay - b\|_2^2 && (\text{holds for } \epsilon < 1/3)
\end{aligned}$$

as desired.

Now, let x^* denote the vector consisting of y^* followed by a single 1 entry. Thus, $Ay^* - b = A'x^*$ and $SAy^* - Sb = SA'x^*$. Thus 1 follows from the subspace embedding property of S applied to x^*. Similarly, letting $\widehat{x}$ denote the vector consisting of $\widehat{y}$ followed by a single 1 entry, we get 2. ∎

The main advantage of this approach is that the solution to the sketched regression problem can be found efficiently (since it is only k-dimensional).

However, calculating SA might still expensive as it is a matrix product — indeed, if S is a random Gaussian matrix, we don't know a fast way to compute SA. (Note that in the proof for subspace embedding, we assumed WLOG that A has orthonormal columns, and we showed that under this assumption, SA has an i.i.d. Gaussian distribution. However, for general A this will not be true, and putting it in orthonormal form is expensive. Conversely, if A is already orthonormal, its pseudoinverse is just its transpose, meaning regression is very easy. So the key idea of subspace embedding is arguably that we can make A have orthonormal columns only in the analysis and not in the algorithm itself.)

3 A better subspace embedding from the Hadamard transform

Can we design a subspace embedding such that calculating the sketched matrix SA is inexpensive? One construction is due to Sárlos [1], called the *subsampled randomized Hadamard transform (SRHT)*. This is another distribution over sketch matrices S; we will show that it is a subspace embedding and that SA is efficiently calculable.

Say n is a power of 2 WLOG. The SRHT matrix $S = PHD$ is a product of three matrices P, H, D, defined as:

- P is an $s \times n$ *random* matrix. Its entries are all 0 except for a uniformly randomly and independently placed $\sqrt{n/s}$ entry in each row.
- H is an $n \times n$ *deterministic* matrix, the (normalized) Hadamard matrix $H_{i,j} = (1/\sqrt{n})(-1)^{\langle i,j \rangle}$ where i, j are viewed as vectors in $\mathbb{F}_2^{\log n}$.
- D is an $n \times n$ *random* diagonal matrix, where each diagonal entry is uniformly and independently sampled from $\{\pm 1\}$.

This distribution also has the advantage that all probability distributions involved are *discrete*.

3.1 Efficient calculation of the embedding

First, we claim that given P, H, D, A, we can output the product $SA = PHDA$ in $O(nd \log n)$ time.[1] This is almost optimal if A is dense, since A is $n \times d$ and it takes $\Theta(nd)$ time to even examine all its entries!

To prove this, we look at each of the d columns of A individually; that is, for P, H, D, a, we can calculate the matrix-vector product Sa in $O(n \log n)$ time. This is because:

- For any a, computing Da takes $O(n)$ time. (Indeed, D's diagonal entries are just ± 1, so D is just re-signing the entries of a.)
- The *fast Hadamard transform* (a.k.a. the fast Fourier transform over $\mathbb{F}_2$) means that you compute any matrix-vector product with H in $O(n \log n)$ time. (This is a divide-and-conquer algorithm using the recursive structure over H: If H' is the $n/2 \times n/2$ Hadamard matrix, then $H = \begin{pmatrix} H' & H' \\ H' & -H' \end{pmatrix}$. So, to compute Ha, we can recursively compute $H'a'$ and $H'a''$ where a' and a'' denote the first and last $n/2$ entries of a, respectively.)
- For any a, computing Sa takes $O(d)$ time. This is because the i-th entry of the product Sa is just the product between S's i-th row and a; S's i-th row only has one nonzero entry, and so the i-th entry of Sa is just an (appropriately scaled) entry of a.

[1] This is a worst-case problem: P can be any matrix with exactly one $\sqrt{n/s}$ entry in each row, and D any diagonal matrix with ± 1's on the diagonal.

3.2 Proof of subspace embedding

Now, we prove the result that the SRHT process gives a subspace embedding.

Theorem 3 (SRHT matrices are subspace embeddings)**.** *Let $\epsilon > 0, s = O(d/\epsilon^2)$. Let $S \in \mathbb{R}^{s\times n}$ be sampled according to the SRHT process described at the beginning of this section. Then for any fixed $A \in \mathbb{R}^{n\times d}$, S is an ϵ-subspace embedding w.h.p. over the choice of S.*

In the proof, we will need the following concentration inequality:

Lemma 1 (Azuma-Hoeffding)**.** *Let $\{Z_j\}_{j\in[n]}$ be a set of independent random variables such that for all $j \in [n]$, $\mathbb{E}[Z_j] = 0$ and $|Z_j| \le \beta_j$ (w.p. 1). Then*

$$\mathbb{P}\left[\left|\sum_{j=1}^{n} Z_j\right| \ge t\right] \le 2\exp\left(\frac{-t^2}{\sum_{j=1}^{n}\beta_j^2}\right).$$

To prove that S is a subspace embedding, we can again assume WLOG that A has orthonormal columns and x is a unit vector. Let $y = Ax$; $y \in \mathbb{R}^n$ is a unit vector since A has orthonormal columns. In this setting, we WTS that $\|PHDy\|_2^2 \in 1 \pm \epsilon$ (i.e., w.h.p. this holds for *all* unit y in the column space of A). Note that D, H are orthogonal: D since it is diagonal with ± 1 entries; and H by definition of the Hadamard matrix (the inner product between two columns in H is $\langle H_{*i}, H_{*j}\rangle = (1/n)\sum_{k=1}^{n}(-1)^{\langle k,i\rangle+\langle k,j\rangle} = (1/n)\sum_{k=1}^{n}(-1)^{\langle k,i+j\rangle}$, and $i+j$ is a nonzero vector since $i \ne j$, and one can setup a bijection based on toggling some coordinate on which i and j disagree).

So why do we need H and D? If they weren't there, and our sketch was just $S = P$, recall that each entry of Sy will just be a randomly selected (appropriately scaled) entry of y. But it's quite possible that y is a very sparse vector, in which case Sy would have decent probability of being all-0's — this is bad, because then the norm is not preserved. (Note that the *expected norm* of Sy is still 1, but the variance is too high.) H and D somehow "spread out y's norm roughly evenly over all the coordinates" by randomly rotating y; this allows a better bound on the concentration of $\|PHDy\|_2$. (Note that there will still be some vectors z such that HDz is sparse (HD just rotates the space, after all), but this is OK as long as they aren't A's column span.)

The fact that HD "spreads out" y's mass over its coordinates is formalized in the following lemma. Recall that for a vector $z \in \mathbb{R}^n$, the ∞*-norm* $\|z\|_\infty$ denotes the magnitude of the largest entry of z. Then:

Lemma 2 (Flattening)**.** *There exists $C > 0$ such that the following holds. Let $y \in \mathbb{R}^n$ be any unit vector and $H \in \mathbb{R}^{n\times n}$ any matrix where all entries are $\pm 1/\sqrt{n}$. If $D \in \mathbb{R}^{n\times n}$ is diagonal with uniform and independent $\{\pm 1\}$ entries, then*

$$\mathbb{P}_D\left[\|Hdy\|_\infty \ge C\sqrt{\frac{\log(nd/\delta)}{n}}\right] \le \frac{\delta}{2d}.$$

Proof. We claim that for fixed $i \in [n]$,

$$\mathbb{P}_D\left[|(HDy)_i| \ge C\sqrt{\frac{\log(nd/\delta)}{n}}\right] \le \frac{\delta}{2nd}.$$

This gives the desired result taking a union bound over i.

To prove the claim, we expand the matrix-(matrix-)vector product

$$(HDy)_i = \sum_{j=1}^{n} H_{i,j} D_{j,j} y_j.$$

Let $Z_j := H_{i,j} D_{j,j} y_j$, so that $(HDy)_i = \sum_{j=1}^n Z_j$. Here are some properties of Z_j: $\mathbb{E}[Z_j] = 0$ (since $H_{i,j}$ and y_j are fixed, and $D_{j,j}$ is uniform in $\{\pm 1\}$); and $|Z_j| \le |y_j|/\sqrt{n}$ w.p. 1 over D, since $|H_{i,j}| = 1/\sqrt{n}$ by assumption. Defining $\beta_j := |y_j|/\sqrt{n}$, we have $\sum_{j=1}^n \beta_j^2 = \frac{1}{n}\sum_{j=1}^n y_j^2 = \frac{1}{n}$ since we assumed y is a unit vector. So we can apply the Azuma-Hoeffding lemma (1) with $t = C\sqrt{\log(nd/\delta)/n}$, giving concentration of $\exp(-t^2/(1/n)^2) = \exp(-C^2 \log(nd/\delta)) < \frac{\delta}{2nd}$ for sufficiently large C. ■

Note that the only fact about the Hoeffding matrix required for this lemma is that all its entries have magnitude $1/\sqrt{n}$. Indeed, the SRHT process uses the Hadamard matrix simply because it (i) has this bounded-entry property, (ii) is orthogonal, and (iii) support efficient matrix-vector products due to the recursive structure.

Since HDA has orthonormal columns, taking any particular column a of A, by Lemma 2 we have $\|HDa\|_\infty \le C\sqrt{\log(nd/\delta)/n}$ w.p. $1 - \delta/(2d)$ over the choice of D. Taking union bound over the d columns of A, we have that $\|HDA\|_\infty \le C\sqrt{\log(nd/\delta)/n}$ w.p $1 - \delta/2$ over the choice of D, where $\|Z\|_\infty$ denotes the largest entry in the entire matrix. So for any *row* $(HDA)_{i*}$ of HDA, we have $\|(HDA)_{i*}\|_2^2 \le C^2 d \log(nd/\delta)/n$ (since there are d entries, each of magnitude at most $C\sqrt{d\log(nd/\delta)/n}$).

Recall that for a matrix $Z \in \mathbb{R}^{n\times m}$, the *Frobenius norm* is $\|Z\|_F := \sqrt{\sum_{i=1}^n \sum_{j=1}^m Z_{ij}^2}$. In this case, $\|HDA\|_F^2 = d$ since each of the columns are orthonormal; thus, the *average* value of $\|(HDA)_{i*}\|_2^2$ over $i \in [n]$ is d/n. So, Lemma 2 implies that no row has much larger norm than average (not much larger than a log factor).

Now, we condition on this event that the norms of rows in HDA are bounded, and we want to deduce that PHD is a good subspace embedding for A, w.h.p. over the choice of P. For this, we'll need another technical tool, the *matrix Chernoff inequality*, which bounds the spectral norm of a random matrix. Recall, the spectral norm of a matrix $Z \in \mathbb{R}^{n\times m}$ is $\|Z\|_2 := \sup_{\|x\|_2=1} \|Zx\|_2$ (and $\|Z\|_2^2$ also equals the largest singular value of Z). In this notation, we want to show e.g. that $\|SA\|_2 \le 1+\epsilon$ (so that SA does not increase the norm of any unit vector by a factor more than $1+\epsilon$).

References

[1] Tamás Sárlos. Improved Approximation Algorithms for Large Matrices via Random Projections. In *47th Annual IEEE Symposium on Foundations of Computer Science*, pages 143–152, October 2006. `doi:10.1109/FOCS.2006.37`.