CS 15-851: Algorithms for Big Data Spring 2024

Lecture 1 — 1/18

Prof. David Woodruff *Scribe: Hoai-An Nguyen*

1 Sketching to Approximate Least Squares Regression

We saw that we can solve least squares regression using singular value decomposition (SVD). However, we desire a faster runtime compared to the naive $O(nd^2)$ (or even the $O(nd^{1.37})$ fast matrix multiplication approach). Therefore, we will use sketching to achieve our faster runtime while settling for an approximation.

Goal: We want to find some x' such that $|Ax' - b|_2 \leq (1+\epsilon)\min_x |Ax - b|_2$ with high probability.

To do this, we will draw our sketching matrix S from a $k \times n$ random family of matrices such that $k << n$. Then, we will compute SA and Sb and use singular value decomposition to output $\min_{x'} |(SA)x - (Sb)|_2$.

Obviously, we will not accomplish our goal of a good approximation for just any arbitrary S. So how do we choose S? We will first consider when S is a $d/\epsilon^2 \times n$ matrix of iid normal random variables [1].

Now, we must prove that this sketch gives us a valid approximation. For this proof, we will use the notion of subspace embeddings.

1.1 Subspace Embeddings

Let S be a $k \times n$ matrix of iid normal random variables drawn from $N(0, 1/k)$ where $k = O(d/\varepsilon^2)$. We want to show that we have a valid subspace embedding. Or in other words, that for any fixed d-dimensional subspace, with high probability, for *all* $x \in \mathbb{R}^d$, $|SAx|_2 = (1 \pm \epsilon)|Ax|_2$. This means that the column space of A is preserved.

Theorem 1. *Let S be a $k \times n$ matrix of iid normal random variables drawn from $N(0, 1/k)$ where $k = O(d/\varepsilon^2)$. For any fixed d-dimensional subspace, with high probability, for* all *$x \in \mathbb{R}^d$, $|SAx|_2 = (1 \pm \epsilon)|Ax|_2$.*

Let us now prove this. First, we make the assumption without loss of generality that the columns of A are orthonormal. This is because we are proving this statement for *all* x. Therefore, $|SAx|_2 = |SU\Sigma V^\intercal|_2$ where our new $A' = U$ and $x' = \Sigma V^\intercal x$ where by SVD, U has orthonormal columns. Here, A' and x' denote our "new" A and x.

We first prove the following claim.

Claim 1. SA is a $k \times d$ matrix of iid random variables drawn from $N(0, 1/k)$.

[1] We note that lots of matrices will work to accomplish our goal. However, this will be a simpler example that will showcase some of the techniques we will continue to use.

Proof. To prove this, we establish two properties of Gaussian distributions.

The first is that for two independent random variables X and Y with X drawn from $N(0, a^2)$ and Y drawn from $N(0, b^2)$, $X + Y$ is drawn from $N(0, a^2 + b^2)$. To see this, note that the probability density function f_Z where $Z = X + Y$ is the convolution of probability density functions f_X and f_Y. We have by definition

$$f_Z(z) = \int f_X(z - y) f_Y(y) dy$$
$$f_X(x) = \frac{1}{a(2\pi)^{0.5}} e^{-x^2/2a^2},\ f_Y(y) = \frac{1}{b(2\pi)^{0.5}} e^{-y^2/2b^2}.$$

Therefore, we can see that

$$\begin{aligned} f_Z(z) &= \int \frac{1}{a(2\pi)^{0.5}} e^{-(z-y)^2/2a^2} \frac{1}{b(2\pi)^{0.5}} e^{-y^2/2b^2} dy \\ &= \frac{1}{(2\pi)^{0.5}(a^2+b^2)^{0}.5} e^{-z^2/2(a^2+b^2)} \int \frac{(a^2+b^2)^{0.5}}{(2\pi)^{0.5}ab} e^{-\frac{(y-b^2z/a^2+b^2)^2}{2((ab)^2/a^2+b^2)}} dy. \end{aligned}$$

The integral evaluates to 1 since this is a probability density function. Then, the rest we can observe is the density function of a Gaussian. So we have proved our first property.

Our second property is rotational invariance which says that if u, v are vectors with $\langle u, v\rangle = 0$, then $\langle g, u\rangle$ and $\langle g, v\rangle$ are independent, where g is a vector of iid $N(0, 1/k)$ random variables.

Why is this? So if g is an n-dimensional vector of iid random variables drawn from $N(0, 1)$, and we fix some matrix R, then the probability density function of Rg is defined as

$$f(x) = \frac{1}{\det(RR^\intercal)(2\pi)^{n/2}} e^{-x^\intercal (RR^\intercal)^{-1} x/2}$$

where $RR^\intercal$ is the covariance matrix. Since we have that for some rotational matrix R that $RR^\intercal = I$ where I is the identity matrix, the distribution of Rg and g are the same.

Now, choose a rotation R which rotates u to αe_1. This is some constant scaling the first standard basis vector. Similarly, R should rotate v to βe_2. Now, define h to be some vector of iid random variables drawn from $N(0, 1/k)$. We can see that

$$\langle g, u\rangle = \langle Rg, Ru\rangle = \langle h, \alpha e_1\rangle = \alpha h_1 \tag{1}$$
$$\langle g, v\rangle = \langle Rg, Rv\rangle = \langle h, \beta e_2\rangle = \beta h_2. \tag{2}$$

Here h_1 and h_2 are the first and second entry in h, respectively. By the definition of h, they are independent, and therefore, we have showed our second property. Now we have everything we need to finish the proof of the claim.

So, we know that the rows of SA are independent since the rows of S are independent. Now, look at each row. Each row looks like $\langle g, A_1\rangle, \langle g, A_2\rangle, ...\langle g, A_d\rangle$. We know by assumption that the columns A_i are orthonormal. Therefore, the entries in each row are independent (by our second property). Furthermore, by our first property since the columns A_i have unit norm since they are orthonormal, we know that the entries in each row are drawn from $N(0, 1/k)$.

■

Let us now continue proving our main theorem concerning subspace embeddings. We will add the additional assumption, again without loss of generality, that x is a unit vector. From our claim, we know that SA is a $k \times d$ matrix of iid random variables drawn from $N(0, 1/k)$.

Look at any fixed unit vector $x \in \mathbb{R}^d$. We have that

$$|SAx|_2^2 = \sum_{i \in [k]} \langle g_i, x \rangle^2$$

where g_i is the i^{th} row of SA. We have proven that for all the i that we consider, each $\langle g_i, x \rangle^2$ is distributed as $N(0, 1/k)^2$. Therefore, $\mathbb{E}[\langle g_i, x \rangle^2] = 1/k$ and therefore $\mathbb{E}[|SAx|_2^2] = 1$. Remember that by assumption the norm of Ax is 1. So, we are partially there. We have shown that we get the right expected value. We now need to show that $|SAx|_2^2$ is concentrated around its expectation. This will give us the high probability bound we need.

First, we will prove that with high probability we get an appropriate approximation for a *fixed* x, then we will generalize it to all x. We use the following theorem.

Theorem 2. *Johnson-Lindenstrauss Theorem Define $G = \sum_i h_i^2$ is a χ^2 random variable where $h_1, ..., h_k$ are iid random variables drawn from $N(0, 1)$. Then, $\mathbb{P}[G \geq k + 2(kx)^{0.5} + 2x] \leq e^{-x}$ and $\mathbb{P}[G \leq k - 2(kx)^{0.5}] \leq e^{-x}$.*

Therefore, if we set $x = \epsilon^2 k/16$, then $\mathbb{P}[G \in k(1 \pm \epsilon)] \geq 1 - 2e^{\epsilon^2 k/16}$. Set $k = \theta(\epsilon^{-2} \log(1/\delta))$, then the probability is at least $1 - \delta$. So, we should think about setting $d = \log(1/\delta)$. So, for fixed x, we have

$$\mathbb{P}[|SAx|_2^2 \in (1 \pm \epsilon)] \geq 1 - 2^{-\theta(d)}.$$

Now, we generalize this to all x. We first introduce a net for a sphere.

1.1.1 Net Sphere

Consider the sphere S^{d-1}.

Definition. A subset N is a γ-net if for all $x \in S^{d-1}$, there is a $\gamma \in N$ such that $|x - y|_2 \leq \gamma$.

We will greedily construct N. Do the following. While there is a point $x \in S^{d-1}$ that has a distance larger than γ from *every* point in N, add x to N. How do we know that this algorithm will terminate? Well, around each point in N, imagine balls of radius $\gamma/2$ with the point as the center. We can see that the balls are disjoint by the construction of N. If they are not, then we have added a point to N that was already within distance γ of an existing point in N. Therefore, there are a finite number of these balls we can pack within the sphere, and therefore a finite number of points that our algorithm adds to N before terminating. Specifically, we can think of the balls of radius $\gamma/2$ around each point in N is contained within one larger ball of radius $1 + \gamma/2$ centered around the origin. Therefore, the ratio of the volume of the d-dimensional ball of radius $1 + \gamma/2$ to the d-dimensional sphere of radius γ is $(1 + \gamma/2)^d/(\gamma/2)^d$ and so $|N| \leq (1 + \gamma/2)^d/(\gamma/2)^d$. We like to see the 2^d factors here (reference the probability bound we got for a fixed x above).

1.1.2 Net for Subspace

We considered a net for a sphere above since a sphere is a simple object. Now we extend this to form a net for a subspace, which is what we need to complete our main theorem.

Let $M = \{Ax | x \in N\}$ so $|M| \leq (1 + \gamma/2)^d/(\gamma/2)^d$.

Claim 2. For every $x \in S^{d-1}$, there is a y in M such that $|Ax - y|_2 \leq \gamma$.

Proof. Let x' in N be such that $|x - x'|_2 \leq \gamma$. Then, $|Ax - Ax'|_2 = |x - x'|_2 \leq \gamma$ using the fact that the columns of A are orthonormal. Now, set $y = Ax'$. ■

1.2 Net Argument

Now, we use a net argument to finish up the proof of our main theorem. Recall that we proved for some fixed x, $\mathbb{P}[|SAx|_2^2 \in (1 \pm \epsilon)] \geq 1 - 2^{-\theta(d)}$. Note therefore that for a fixed pair x, x', $|SAx|_2^2$, $|SAx'|_2^2$, and $|SA(x - x')|_2^2$ are preserved up to a $(1 \pm \epsilon)$-factor with probability $1 - 2^{-\theta(d)}$. So, consider the following two quantities.

$$|SA(x - x'|_2^2 = |SAx|_2^2 + |SAx'|_2^2 - 2\langle SAx, SAx'\rangle \tag{3}$$

$$|A(x - x'|_2^2 = |Ax|_2^2 + |Ax'|_2^2 - 2\langle Ax, Ax'\rangle. \tag{4}$$

We can conclude that $\mathbb{P}[\langle Ax, Ax'\rangle = \langle SAx, SAx'\rangle \pm O(\epsilon)] \geq 1 - 2^{-\theta(d)}$.

Now, choose a 1/2-net $M = \{Ax | x \in N\}$ of size 5^d. Using a union bound, for all pairs y, y' in M, $\langle y, y'\rangle = \langle Sy, Sy'\rangle + O(\epsilon)$. If we condition on this event, then by linearity, if it holds for y, y' in M, then for $\alpha y, \beta y'$, $\langle \alpha y, \beta y'\rangle = \alpha\beta\langle Sy, Sy'\rangle \pm O(\epsilon\alpha\beta)$.

Now, let $y = Ax$ for some arbitrary $x \in S^{d-1}$. Let $y_1 \in M$ be such that $|y - y_1|_2 \leq \gamma$. Let α be such that $|\alpha(y - y_1)|_2 = 1$. Note here that $\alpha \geq 1/\gamma$ which could be infinite, and if so, we should stop at this point. Otherwise, let $y_2' \in M$ be such that $|\alpha(y - y_1) - y_2'|_2 \leq \gamma$ which gives us

$$|y - y_1 - y_2'/\alpha| \leq \gamma/\alpha \leq \gamma^2.$$

Now, set $y_2 = y_2'/\alpha$. Keep on repeating to get $y_1, y_2, y_3, ...$ We can see that

$$|y - y_1 - y_2 - y_3 ... - y_i|_2 \leq \gamma^i$$

which gives us

$$|y_i|_2 \leq \gamma^{i-1} + \gamma^i \leq 2\gamma^{i-1}$$

by the triangle inequality.

So, we have $y = \sum_i y_i$. Therefore,

$$\begin{aligned} |Sy|_2^2 &= |S\sum_i y_i|_2^2 \\ &= \sum_i |Sy_i|_2^2 + 2\sum_{i,j}\langle Sy_i, Sy_j\rangle \\ &= \sum_i |y_i|_2^2 + 2\sum_{i,j}\langle yi, yj\rangle + O(\epsilon)\sum_{i,j}|y_i|_2|y_j|_2. \end{aligned}$$

Since we see that $|y_i|_2$ is a geometric sequence, it converges to sum constant. Therefore, we further have

$$\begin{aligned} |Sy|_2^2 &= |\sum_i y_i|_2^2 \pm O(\epsilon) \\ &= |y|_2^2 \pm O(\epsilon) = 1 \pm O(\epsilon). \end{aligned}$$

Since we proved this for any arbitrary $y = Ax$ for unit vector x, by linearity, we have that for all x, $|SAx|_2 = (1 \pm \epsilon)|Ax|_2$.