CS 15-851: Algorithms for Big Data Spring 2025

Lecture 1 Part 2 — Jan 16

Prof. David Woodruff *Scribe: Samvitti Sharma*

1 Improving the time complexity of least squares regression

In Part 1 of this lecture, we discussed solving least squares regression via normal equations. Given a $n \times d$ matrix A, where $n \geq d$, and a vector $\mathbf{b} \in \mathbb{R}^n$, an optimal least squares solution is $\mathbf{x} \in \mathbb{R}^n$ such that $|A\mathbf{x} - \mathbf{b}|_2$ is minimized. We wish to find the optimal solution $\mathbf{x}$ with the smallest norm amongst all optimal solutions. We saw in Part 1 that to do so, we can simply compute $\mathbf{x} = A^{-1}\mathbf{b}$, where A^- is the Moore-Penrose pseudoinverse of A.

What are some limitations of this method? Computing SVD is slow: naively, its runtime is $O(nd^2)$. We can improve it to $O(nd^{1.376})$ using fast matrix multiplication, but this is still much too slow. Our motivation for Part 2 of this lecture is to achieve a better running time for least squares regression.

To achieve this, we will first relax our goal to finding an *approximate* solution $\mathbf{x}$ to $\min_{\mathbf{x}} |A\mathbf{x} - \mathbf{b}|_2$.

Relaxed goal.
Output $\mathbf{x}'$ for which $|A\mathbf{x}' - \mathbf{b}|_2 \leq (1+\epsilon) \min_{\mathbf{x}} |A\mathbf{x} - \mathbf{b}|_2$ with high probability, for some small $\epsilon \in \mathbb{R}$.

To achieve this goal, we're going to apply a technique called *sketching*, where we apply randomization to reduce the dimensions of our original matrices and then proceed to compute least squares regression for the smaller matrices. This achieves a better running time while still maintaining a good approximation of the solution to our original matrices. An outline of this procedure is:

1. Choose a $k \times n$ matrix S from a random family of matrices, where $k << n$.
2. Compute SA, which is a $k \times d$ matrix. Thus A has been "squashed down" into a smaller matrix.
3. Compute $S\mathbf{b}$, which is in $\mathbb{R}^k$. Thus $\mathbf{b}$ has been "squashed down" into a smaller vector.
4. Now compute $\mathbf{x}' = (SA)^-(S\mathbf{b})$, which minimizes $|(SA)\mathbf{x} - (S\mathbf{b})|_2$.

Note that the fourth step can be computed in $O(kd^2)$ naively, or $O(kd^{1.376})$ with fast matrix multiplication.

The goal for the rest of lecture is to show that this $\mathbf{x}'$ satisfies our relaxed condition: that $|A\mathbf{x}' - \mathbf{b}|_2 \leq (1+\epsilon) \min_{\mathbf{x}} |A\mathbf{x} - \mathbf{b}|_2$ with high probability.

First, we must think about how to choose an appropriate sketching matrix S: what random family of matrices should we use? While many families work, a natural choice is to draw all entries of S independently from a normal distribution. We're also going to pick $k = \frac{d}{\epsilon^2}$, and we're going to show that this choice of S and k works for our purposes via a concept called *subspace embeddings*.

1.1 Subspace Embeddings

Theorem 1.
Let $k = O(\frac{d}{\epsilon^2})$. Let S be a $k \times n$ matrix of i.i.d normal $\mathcal{N}(0, \frac{1}{k})$ random variables.
For any fixed d-dimensional subspace, i.e. the column space of a $n \times d$ matrix A, for all vectors $\mathbf{x} \in \mathbb{R}^d$, $|SA\mathbf{x}|_2 = (1 \pm \epsilon)|A\mathbf{x}|_2$ with high probability.

This is equivalent to showing that S is a *subspace embedding* of the column space of A. This also means that the column space of A is preserved.

How do we prove Theorem 1?
First, we can assume that the columns of A are orthonormal. This is because we can write A in its SVD form: $A = U\Sigma V^T$. We can substitute this into Theorem 1, so we WTS that $|SU\Sigma V^T\mathbf{x}|_2 = (1 \pm \epsilon)\,|U\Sigma V^T\mathbf{x}|_2$. Since we are proving Theorem 1 for all $\mathbf{x} \in \mathbb{R}$, we can do a change of variables $\mathbf{y} = \Sigma V^T\mathbf{x}$. So now we can simply show that $|SU\mathbf{y}|_2 = (1 \pm \epsilon)\,|U\mathbf{y}|_2$. Because U has orthonormal columns, we can make the assumption that the columns of A are orthonormal.

To prove Theorem 1, we first need to prove the following lemma:

Lemma 1.
SA is a $k \times d$ matrix of i.i.d $\mathcal{N}(0, \frac{1}{k})$ random variables.

We need to prove two properties to prove Lemma 1.

Property 1.
Given two independent random variables X and Y, where X is drawn from $\mathcal{N}(0, a^2)$, and Y is drawn from $\mathcal{N}(0, b^2)$, $X + Y$ is drawn from $\mathcal{N}(0, a^2 + b^2)$.

Proof of Property 1.
First, note that the probability density function f_z of $Z = X + Y$ is the convolution of the probability density functions f_x and f_y:

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

Plugging in $f_X(x), f_Y(y)$ into $f_Z(z)$, we get:

$$f_Z(z) = \int \frac{1}{a(2\pi)^{0.5}} e^{-\frac{(z-y)^2}{2a^2}} \frac{1}{b(2\pi)^{0.5}} e^{-\frac{y^2}{2b^2}}\, dy$$
$$= \frac{1}{(2\pi)^{0.5}(a^2+b^2)^{0.5}} e^{\frac{-z^2}{2(a^2+b^2)}} \int \frac{(a^2+b^2)^{0.5}}{(2\pi)^{0.5}ab} e^{\frac{\left(y - \frac{b^2 z}{a^2+b^2}\right)^2}{2\left(\frac{(ab)^2}{a^2+b^2}\right)}}\, dy$$

This integral on the right is the integral of another probability density function: that of a random variable $\sim \mathcal{N}(\frac{b^2 z}{a^2+b^2}, \frac{(ab)^2}{a^2+b^2})$. So by the fact that any probability density function integrated over the whole real line evaluates to 1, the right hand side of this multiplication is equal to 1.

So we derive that

$$f_Z(z) = \frac{1}{(2\pi)^{0.5}(a^2+b^2)^{0.5}} e^{\frac{-z^2}{2(a^2+b^2)}}$$

This corresponds to the probability density function of a random variable $\sim \mathcal{N}(0, a^2 + b^2)$. So we have shown that $Z = X + Y$ is drawn from $\mathcal{N}(0, a^2 + b^2)$. $\square$

Property 2.
If $\mathbf{u}, \mathbf{v}$ are vectors with $\langle \mathbf{u}, \mathbf{v} \rangle = 0$, and you have a Gaussian vector $\mathbf{g}$ with i.i.d $\mathcal{N}(0, \frac{1}{k})$ random variables as its entries, then $\langle \mathbf{g}, \mathbf{u} \rangle$ and $\langle \mathbf{g}, \mathbf{v} \rangle$ are independent.

This property is also called the *rotational invariance* of Gaussian distributions.

Proof of Property 2.
Let $\mathbf{g}$ be a n-dimensional vector of i.i.d $\mathcal{N}(0, 1)$ random variables, and let R be a fixed matrix. Then the probability density function of $R\mathbf{g}$ is:

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

Note that RR^T is also called the covariance matrix.

What happens if R is a rotation matrix, meaning it has orthonormal rows and columns?
RR^T will evaluate to the identity matrix, and $\det(RR^T)$ will evaluate to 1, so we end up with the probability density function corresponding with $\mathbf{g}$, meaning $R\mathbf{g}$ and $\mathbf{g}$ have the same distribution. So rotations don't change the distribution.

This means we can choose a rotation R that sends $\mathbf{u}$ to $\alpha \cdot \mathbf{e_1}$, where $\mathbf{e_1}$ is the first standard basis vector, and sends $\mathbf{v}$ to $\beta \cdot \mathbf{e_2}$, where $\mathbf{e_2}$ is the second standard basis vector.

Let $\mathbf{h}$ be a vector of i.i.d $\mathcal{N}(0, \frac{1}{k})$ random variables. We find that

$\langle \mathbf{g}, \mathbf{u} \rangle = \langle R\mathbf{g}, R\mathbf{u} \rangle = \langle \mathbf{h}, \alpha \mathbf{e_1} \rangle = \alpha \mathrm{h}_1$, and
$\langle \mathbf{g}, \mathbf{v} \rangle = \langle R\mathbf{g}, R\mathbf{v} \rangle = \langle \mathbf{h}, \beta \mathbf{e_2} \rangle = \beta \mathrm{h}_2$

Since h_1 and h_2 are different entries of a vector of independent Gaussians, they must be independent. So we have shown that $\langle \mathbf{g}, \mathbf{u} \rangle$ and $\langle \mathbf{g}, \mathbf{v} \rangle$ are independent. $\square$

Now that we've proved these 2 properties, we can proceed to prove Lemma 1.

Proof of Lemma 1.
We WTS that SA is a $k \times d$ matrix of i.i.d $\mathcal{N}(0, \frac{1}{k})$ random variables. First, note that an arbitrary entry of SA is derived by computing the dot product of a row of S and a column of A. Entries across rows of SA are independent, because their dot products involve different rows of S, which are independent by definition.
What about entries in the same row of SA? Consider an arbitrary row in SA. Its entries look like $\langle \mathbf{g}, A_1 \rangle$, $\langle \mathbf{g}, A_2 \rangle$, ..., $\langle \mathbf{g}, A_d \rangle$. Note that the columns A_i are orthonormal. This means that they have unit norm, so by the first property, each entry is drawn from $\mathcal{N}(0, \frac{1}{k})$. This also means that they are orthogonal to each other, so by the second property, the entries in this row are independent. Thus, we have shown Lemma 1. $\square$

Now, let's go back to Theorem 1. We WTS that S is a subspace embedding, meaning that $|SA\mathbf{x}|_2 = (1 \pm \epsilon)|A\mathbf{x}|_2$ for all $\mathbf{x}$.

Note that we can assume that $\mathbf{x}$ is a unit vector, since we can scale both sides of our equation by the norm of $\mathbf{x}$. We also know that the columns of A are orthonormal, so $|A\mathbf{x}|_2 = |\mathbf{x}|_2 = 1$. We have also shown that SA is a Gaussian matrix with i.i.d $\mathcal{N}(0, \frac{1}{k})$ random variables as its entries.

Let's first try to achieve a less ambitious goal: proving Theorem 1 for a fixed unit vector $\mathbf{x} \in \mathbb{R}^d$. Given a fixed unit vector $\mathbf{x}$, we derive that

$$|SA\mathbf{x}|_2^2 = \Sigma_{i\in[k]}\langle \mathbf{g}_i, \mathbf{x}\rangle^2, \text{where } \mathbf{g}_i \text{ is the } i\text{-th row of } SA.$$

We want to show that this sum is within $(1 \pm \epsilon)$ w.h.p, since $|A\mathbf{x}|_2 = 1$.

Note that each $\langle \mathbf{g}_i, \mathbf{x}\rangle^2 \sim \mathcal{N}(0, \frac{1}{k})^2$.

$\mathbb{E}[\langle \mathbf{g}_i, \mathbf{x}\rangle^2] = \text{Var}[\langle \mathbf{g}_i, \mathbf{x}\rangle] + \mathbb{E}^2[\langle \mathbf{g}_i, \mathbf{x}\rangle] = \frac{1}{k} + 0^2 = \frac{1}{k}$.

Therefore, by linearity of expectation, $\mathbb{E}[|SA\mathbf{x}|_2^2] = 1$. But how concentrated is $|SA\mathbf{x}|_2^2$ around its expectation? To answer this question, we can use the following theorem:

Johnson-Lindenstrauss Theorem.
Let $h_1, ..., h_k$ be i.i.d $\mathcal{N}(0, 1)$ random variables. Then $G = \sum_i h_1^2$ is a χ^2 random variable.
By applying tail bounds to G, we derive that:

$$\Pr[G \geq k + 2(kx)^{0.5} + 2x] \leq e^{-x}.$$

$$\Pr[G \leq k - 2(kx)^{0.5}] \leq e^{-x}.$$

If we choose $x = \frac{\epsilon^2 k}{16}$, then plugging that into both tail bounds and applying the union bound gives us:

$$\Pr[G \in k(1 \pm \epsilon)] \geq 1 - 2e^{-\frac{\epsilon^2 k}{16}}.$$

If we set $k = \Theta(\epsilon^{-2}\log(1/\delta))$, this probability is at least 1 - δ. Earlier, we set $k = d/\epsilon^2$, so if we set $d = \log(1/\delta)$, then $\delta = 2^{-d}$. So we get that:

$$Pr[|SA\mathbf{x}|_2^2 \in (1 \pm \epsilon)] \geq 1 - 2^{-\Theta(d)}.$$

1.2 Nets for Spheres and Subspaces

We're still not done, because we want to show that this property holds for all $\mathbf{x}$. But we have an infinite number of vectors to work with, and we can't apply the union bound to an infinite number of tail bounds. How can we convert an infinite number of events to a finite number of events that we can then apply the union bound to?
We're going to do so by using a *net*.

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

First, let's think about how we can construct a γ-net. We can use the following greedy algorithm:

1. Pick a random point in S^{d-1} and add it to N.
2. While there exists $\mathbf{x} \in S^{d-1}$ with distance larger than γ from every point in N, include $\mathbf{x}$ in N.

We need to show that this algorithm terminates. Imagine a ball of radius $\frac{\gamma}{2}$ around every point in N. All these balls are disjoint, because the distance between two corresponding points in N must be larger than γ by the definition of our greedy algorithm, so no two balls of radius $\frac{\gamma}{2}$ can intersect with each other. Also, all of these balls are contained within a larger ball of radius $1 + \frac{\gamma}{2}$ centered at 0^d (the origin). This is because the radius of the sphere is 1, so the distance between the origin and a point in N is at most 1. So we can only pack a finite number of balls within this larger ball, meaning that N has a finite size and this algorithm terminates.
Furthermore, the number of smaller balls multiplied by the volume of each smaller ball cannot exceed the volume of our larger ball of radius $1 + \frac{\gamma}{2}$ centered at the origin. This means that $|N| \cdot (\frac{\gamma}{2})^d \leq (1 + \frac{\gamma}{2})^d$, so $|N| \leq \frac{(1+\gamma/2)^d}{(\gamma/2)^d}$.

What we need for our subspace embedding proof is a net for a subspace rather than a sphere, which we will define as follows:
Let $M = \{A\mathbf{x} \mid \mathbf{x} \in N\}$, so $|M| \leq \frac{(1+\gamma/2)^d}{(\gamma/2)^d}$.

Claim.
For every $\mathbf{x}$ in S^{d-1}, there is a $\mathbf{y} \in M$ for which $|A\mathbf{x} - \mathbf{y}|_2 \leq \gamma$. So M is a net for the subspace.
Proof.
Pick an arbitrary $\mathbf{x} \in S^{d-1}$. Let $\mathbf{x}' \in N$ be such that $|\mathbf{x} - \mathbf{x}'|_2 \leq \gamma$. Then $|A\mathbf{x} - A\mathbf{x}'|_2 = |\mathbf{x} - \mathbf{x}'|_2 \leq \gamma$, since A has orthonormal columns. So we can set $\mathbf{y} = A\mathbf{x}'$, and $\mathbf{y}$ will be in M by the definition of M. $\square$

1.3 Net Argument

We've shown for a fixed unit vector $\mathbf{x}$, $\Pr[|SA\mathbf{x}|_2^2 \in (1 \pm \epsilon)] \geq 1 - 2^{-\Theta(d)}$. We want to generalize this to all unit vectors $\mathbf{x}$.

Note that for a fixed pair of unit vectors $\mathbf{x}, \mathbf{x}'$, the three quantities $|SA\mathbf{x}|_2^2$, $|SA\mathbf{x}'|_2^2$, and $|SA(\mathbf{x} - \mathbf{x}')|_2^2$ are all preserved up to a $(1 \pm \epsilon)$ factor with probability $1 - 2^{-\Theta(d)}$, by the union bound.
Also note that

$$|SA(\mathbf{x} - \mathbf{x}')|_2^2 = |SA\mathbf{x}|_2^2 + |SA\mathbf{x}'|_2^2 - 2\langle SA\mathbf{x}, SA\mathbf{x}'\rangle$$
$$|A(\mathbf{x} - \mathbf{x}')|_2^2 = |A\mathbf{x}|_2^2 + |A\mathbf{x}'|_2^2 - 2\langle A\mathbf{x}, A\mathbf{x}'\rangle$$

So we can derive that $\Pr[\langle A\mathbf{x}, A\mathbf{x}'\rangle = \langle SA\mathbf{x}, SA\mathbf{x}'\rangle \pm O(\epsilon)] \geq 1 - 2^{-\Theta(d)}$.

Now, we can choose a $\frac{1}{2}$-net $M = \{A\mathbf{x} \mid \mathbf{x} \in N\}$, which has size 5^d.
By the union bound, for all pairs $\mathbf{y}, \mathbf{y}' \in M$, $\langle \mathbf{y}, \mathbf{y}'\rangle = \langle \mathbf{Sy}, \mathbf{Sy}'\rangle \pm O(\epsilon)$. We have 25^d such pairs, which is finite. By linearity, if this property holds for $\mathbf{y}, \mathbf{y}' \in M$, then for $\alpha\mathbf{y}, \beta\mathbf{y}'$, we have $\langle \alpha\mathbf{y}, \beta\mathbf{y}'\rangle = \alpha\beta\langle \mathbf{Sy}, \mathbf{Sy}'\rangle \pm O(\epsilon\alpha\beta)$.

Consider an arbitrary unit vector $\mathbf{x} \in S^{d-1}$, and let $\mathbf{y} = A\mathbf{x}$. We can pick $\mathbf{y_1} \in M$ such that $|\mathbf{y} - \mathbf{y_1}|_2 \leq \gamma$.
Now, let α be such that $|\alpha(\mathbf{y} - \mathbf{y_1})|_2 = 1$, i.e. we scale the difference vector $\mathbf{y} - \mathbf{y_1}$ to become a unit vector. Note that $\alpha \geq 1/\gamma$, which could be infinite. If so, we stop. Otherwise, we continue by picking another $\mathbf{y_2'} \in M$ such that $|\alpha(\mathbf{y} - \mathbf{y_1}) - \mathbf{y_2'}|_2 \leq \gamma$. Then $|\mathbf{y} - \mathbf{y_1} - \frac{\mathbf{y_2'}}{\alpha}| \leq \frac{\gamma}{\alpha} \leq \gamma^2$.
We can set $\mathbf{y_2} = \frac{\mathbf{y_2'}}{\alpha}$, and repeat this process to obtain $\mathbf{y_1}, \mathbf{y_2}, \mathbf{y_3}$, and so on, such that for all i, $|\mathbf{y} - \mathbf{y_1} - \mathbf{y_2} - ... - \mathbf{y_i}|_2 \leq \gamma^i$.

We will finish this net argument next lecture.