CS 15-859: Algorithms for Big Data Fall 2019

Lecture 2 — Sep 5

Prof. David Woodruff *Scribe: Hang Liao*

1 Sketching to Solve Least Squares Regression

Although A^-b gives the solution to the least squares regression problem $Ax = b$, it is still not satisfying because calculate $A^- = V\Sigma^{-1}U^T$ can naively take $O(nd^2)$ time and use fast matrix multiplication it still takes $O(nd^{1.376})$ time. However, with sketching and randomization, we can find a multiplicative approximation algorithm that outputs x' for which $|Ax' - b|_2 = (1 \pm \epsilon) \min_x |Ax - b|_2$ with high probability in a much shorter time.

Claim 1. The following procedure produces such an x':

- Draw S from a **certain** $k \times n$ random family of matrices, for a value $k \ll n$
- Compute $S \times A$ and $S \times b$
- Output the solution $x' = (SA)^- Sb$ to $\min_x |(SA)x - (Sb)|_2$

Remark 1. Notice that not every distribution of random matrices family can work as S. Also note normally calculating $S \times A$ still require $O(knd)$ work. However, for some distribution of S, SA can be computed in $O(nd \log n)$ time, which will be covered in the next lecture.

Definition. (subspace embedding) Let A be a $n \times d$ matrix. S is a subspace embedding for the column space of A if with high probability $\forall x \in \mathbb{R}^d$, $|SAx|_2 = (1 \pm \epsilon)|Ax|_2$.

Lemma 1. *Given two independent random variables X, Y, with $X \sim N(0, a^2)$ and $Y \sim N(0, b^2)$, $Z = X + Y$ is a variable drawn from $N(0, a^2 + b^2)$.*

Proof. Note the probability density function f_Z of Z is the convolution of probability density functions f_X and f_Y.

$$
\begin{aligned}
f_Z(z) &= \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) \mathrm{d}y \\
&= \int_{-\infty}^{\infty} \frac{1}{a(2\pi)^{1/2}} e^{-\frac{(z-y)^2}{2a^2}} \frac{1}{b(2\pi)^{1/2}} e^{-\frac{y^2}{2b^2}} \mathrm{d}y \\
&= \frac{1}{(a^2+b^2)^{1/2}(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{(a^2+b^2)^{1/2}}{(2\pi)^{1/2}ab} e^{-\frac{(z-y)^2}{2a^2} - \frac{y^2}{2b^2}} \mathrm{d}y \\
&= \frac{1}{(a^2+b^2)^{1/2}(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{(a^2+b^2)^{1/2}}{(2\pi)^{1/2}ab} e^{-\frac{z^2}{2(a^2+b^2)} - \frac{(y-\frac{b^2 z}{a^2+b^2})^2}{2(\frac{(ab)^2}{a^2+b^2})}} \mathrm{d}y \\
&= \frac{1}{(a^2+b^2)^{1/2}(2\pi)^{1/2}} e^{-\frac{z^2}{2(a^2+b^2)}} \int_{-\infty}^{\infty} \frac{(a^2+b^2)^{1/2}}{(2\pi)^{1/2}ab} e^{-\frac{(y-\frac{b^2 z}{a^2+b^2})^2}{2(\frac{(ab)^2}{a^2+b^2})}} \mathrm{d}y
\end{aligned}
$$

Note the last integral is the integral of the probability density function of a Gaussian variable in $N(\frac{b^2 z}{a^2+b^2}, \frac{(ab)^2}{a^2+b^2})$, which equals to 1. So

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

and z is a random variable drawn from $N(0, a^2+b^2)$. ■

Lemma 2. *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 i.i.d. $N(0, 1/k)$ random variables.*

Proof. Since u, v are orthogonal, there exists a rotation matrix R_0 that sends u to αe_1 and v to βe_2, where α, β are constants and e_1, e_2 are the standard bases with their first and second coordinates equals to 1 respectively. Note for any fixed $n \times n$ matrix R and n-dimensional vector of i.i.d. $N(0, 1)$ random variables g, the probability density function $f(x)$ of Rg is

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

Since R_0 is a rotation matrix, $R_0 R_0^T = I$,

$$f_{R_0 g}(x) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{x^T x}{2}} = f_g(x)$$

Therefore, for rotation matrix R_0, the distribution of $R_0 g$ and of g are the same. Now

$$\begin{aligned}
\langle g, u\rangle &= g^T u && (1)\\
&= g^T R_0^T R_0 u && (2)\\
&= \langle R_0 g, R_0 u\rangle && (3)\\
&= \langle h, \alpha e_1\rangle && (4)\\
&= \alpha h_1 && (5)
\end{aligned}$$

and

$$\begin{aligned}
\langle g, v\rangle &= \langle R_0 g, R_0 v\rangle && (6)\\
&= \langle h, \beta e_2\rangle && (7)\\
&= \beta h_2 && (8)
\end{aligned}$$

where h is a vector of i.i.d. $N(0, 1/k)$ random variables and h_i is the i-th entry of h.

By definition, $\langle g, u\rangle$ and $\langle g, v\rangle$ are independent. ■

Lemma 3. *Suppose S is a $k \times n$ matrix of i.i.d. Normal random variables. Then with high probability, $\forall x \in \mathbb{R}^d$, $|SAx|_2 = (1 \pm \epsilon)|Ax|_2$ or equivalently, S is a subspace embedding for the column space of A.*

Proof. We make two assumptions:

1. We can assume the columns of A are orthonormal. Since $A = U\Sigma V^T$ (SVD), we achieve this by replacing A with U and replacing x with $\Sigma V^T x$.

2. We can also assume x is a unit vector by dividing both sides by the norm of x.

Let S_{i*} be the i-th row of S, A_j be the j-th column of A. The i-th row of SA is $\langle\langle S_{i*}, A_1\rangle, \langle S_{i*}, A_2\rangle, ..., \langle S_{i*}, A_d\rangle$. For each $\langle S_{i*}, A_j\rangle$ term, as scaling a normal distribution by c will scale the variance by c^2, element in S_{i*} is in $N(0, 1/k)$, and A_j is a unit vector, by lemma 1, we conclude that $\langle S_{i*}, A_j\rangle \sim N(0, 1/k)$.

Since A_i is orthogonal to A_j for $i \neq j$, by lemma 2, elements in the same row are independent.

The above analysis holds for all rows of SA, so SA is a $k \times d$ matrix of i.i.d. $N(0, 1/k)$ random variables.

Consider a fixed unit vector $x_0 \in \mathbb{R}^d$. $|SAx_0|_2^2 = \sum_{i\in[k]}\langle g_i, x_0\rangle^2$ where g_i is the i-th row of SA. By lemma 1, $\langle g_i, x_0\rangle \sim N(0, \frac{1}{k})$. So $\langle g_i, x_0\rangle^2 \sim N(0, \frac{1}{k})^2$. Hence $\langle g_i, x_0\rangle \sim \frac{1}{k}\chi^2(1)$, $\mathbb{E}[\langle g_i, x_0\rangle] = \frac{1}{k}$ and $\mathbb{E}[|SAx_0|_2^2] = k\frac{1}{k} = 1$.

Theorem 1. *(Johnson-Lindenstrauss) Let G be the sum of k i.i.d. $N(0, 1)$ random variables. We have*

$$\mathbf{Pr}\Big[G \geq k + 2(kx)^{1/2} + 2x\Big] \leq e^{-x}$$

$$\mathbf{Pr}\Big[G \leq k - 2(kx)^{1/2}\Big] \leq e^{-x}$$

With Johnson-Lindenstrauss Theorem, we can show $|SAx_0|_2^2$ is concentrated to its mean. Let $x = \frac{\varepsilon^2 k}{16}$. The two inequality becomes

$$\mathbf{Pr}\Big[G \geq k + \frac{\varepsilon k}{2} + \frac{\varepsilon^2 k}{8}\Big] \leq e^{-\frac{\varepsilon^2 k}{16}}$$

$$\mathbf{Pr}[G \leq k - \frac{\varepsilon k}{2}] \leq e^{-\frac{\varepsilon^2 k}{16}}$$

respectively. Therefore

$$\mathbf{Pr}[G = (1 \pm \varepsilon)k] \geq 1 - 2e^{\frac{\varepsilon^2 k}{16}}$$

If $k = \Theta(\varepsilon^{-2}\log\frac{1}{\delta})$, we have

$$\mathbf{Pr}[G = (1 \pm \varepsilon)k] \geq 1 - \delta$$

Note G is a sum of i.i.d. $N(0, 1)$ variables whereas $|SAx_0|$ is a sum of i.i.d. $N(0, \frac{1}{k})$ variables. Therefore,

$$\mathbf{Pr}\Big[|SAx_0|_2^2 = (1 \pm \varepsilon)\Big] \geq 1 - \delta$$

Let $\Theta(\log(\frac{1}{\delta})) = d$, we have $\delta = 2^{-\Theta(d)}$. So

$$\mathbf{Pr}\Big[|SAx_0|_2^2 = (1 \pm \varepsilon)\Big] \geq 1 - 2^{-\Theta(d)}$$

with $k = \Theta(\frac{d}{\varepsilon^2})$.

To argue $|SAx_0|_2^2 = (1 \pm \varepsilon)$ holds for all x, we introduce the definition of net. Consider the d-dimensional unit sphere S^{d-1}.

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

We use greedy construction to build the net. First we pick a random point on the sphere. If there is a point on the sphere that has distance larger than γ from any other points, we include the point in N. Keep this process until we can't find such point.

Now we argue that the algorithm terminates. Imagine there is a $\frac{\gamma}{2}$ ball around all the points on our net. Then every ball is disjoint because by our greedy algorithm we wouldn't pick a point if it has distance within γ to another point. Also notice that the union of the unit sphere and imaginary balls are all contained in a larger ball with radius $1+\frac{\gamma}{2}$. The total number of balls we can pick is at most $\frac{(1+\frac{\gamma}{2})^d}{(\frac{\gamma}{2})^2}$. So $|N| \leq \frac{(1+\frac{\gamma}{2})^d}{(\frac{\gamma}{2})^2}$.

This actually translates to subspace. Let $M = \{Ax | x \in N\}$. Note for every $x \in N$, we can find $x' \in N$ such that $|x - x'| \leq \gamma$ and so $y = Ax'$ such that $|Ax - y| = |Ax - Ax'| = |x - x'| \leq \gamma$ since the column of A are orthonormal.

For a fixed pair of unit x, x', $|SAx|_2^2, |SAx'|_2^2$ and $|SA(x-x')|_2^2$ are preserved (compared with $|Ax|_2^2, |Ax'|_2^2$ and $|A(x-x')|_2^2$) up to a $1 \pm \varepsilon$ factor with probability $1 - 2^{-\Theta(d)}$. Since

$$|SA(x-x')|_2^2 = |SAx|_2^2 + |SAx'|_2^2 - 2\langle SAx, SAx' \rangle$$

we know $\mathbf{Pr}[\langle Ax, Ax' \rangle = \langle SAx, SAx' \rangle \pm O(\varepsilon)]$ with probability $1 - 2^{-\Theta(d)}$. Now let $\gamma = \frac{1}{2}$, and so $|M| \leq |N| = \frac{(1+\frac{\gamma}{2})^d}{(\frac{\gamma}{2})^d} = 5^d$. We can condition on the event for all $y, y' \in M$, $\langle y, y \rangle = \langle Sy, Sy' \rangle \pm O(\varepsilon)$ and union bound the error probability.

Consider an arbitrary $x \in S^{d-1}$ and $y = Ax$, choose an y_1 such that $|y - y_1|_2 < \gamma$. Let α be a constant such that $|\alpha(y - y_1)| = 1$ and so $\alpha \geq \frac{1}{\gamma}$. Note $\alpha(y - y_1)$ is still in the column space of A. Let $y_2' \in M$ such that $|\alpha(y - y_1) - y_2'|_2 \leq \gamma$. Then $|y - y_1 - \frac{y_2'}{\alpha}|_2 \leq \frac{\gamma}{\alpha} \leq \gamma^2$. Set $y_2 = \frac{y_2'}{\alpha}$ and repeat, we obtain y_i that for all i, $|y - y_1 - ... - y_i|_2 \leq \gamma^i$. By triangle inequality,

$$\begin{aligned}
|y_i|_2 &= |(y - y_1 - ... - y_{i-1}) - (y - y_1 - ... - y_{i-1} - y_i)|_2 \\
&\leq \gamma^{i-1} + \gamma^i \\
&\leq 2\gamma^{i-1}
\end{aligned}$$

Now suppose we have $y = \sum_i y_i$,

$$\begin{aligned}
|Sy|_2^2 &= |S\sum_i y_i|_2^2 \\
&= \sum_i |Sy_i|_2^2 + 2\sum_{i,j,i\neq j} \langle Sy_i, Sy_j \rangle \\
&= \sum_i |y_i|_2^2 + 2\sum_{i,j,i\neq j} \langle y_i, y_j \rangle + O(\varepsilon)\sum_{i,j} |y_i|_2 |y_j|_2 \\
&= \sum_i |y_i|_2^2 + 2\sum_{i,j,i\neq j} \langle y_i, y_j \rangle \pm O(\varepsilon) \\
&= |\sum_i y_i|_2^2 \pm O(\varepsilon) \\
&= 1 \pm O(\varepsilon)
\end{aligned}$$

Since this held for an arbitrary $y = Ax$ for unit x, by linearity it follows that $\forall x$, $|SAx|_2 = (1 \pm \varepsilon)|Ax|_2$. ■

Theorem 2. *Suppose S is a $d/\epsilon^2 \times n$ matrix of i.i.d. Normal random variables. Then with high probability the solution to* $\operatorname{argmin}_x |(SA)x - (Sb)|_2$ *gives us* $|Ax' - b|_2 = (1 \pm \epsilon) \min |Ax - b|_2$.

Proof. By lemma 3, for any matrix with size $n \times d$, we can find S such that for any vector $x \in \mathbb{R}^d$, $|SAx|_2 = (1 \pm \varepsilon)|Ax|_2$.

Now consider subspace L spanned by columns of A together with b. For any $y = Ax - b \in L$, $|Sy|_2 = (1 \pm \varepsilon)|y|_2$. Therefore, for all x, $|S(Ax - b)|_2 = (1 \pm \varepsilon)|Ax - b|_2$. Note the solution x' to $\operatorname{argmin}_x |(SA)x - (Sb)|_2$ satisfies $|Ax' - b| = (1 \pm \varepsilon) \min |Ax - b|_2$. Given SA and Sb, we can calculate $\operatorname{argmin}_x |(SA)x - (Sb)|_2$ in $poly(\frac{d}{\varepsilon})$ time.

■