CS 15-859: Algorithms for Big Data Fall 2022

Lecture 1 — 09/08/2022

Prof. David Woodruff *Scribe: Alper Çakan*

1 Part 2

Definition (Singular Value Decomposition)**.** Let $A_{n\times d}$ be any matrix. Then, there is $U_{n\times d}, \Sigma_{d\times d}, V_{d\times d}$ such that

- $A = U\Sigma V^T$,
- Columns of U are orthonormal, that is, $U^T U = I$,
- Σ is a non-negative diagonal matrix[1]
- Columns of V are orthonormal

Using singular value decomposition, we define a *psuedoinverse*.

Definition (Moore-Penrose psuedoinverse)**.** Let $A = U\Sigma V^T$ be a SVD. Define the Moore-Pensore pseudoinverse A^- of A as

$$A^- = V\Sigma^- U^T$$

where

$$(\Sigma^-)_{ii} = \begin{cases} \frac{1}{\Sigma_{ii}} & \text{, if } \Sigma_{ii} \neq 0, \\ 0 & \text{, otherwise} \end{cases}$$

It is easy to show that AA^- is projector onto column space of A.

Remember that when columns of A are not linearly independent, then there is no unique solution to $\min_x |Ax - b|_2^2$. However, $x = A^- b$ is a solution with a useful property.

Proposition 1. Let $x^* = A^- b$. Then,

- x^* is an optimal solution. That is, for any x, we have $|Ax - b|_2^2 \geq |Ax^* - b|_2^2$.
- x^* has minimum norm. That is, for any x' with $|Ax^{'} - b|_2^2 = |Ax^* - b|_2^2$, we have $||x^*|| \leq ||x^{'}||$.

Remark 1. We can indeed solve least squares regression via normal equations and the SVD method above. However, computing SVD naively takes $O(nd^2)$ time. Even with best known algorithms it takes $O(nd^{\sim 1.3})$ time. However, in the next section we will have much faster algorithms by allowing approximate solutions.

[1] In fact, we can even pick a Σ such that $\Sigma_{1,1} \geq \Sigma_{2,2} \geq \cdots \geq \Sigma_{d,d}$

1.1 Skecth-and-Solve

High level: Instead of solving $\min_x |Ax - b|$, we will solve $\min_x |S_{k\times n}Ax - Sb|$ where $k \ll n$. Of course, we need S to satisfy some properties. For example, the multiplication SA should be easy or there is no point in the reduction. Also, to have some kind of approximation guarantee, there must be some *relation* between $\min_x |Ax - b|$ and $\min_x |S_{k\times n}Ax - Sb|$.

Candidate: $\frac{d}{\varepsilon^2} \times n$ matrix of iid normal variables.

Theorem 1 (Subspace embedding)**.**

$$k = O(\frac{d^2}{\varepsilon^2})$$
$$S : \frac{d}{\varepsilon^2} \times n \textit{ matrix of iid } N(0, \frac{1}{k}) \textit{ normal variables}$$

Then, for any fixed d-dimensional subspace, i.e., column space of some $A_{n\times d}$, we have that with high probability,

$$\forall x \in \mathbb{R}^d \quad ||SAx||_2 = (1 \pm \varepsilon)||Ax||_2$$

Without loss of generality[2] we can assume that A has orthonormal columns. We can also assume that x is a unit vector since we can scale each side by $||x||_2$ otherwise.

First, we will prove a statement about distribution of SA.

Claim 1. Let S be defined as in Theorem 1. Then, for any A, entries of SA are iid with distribution $N(0, \frac{1}{k})$

Proof. To prove this, we will prove two simpler claims.

Claim 2. For independent $X \sim N(0, a^2)$ and $Y \sim N(0, b^2)$, we have $(X + Y) \sim N(0, a^2 + b^2)$.

Proof. Probability density function of $Z = X + Y$ is convolution of pdfs of X and Y.

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

since the integral in the last step is the pdf of a Gaussian integrated over the whole real line. ■

Claim 3.
- u, v vectors with $\langle u, v\rangle = 0$

[2] By picking an A' whose columns are an orthonormal basis of $Col(A)$

- g vector with iid normally distributed entries

Then, $\langle g, u\rangle$, $\langle g, v\rangle$ are independent Gaussians.

Proof. By above, it is easy to see that $\langle g, u\rangle$ and $\langle g, v\rangle$ are Gaussians. Now, we need to prove that they are independent. First, observe that rotating a Gaussian vector preserves its distribution. More precisely,

Lemma 1 (Rotational invariance)**.** *If R is a fixed matrix and g is an n dimensional vector of iid $N(0, 1)$ random variables, then pdf of Rg is*

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

In particular, when R is a rotation matrix, $RR^T = I$ and hence $Rg \sim g$

Now, let's pick a rotation takes u to αe_1 and v to βe_2. We can do this since $\langle u, v\rangle = 0$. Since rotations preserves inner products, we have $\langle g, u\rangle = \langle Rg, Ru\rangle = \alpha h_1$ and similarly $\langle g, v\rangle = \beta h_2$ where $h = Rg$. Since g and hence h has iid entries, h_1, h_2 are independent Gaussians. ■

Then, observe that each entry of SA is a dot product of a row of S and a column of A. Since A has orthonormal columns, and since each row of S is independent, we get that entries of SA are iid with distribution $N(0, \frac{1}{k})$ ■

Now, we move to the proof our theorem.

Proof. Consider any fixed vector x for now, and we will at a later stage use union bound in combination with another technique.

Then, we have $|SAx|_2^2 = \sum_{i\in[k]} \langle g_i, x\rangle^2$ where g_i is the i^{th} row of SA. Each $\langle g_i, x\rangle$ is distributed as $N(0, \frac{1}{k})$. Therefore, $\mathbb{E}[\langle g_i, x\rangle^2] = \frac{1}{k}$ and hence $\mathbb{E}[|SAx|_2^2] = 1$[3]. While on expectation we have what we want, we had a stronger claim that our good event happens with high probability. So, we need analyze how concentrated $|Ax|_2^2 = 1$ is around its expectation.

Theorem 2 (Johnson-Lindenstrauss)**.** *$h_1, \ldots, h_k$: iid $N(0, 1)$ random variables*

Then, $G = \sum_{i=1}^k h_i^2$ is a χ^2 random variable.

When we apply known tail bounds to G,

$$\mathbb{P}[G \geq k + 2\sqrt{kx} + 2x] \leq e^{-x}$$
$$\mathbb{P}[G \leq k - 2\sqrt{kx}] \leq e^{-x}$$

By plugging in $x = \frac{\varepsilon^2 k}{16}$ above, we get

$$\mathbb{P}[G \in k(1 \pm \varepsilon)] \geq 1 - 2e^{-\varepsilon^2 k/16}$$

[3] Remember that we have $|Ax|_2^2 = 1$

By choosing $k = \Theta(\varepsilon^{-2}\log(\frac{1}{\delta}))$, we get

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

While this is *almost* what we wanted, remember that we showed this for any arbitrary x, while we actually wanted to show that this holds for all x at the same time. To achieve this, one might consider using a union bound argument. However, there are infinitely many points x. But, we can still make this work by using a union bound type argument in a smart way.

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

We can construct a γ-net N by keeping greedily choosing a point that is not yet covered. It is easy to see that this yields a net N with $|N| \leq \left(\frac{1+\gamma/2}{\gamma/2}\right)^d$. In fact, we can construct a net for our subspace Ax by first constructing N and then $M = \{Ax : x \in N\}$ is a net for $\{Ax\}$. To see this, observe that for every $x \in S^{d-1}$, there is a y in M for which $|Ax - y|_2 \leq \gamma$. To see the second part, let x' in S^{d-1} be such that $|x - x'|_2 \leq \gamma$. Then, since A is orthonormal, we have $|Ax - Ax'| = |x - x'| \leq \gamma$. Set $y = Ax'$

Now, using nets, we will finish our proof. Take a fixed pair of (unit) x, x'. Then, $|SAx|^2$, $|SAx'|^2$, $|SAx - SAx'|^2$ are preserved up to $(1 \pm \varepsilon)$ factor with probability $1 - 2^{-\Theta(d)}$. Observe that

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

Therefore, by subtracting each side,

$$\mathbb{P}[\langle Ax, Ax'\rangle = \langle SAx, SAx'\rangle \pm O(\varepsilon)] = 1 - 2^{-\Theta(d)}$$

Basically, S also preserves inner products (between any fixed pair) with high probability. Choose a $\frac{1}{2}$-net $M = \{Ax : x \in N\}$ of size 5^d. By a union bound, for all pairs $y, y' \in M$[4]

$$\langle y, y'\rangle = \langle Sy, Sy'\rangle \pm O(\varepsilon)$$

Now, condition on this event. By linearity, this also holds for $\alpha y, \beta y'$ with error $\alpha\beta\varepsilon$. Finally, take any $x \in S^{d-1}$ and consider $y = Ax$. Let $y_1 \in M$ be such that $|y - y_1|_2 \leq \gamma$. If this error norm is zero, we can stop. If not, let α be such that $|\alpha(y - y_1)|_2 = 1$. Now $\alpha(y - y_1)$ is a unit vector in the column space of A. Then we can repeat and approximate this difference vector with a net vector. Let y_2' be 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 $y_2 = \frac{y_2'}{\alpha}$. Repeat, obtaining $y_1, y_2, \ldots$ such that for all integers i we have

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

But by using the same inequality for the previous step $i - 1$ and triangle inequality, this implies $|y_i|_2 \leq \gamma^{i-1} + \gamma^i \leq 2\gamma^{i-1}$

[4] In particular, lengths of y are also preserver by picking $y' = y$

Now, we have $y_1, y_2, \dots$ with $|y_i|_2 \leq \gamma^{i-1} + \gamma^i \leq 2\gamma^{i-1}$. Then,

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

■