Advanced Linear Algebra Week 10 Day 1

2018/11/19 – Jonathan Hayase, updated by Prof. Weiqing Gu

1 Application of Advanced Linear Algebra to Big Data

Consider the following:

$$\underbrace{L}_{\text{Linear}} \longleftrightarrow \underbrace{A}_{P^{-1}AP}$$

We are interested in invariants. Some examples of invariants are: $\det\left(P^{-1}AP\right) = \det A$, $\operatorname{tr}\left(P^{-1}AP\right) = \operatorname{tr} A$, and $\operatorname{rank}\left(P^{-1}AP\right) = A$. If these are the things that we care about then we have the freedom to choose P to fit our needs, and the quantities we are interested in do not change. This is a general technique that we would like to explore.

Suppose A (say, a data design matrix) is very big and can't fit into memory. Consider $\Omega : \mathcal{M}_{n\times n}(\mathbb{K}) \to \mathrm{BlockD}^n(\mathbb{K})$

$$A = \begin{bmatrix} & \end{bmatrix} \mapsto \begin{bmatrix} \boxed{h_1 \times h_1} & & 0 \\ & \ddots & \\ 0 & & \boxed{h_k \times h_k} \end{bmatrix} \leftarrow B$$

$$A\boldsymbol{x} = \boldsymbol{b} \qquad A_i\boldsymbol{x}_i = \boldsymbol{b}_i \qquad i = 1, 2, \ldots, k$$

$$B = P^T AP$$

We wish to find $P \in \operatorname{perm}(n)$ which minimizes $\|PAP^* - B_D\|$ with large probability.

1.1 Norms of Matrices

$$A \to \text{view this as an operator}$$

We can talk about the norm that is induced by the vector norm like so:

$$\|A\| \triangleq \sup_{\|\boldsymbol{x}\|=1} \|A\boldsymbol{x}\| = \sup_{\boldsymbol{x}\in\mathbb{R}^n} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$$

However, there are other norms we might consider, for example:

$$\|A\| = \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{\operatorname{tr}\left(A^T A\right)} = \|A\|_F$$

1.2 Johnson-Lindenstrauss lemma

Suppose we have $\|A\boldsymbol{x}\| = \|\boldsymbol{x}\|$, then A is orthogonal. Then the mapping from JL, if viewed as a matrix, could be thought of as "almost orthogonal".

1.3 Sketching method as a tool for Linear Algebra

1.3.1 Problem 1

Approximating Leverage scores: A d dimensional subspace W contained in $\mathbb{R}^n$ can be expressed written as $\left\{x \mid \exists y \in \mathbb{R}^d, x = Uy\right\}$ for some $U \in \mathbb{R}^{n\times d}$ with orthonormal columns. The squared Euclidean norms of rows of U are unique up to permutation. i.e. they depend only on A and are known as the Leverage scores of A.

Given A, we would like to output a list of its leverage scores up to $1 \pm \epsilon$.

e.g. $A = \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_n \end{bmatrix}$. Consider the column space of A. $\implies$ can find an orthonormal basis of the columns space of A.

1.3.2 Problem 2

Least squares regression:

Given $A \in \mathbb{R}^{n\times d}$ and $\boldsymbol{b} \in \mathbb{R}^n$ want to compute $\|A\tilde{\boldsymbol{x}} - \boldsymbol{b}\| \leq (1+\epsilon)\min_{x\in\mathbb{R}^d}\|A\boldsymbol{x} - b\|$.

1.3.3 Problem 3

Given $A \in \mathbb{R}^{n\times d}$ and integer $k > 0$. Compute $\tilde{A}_k \in \mathbb{R}^{n\times d}$ with $\operatorname{rank}\left(\tilde{A}\right) \leq k$ so that $\left\|A - \tilde{A}_k\right\|_F \leq (1+\epsilon)\min_{\operatorname{rank}(Ak)<k}\|A - A_k\|$.

Today, we will focus on Problem 2.

1.4 More on Least Squares Regression

Q: How to find an approximate solution $\boldsymbol{x}$ to $\min_x\|A\boldsymbol{x} - \boldsymbol{b}\|$. Goal: output $\tilde{\boldsymbol{x}}$ for which $\|A\tilde{\boldsymbol{x}} - b\|_2 \leq (1+\epsilon)\min\|A\boldsymbol{x} - \boldsymbol{b}\|$.

Idea: Draw S from a $k \times n$ random family of matrices for value $k \ll n$: Compute SA and Sb and output the solution $\tilde{\boldsymbol{x}}$ to $\min\left\|(SA)\boldsymbol{x} - S\boldsymbol{b}\right\|$. e.g. S is $d^2/\epsilon \times n$ matrix of i.i.d. normal random variables. E.g.

$$S = \begin{bmatrix} \pm e_{i_1} & \pm e_{i_2} & \dots & \pm e_{i_k} \end{bmatrix}$$

$$\begin{bmatrix} \boldsymbol{e}_2 & -\boldsymbol{e}_1 & \boldsymbol{e}_1 & \dots \end{bmatrix} = \begin{bmatrix} 0 & -1 & 1 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} RA_1 \\ \vdots \\ RA_n \end{bmatrix}$$

Original approach/analysis was very long and complicated, then Nelson and Nugyen[1] used Advanced Linear Algebra + Probability to give a much simpler proof.

Key idea: Consider $\left[A \mid b\right] \triangleq B$ and consider the column space of B. Let U be a matrix with columns which form an orthonormal basis of the column space of B. Claim: It sufficies to show $\|\Pi U\boldsymbol{x}\|_2 = (1\pm\epsilon)\|\boldsymbol{x}\|_2$. This will imply $\left\|S(A\boldsymbol{x} - \boldsymbol{b})\right\|_2 = (1\pm\epsilon)\|A\boldsymbol{x} - \boldsymbol{b}\|_2$ for all $\boldsymbol{x}$.

[1] is this right...

1.5 Basics of Least Squares

$$\overbrace{\begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(i)} \\ \vdots \\ y^{(N)} \end{bmatrix} = \underbrace{\begin{bmatrix} \vdots & & & \vdots \\ 1 & x_1^{(i)} & \cdots & x_n^{(i)} \\ \vdots & & & \vdots \end{bmatrix}}_{X} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}}^{y^{(i)}=\theta_0+\theta_1 x_1^{(i)}+\cdots+\theta_n x_n^{(n)}}$$

or

$$\boldsymbol{y} = X\boldsymbol{\theta} = \underbrace{\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m}_{\text{column space of } X}$$

where we wish to find $\boldsymbol{\theta}$. Then we may write

$$X^T\boldsymbol{y} = X^T X\boldsymbol{\theta} \implies \boldsymbol{\theta} = (X^T X)^{-1}(X^T\boldsymbol{y})$$

1.6 An aside on the product of a matrix and its adjoint

$$\boldsymbol{x}^* A^* A\boldsymbol{x} = (A\boldsymbol{x})^*(Ax) = \|Ax\|^2 \geq 0 \implies \lambda_i \geq 0$$

so

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^*A)} = \sigma_{\max}(A)$$

then

$$P^{-1}AP = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$$

$$\operatorname{tr}(P;-1)\, A^*AP = \lambda_1 + \cdots + \lambda_n$$

and

$$\|A\|_F = \sqrt{\operatorname{tr}(A^*APP^{-1})} = \sqrt{\lambda_1 + \cdots + \lambda_n} = \sqrt{\sum \sigma_i^2} = \left(\sum \sigma_i\right)^{1/2}$$

1.7 Why orthonormal bases

Suppose we have an orthonormal basis $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$. Then any $\boldsymbol{x}$ can be written as $\boldsymbol{x} = x_1\boldsymbol{u}_1 + \cdots + x_n\boldsymbol{u}_n$ and $A\boldsymbol{x} = x_1 A\boldsymbol{u}_1 + \cdots + x_n A\boldsymbol{u}_n = x_1\lambda_1\boldsymbol{u}_1 + \cdots + x_n\lambda_n\boldsymbol{u}_n$.

$$\|Ax\| = \sqrt{(\lambda_1 x_1)^2 + \cdots + (\lambda_n x_n)^2} \geq \sqrt{(\lambda_n x_1)^2 + \cdots + (\lambda_n x_n)^2} = \sqrt{\lambda_n^2(x_1^2 + \cdots + x_n^2)}$$

$$= \sqrt{\lambda_n^2\|x\|} = \sqrt{\lambda_n^2}$$

If A is hermitian then $\|Ax\| \geq \lambda_{\min}$. Similarly,

$$\|Ax\| \leq (\lambda_1 x_1)^2 + \cdots + (\lambda_n x_n)^2 = \lambda_{\max} = (1 \pm \epsilon)\|Ax - b\|_2$$

for all x.

1.8 Picking up from before...

Claim: SU is $\frac{(d+1)^2}{\epsilon^2} \times (d+1)$ matrix.

Claim: $\left\|(SU)^T SU - I\right\|_2 \leq \left\|U^T S^T SU - I\right\|_F \leq \epsilon$.

Definition: An oblivious subspace embedding (OSE) is a distribution D over matrices $\Pi \in \mathbb{R}^{n \times m}$ given some parameters ϵ, δ such that for any linear subspace $W \subseteq \mathbb{R}^n$ with $\dim W = d$, the following holds:

$$\mathcal{P}_{\Pi \sim D}\left(\forall x \in W, \|\Pi x\|_2 \in (1 \pm \epsilon)\|x_2\|\right) > \frac{2}{3}$$

N N showed that an OSE exists with $m = O(d^2/\epsilon^2)$ and where $\Pi \in \text{supper}(\mathcal{O})$ has exactly $s = 1$ nonzero entries per column. (This improves Woodruff's result.)

Goal: Obtain a fast randomized Algorithm for several numerical linear algebra problems. We focus on Least Squares problem $\text{argmin}_{x \in \mathbb{R}^d} \|\boldsymbol{Ax} - \boldsymbol{b}\|$.

Key idea: Use sketch as application to ℓ_2-estimation in data streams only require h to be pairwise independent and σ 4-wise independent. Claim: A matrix Π preserving the Euclidean norm of all vectors $x \in W$ up to $1 \pm \epsilon$ is equivalent to

$$\Pi U \boldsymbol{y} = (\pm \epsilon)\|\boldsymbol{y}\|$$

simultaneously for $y \in \mathbb{R}^d$ if and only if all singular values lie in the interval $[1-\epsilon, 1+\epsilon]$.

Claim/Recall: Eigenvalues of orthonormal matrices have norm 1. E.g.

$$A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

has no eigenvalues geometrically, but with complex eigenvalues $\lambda_i, |\lambda_i| = 1$.

Proof: Let $\lambda, \boldsymbol{x}$ be an eigenvalue/vector pair, so $A\boldsymbol{x} = \lambda \boldsymbol{x}$.

$$\begin{aligned}
(Ax)^*(Ax) &= (\overline{Ax})^T(Ax) \\
&= \overline{x}^T \overline{A}^T Ax \\
&= \overline{x}^t x = \|x\|^2
\end{aligned}$$

But

$$(Ax)^*(Ax) = \|Ax\|^2 = \|\lambda x\|^2 = |\lambda|^2 \|x\|^2$$

so $|\lambda| = 1$.

Write $S = (\Pi U)^*(\Pi U)$ (note, S here is not the same as before) so that we want to show that all of the eigenvalues of S lie in $\left[(1-\epsilon)^2, (1+\epsilon)^2\right]$.

Trick: $S = I + (S - I)$. Use Weyl's inequality (described in the next section).

Let $M = S$ with eigenvalues μ_i, $H = I$ with eigenvalues 1, and $P = S - I$ with eigenvalues ρ_i.

$$-\|S - I\|_2 \leq \rho_n \leq \mu_i - 1 \leq \rho_1 \leq \|S - I\|_2$$

Because $S = I + (S - I)$, we can show all eigenvalues of S are $1 \pm \|S - I\|$. And we want to bound $\|S - I\|$. We ultimately are showing that S is the JL transformation.

1.8.1 Weyl's inequality

Let M, H, P be $n \times n$ hermitian matrices with only real eigenvalues. where M has eigenvalues $\mu_1 \geq \cdots \geq \mu_n$, H had eigenvalues $\gamma_1 \geq \cdots \geq \gamma_n$, and P has eigenvalues $\rho_1 \geq \cdots \geq \rho_n$. Then for $1 \leq i \leq n$, it holds that $\rho_n \leq \mu_i - \gamma_i \leq \rho_1$. Proof [Tao 12].

1.9 Continuing from before

We want to show that $\|S - I\|$ is small with good probability by Markov's inequality.

$$\mathbb{P}\left(\|S_I\| \geq t\right) = \mathbb{P}\left(\|S - I\|^2 \geq t^2\right) \leq \frac{\|S - I\|^2}{t^2} \leq \frac{\|S - I\|_F}{t^2}$$

and then we proceed to bound.