Recitation #1: Linear Algebra & Yao's Min-Max Theorem

TA: Rajesh Jayaram

September 6, 2019

Vectors

- A set of vectors $v_1, \dots, v_k \in \mathbb{R}^n$ are *linearly independent* if for all $\alpha_1, \dots, \alpha_k \in \mathbb{R}$,

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_k v_k = 0$$

  implies that $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$.

- A subspace S of $\mathbb{R}^n$ is a set of vectors such that:
  - If $y \in S$, then $\alpha y \in S$ for all $\alpha \in \mathbb{R}$.
  - If $x, y \in S$, then $x + y \in S$.

- The *inner-product* of vectors $v, u \in \mathbb{R}^n$ is given by $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$. If $\langle x, y \rangle = 0$, x and y are *orthogonal*: $x \perp y$.

Vectors

- The *Span* of $V = \{v_1, \dots, v_k\} \subset \mathbb{R}^n$ is a subspace:

$$\mathtt{Span}(V) = \{y \in \mathbb{R}^n \mid y = \sum_{i=1}^{k} \alpha_i v_i,\ \alpha_1, \dots, \alpha_k \in \mathbb{R}\}$$

- if $v_1, \dots, v_k$ are linearly independent, we say that the dimension of $\mathtt{Span}(V)$ is k, $\dim(\mathtt{Span}(V)) = k$,, and $v_1, \dots, v_k$ are a basis for $\mathtt{Span}(V)$.

- If $v_1, \dots, v_k$ are linearly independent and $\langle v_i, v_j \rangle = 1$ if $i = j$, otherwise $\langle v_i, v_j \rangle = 0$, then $v_1, \dots, v_k$ are a *orthonormal* basis for Span$(v_1, \dots, v_k)$.

Vectors

- The p-norm of a vector $v \in \mathbb{R}^n$ is given by $\|v\|_p = \left(\sum_{i=1}^n |v_i|^p\right)^{1/p}$.
  - Euclidean Norm: $\|v\|_2$
  - L_1 or Manhattan Norm: $\|v\|_1$.
- **Cauchy Schwartz Inequality:** for any $x, y \in \mathbb{R}^n$:

$$|\langle x, y\rangle|^2 \leq \|x\|_2^2 \|y\|_2^2$$

- **Pythagorean Theorem:** If $x_1, \ldots, x_k \in \mathbb{R}^n$ are pair-wise orthogonal, then

$$\left\| \sum_{i=1}^k x_i \right\|_2^2 = \sum_{i=1}^k \|x_i\|_2^2$$

Matrices

- A matrix $A \in \mathbb{R}^{d \times d}$ is *invertible* if there exists a $A^{-1} \in \mathbb{R}^{d \times d}$ with
$$A^{-1}A = AA^{-1} = I_d$$
where I_d is the $d \times d$ identity matrix.

- If $A \in \mathbb{R}^{d \times d}$, then A is invertible iff 1) the rows (or columns) of A are linearly independent.

- $A \in \mathbb{R}^{d \times d}$ is called orthogonal if both the columns and rows are orthonormal basis of $\mathbb{R}^d$.

- **Orthogonal matrices preserves 2-norm:** if $A \in \mathbb{R}^{n \times d}$, $d \leq n$ has orthonormal columns $A_1, \ldots, A_d \in \mathbb{R}^n$, then $\|Ax\|_2 = \|x\|_2$ for any $x \in \mathbb{R}^n$
  - Proof: use Pythagorean theorem:

$$\|Ax\|_2^2 = \|\sum_i A_i x_i\|_2^2 = \sum_i x_i^2 \|A_i\|_2^2 = \sum_i x_i^2 = \|x\|_2^2$$

Matrices

- Given $A \in \mathbb{R}^{n \times d}$ the column (row) rank of A is the size of the largest subset of linearly independent columns (rows).
- **Theorem:** Row rank $=$ column rank, so we call both the rank of A.
- **Rank Decomposition:** If A is rank k, we can always factor $A = U \cdot V$ where $U \in \mathbb{R}^{n \times k}$ and $V \in \mathbb{R}^{k \times d}$.

Eigenvalues and Singular Values

- If $Ax = \lambda x$, then λ is an *eigenvalue* of A, and x is an *eigenvector*.
- The rank of A is the number of non-zero eigenvalues.
- If A is symmetric, all eigenvalues are real (otherwise they may not be).
- The singular values of A are the square roots of the eigenvalues of $A^T A$.
- If $\sigma_{max}, \sigma_{min}$ is the largest (smallest) singular value of A, then for all x:

$$\sigma_{min}\|x\|_2 \leq \|Ax\|_2 \leq \sigma_{max}\|x\|_2$$

Singular Value Decomposition (SVD)

Given any matrix $A \in \mathbb{R}^{n \times d}$, (say $d \leq n$) we can write it as the product of three matrices $A = U\Sigma V$ where:

- $U \in \mathbb{R}^{n \times d}$ has Orthonormal columns:
  - $U^T U = I$ and $\|Ux\|_2 = \|x\|_2$.
- $V \in \mathbb{R}^{d \times d}$ has orthonormal rows:
  - $VV^T = I$ and $\|yV\|_2 = \|y\|_2$.
- $\Sigma \in \mathbb{R}^{d \times d}$ is a diagonal matrix with the diagonal entries $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_d$, where σ_i is the i-th singular value of A.

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d \end{bmatrix}$$

Singular Value Decomposition (SVD)

If $A \in \mathbb{R}^{n \times d}$, then the SVD of A can be computed in $\min\{nd^2, dn^2\}$ time.

- If $n >> d$, this can be expensive.
- In this course: (roughly) reduce $n \to$ poly(d), then apply SVD.

Yao's Minimax Principle

A tool to prove lower bounds for randomzied algorithms:

Theorem

Let $\mathcal{X}$ be a set on inputs to a problem, and $\mathcal{A}$ the set of all possible deterministic algorithms that solve the problem. For any $a \in \mathcal{A}$ and $x \in \mathcal{X}$, let $c(a, x)$ be the cost of algorithm a on input x. Let p, q be a distribution over $\mathcal{A}, \mathcal{X}$ respectively. Then

$$\max_{x \in \mathcal{X}} \left(\mathbb{E}_{A \sim_p \mathcal{A}} \left[c(A, x) \right] \right) \geq \min_{a \in \mathcal{A}} \left(\mathbb{E}_{X \sim_q \mathcal{X}} \left[c(a, X) \right] \right)$$

Informally: *the expected cost of a randomized algorithm on the worst-case input is at least the expected cost of the best deterministic algorithm for any fixed distribution over the inputs.*

Yao's Minimax Principle

How to use Yao's Minimax: Suppose you have a problem with possible inputs $\mathcal{X}$, and you want to prove a lower bound of B on the cost of any randomized algorithm solving the problem.

- Cost $c(a, x)$ can e.g. be the number of queries, runtime, space complexity.

Instead, we can choose a distribution q over the *inputs* $\mathcal{X}$, and show that any fixed deterministic algorithm pays cost at least B in expectation when solving a random $x \sim_q \mathcal{X}$ drawn according to q.

Yao's Minimax Principle: Example Problem

We now prove a $\Omega(\text{nnz}(A))$-time lower bound for any approximate regression algorithm.

Problem: For every n, d with $n \geq 2d$, and $n \leq m \leq nd/2$, prove that there is a set $\mathcal{X}$ of problem inputs (A, b) with the following properties. For each $(A, b) \in \mathcal{X}$, $\text{nnz}(A) = m$, $b \in \mathbb{R}^n$, and every randomized algorithm that, for each input $(A, b) \in \mathcal{X}$, outputs an x' with

$$\|Ax' - b\|_2 \leq 2 \min_x \|Ax - b\|_2$$

must read $\Omega(m)$ entries of A in expectation.

Yao's Minimax Principle: Example Problem

Solution: (assume for simplicity d divides m)

- **Intuition:** Hide a large entry in a random location inside A.
- Set $b = \mathbf{1} \in \mathbb{R}^n$, and put I_d in the first d rows of A_0
- For the next $(m-d)/d$ rows of A_0, set each entry equal to $1/d$, and all other entries equal to 0.
  - If $x = \vec{1}$ then $(A_0 x - b) = 0$.
- Now create a family $\mathcal{F} = \{(A^{i,j}, b) \mid i \in [(m-d)/d], j \in [d]\}$, where $A^{i,j} = A_0 + (10n - \frac{1}{d}) e_i e_j^T$.

Let $\mathcal{X} = (A_0, b) \cup \mathcal{F}$. Fix the distribution p which selects (A_0, b) /w prob $1/2$, otherwise selects a uniform instance from $\mathcal{F}$. Fix any deterministic algorithm that is correct with prob $3/4$ on $\mathcal{X}$ over p.

- With probability $2/3$, $\mathcal{X}$ must be able to correctly distinguish between an instance from $\mathcal{F}$ and (A_0, b) (why?).
- To do this, it must be able to find a large entry of size $10n$ with prob $\geq 1/6$ if it exists, in A $\implies$ must read $\Omega(m)$ entries.

Yao's Minimax Principle: Example Problem

$$A_0 = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ \frac{1}{d} & \frac{1}{d} & \frac{1}{d} & \frac{1}{d} \\ \vdots & \vdots & \dots & \vdots \\ \frac{1}{d} & \frac{1}{d} & \frac{1}{d} & \frac{1}{d} \\ \vdots & \vdots & \dots & \vdots \\ \frac{1}{d} & \frac{1}{d} & \frac{1}{d} & \frac{1}{d} \end{bmatrix}, \quad A^{i,j} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ \frac{1}{d} & \frac{1}{d} & \frac{1}{d} & \frac{1}{d} \\ \vdots & \vdots & \dots & \vdots \\ \frac{1}{d} & \frac{1}{d} & 10n & \frac{1}{d} \\ \vdots & \vdots & \dots & \vdots \\ \frac{1}{d} & \frac{1}{d} & \frac{1}{d} & \frac{1}{d} \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^n$$

If $A \in \mathcal{F}$, then the OPT error is > 0, if $A = A_0$, OPT error is 0.

Yao's Minimax Principle: Example Problem

- Suppose $\exists$ randomized algorithm that reads $o(m)$ entries in expectation and is correct with prob $3/4$. Then there is a deterministic algorithm which is correct with prob $3/4$ over p, and reads $o(m)$ entries in expectation.
- By Markov's, there is a deterministic algorithm which is correct with prob $2/3$ and *always* reads at most $o(m)$ entries. Call this Alg $\mathcal{A}$.
- If $\mathcal{A}$ is correct with prob $2/3$ over (A, b) drawn from distribution p, then it must distinguish $A = A_0$ vs $A \in \mathcal{F}$ with prob $2/3$.
- But the probability that it finds a heavy coordinate in A if $A \in \mathcal{F}$ is $\frac{\text{\# of entries read by } \mathcal{A}}{m-d} = o(1)$, so it cannot distinguish the two cases with constant prob.