CSCE 689: Special Topics in Modern Algorithms for Data Science

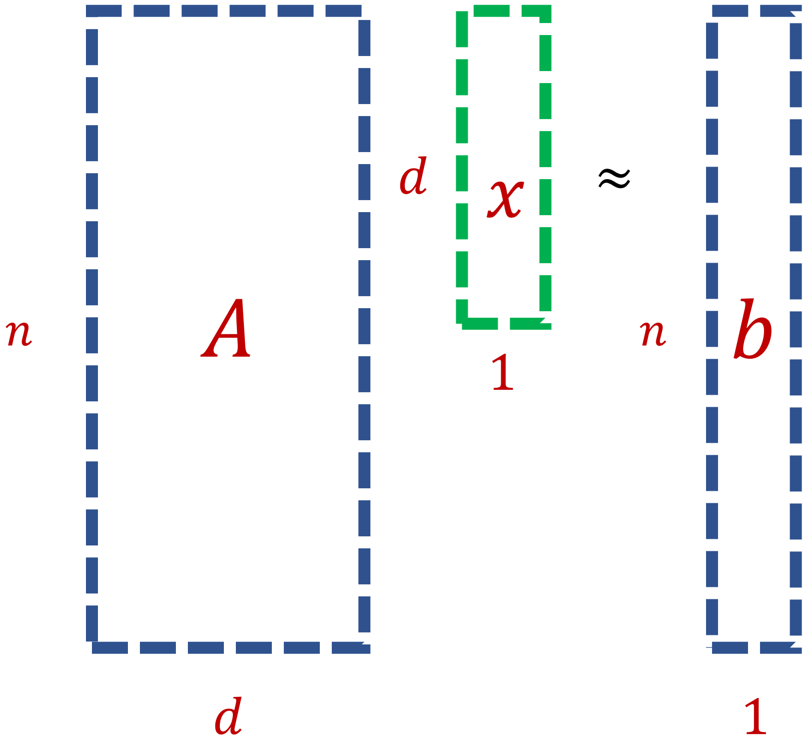
Lecture 31

Samson Zhou

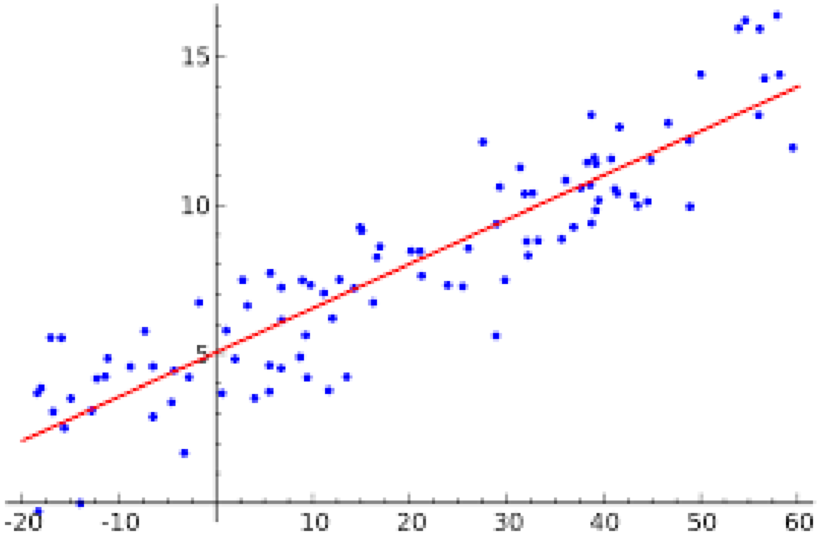
Presentation Schedule

- November 27: Chunkai, Jung, Galaxy AI
- November 29: STMI, Anmol, Jason
- December 1: Bokun, Ayesha, Dawei, Lipai

Last Time: Linear Regression



- Find the vector x that minimizes $\|Ax - b\|_2$
- "Least squares" optimization



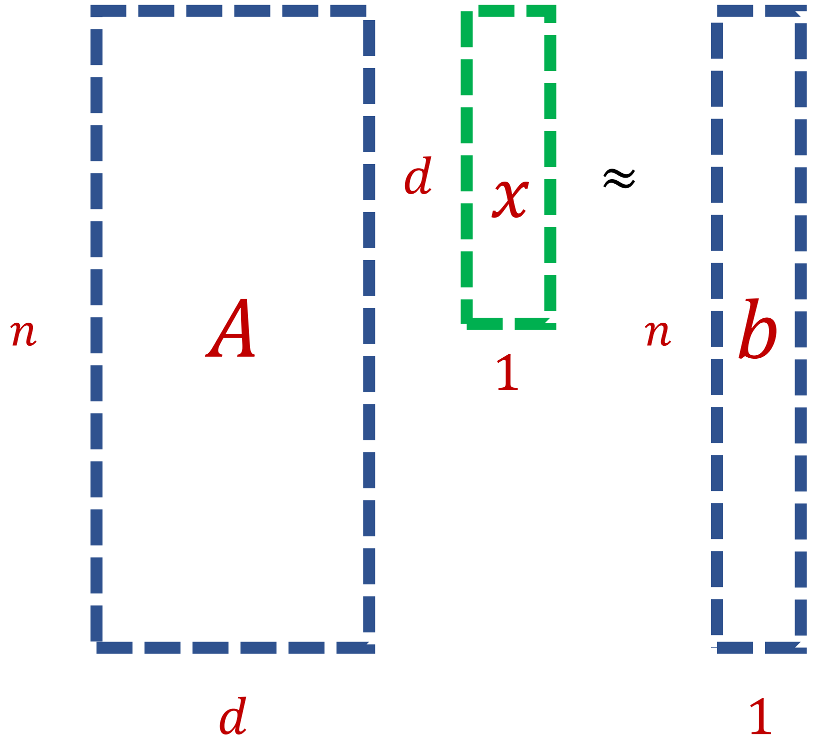
Last Time: Linear Regression

- We have $\arg\min_{x \in \mathbb{R}^d} \|Ax - b\|_2 = \arg\min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2$

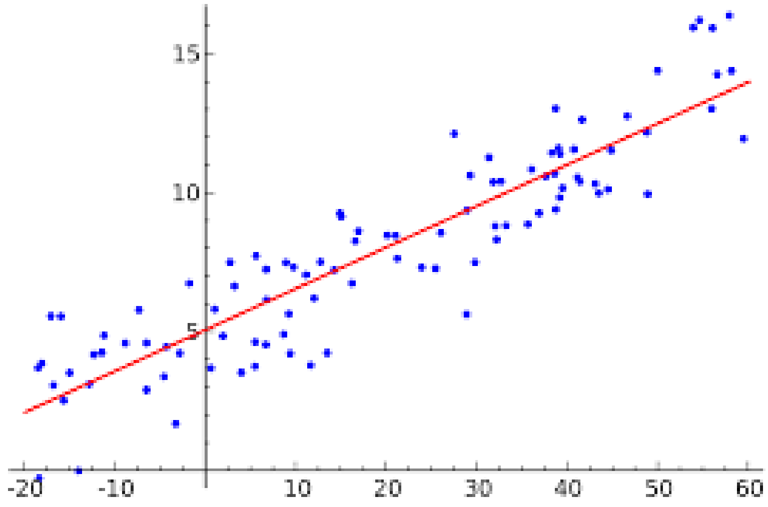
- $$\begin{aligned} \|Ax - b\|_2^2 &= \left\|Ax - b^\perp - b^\parallel\right\|_2^2 \\ &= \left\|Ax - b^\parallel\right\|_2^2 - 2\left\langle Ax - b^\parallel, b^\perp \right\rangle + \|b^\perp\|_2^2 \\ &= \left\|Ax - b^\parallel\right\|_2^2 + \|b^\perp\|_2^2 \end{aligned}$$

- Minimized for $\left\|Ax - b^\parallel\right\|_2^2 = 0$ when $x = A^\dagger b^\parallel = A^\dagger b$

Last Time: Linear Regression



- Find the vector x that minimizes $\|Ax - b\|_2$
- “Least squares” optimization
- MLE under Gaussian noise
- Closed form solution: $x = A^{\dagger} b$



Previously: Coreset Construction and Sampling

- Importance sampling only needs X' to have size $O\left(\frac{1}{\varepsilon^2}\right)$ to achieve $(1+\varepsilon)$-approximation to $\mathrm{Cost}(X, C)$
- To handle all possible sets of k centers:
  - Need to sample each point x with probability $\max_C \frac{\mathrm{Cost}(x,C)}{\mathrm{Cost}(X,C)}$ instead of $\frac{\mathrm{Cost}(x,C)}{\mathrm{Cost}(X,C)}$
  - Need to union bound over a net of all possible sets of k centers

Net with size $\left(\frac{n\Delta}{\varepsilon}\right)^{O(kd)}$

Previously: Sensitivity Sampling

- The quantity $s(x) = \max_C \frac{\mathrm{Cost}(x,C)}{\mathrm{Cost}(X,C)}$ is called the *sensitivity* of x and intuitively measures how "important" the point x is

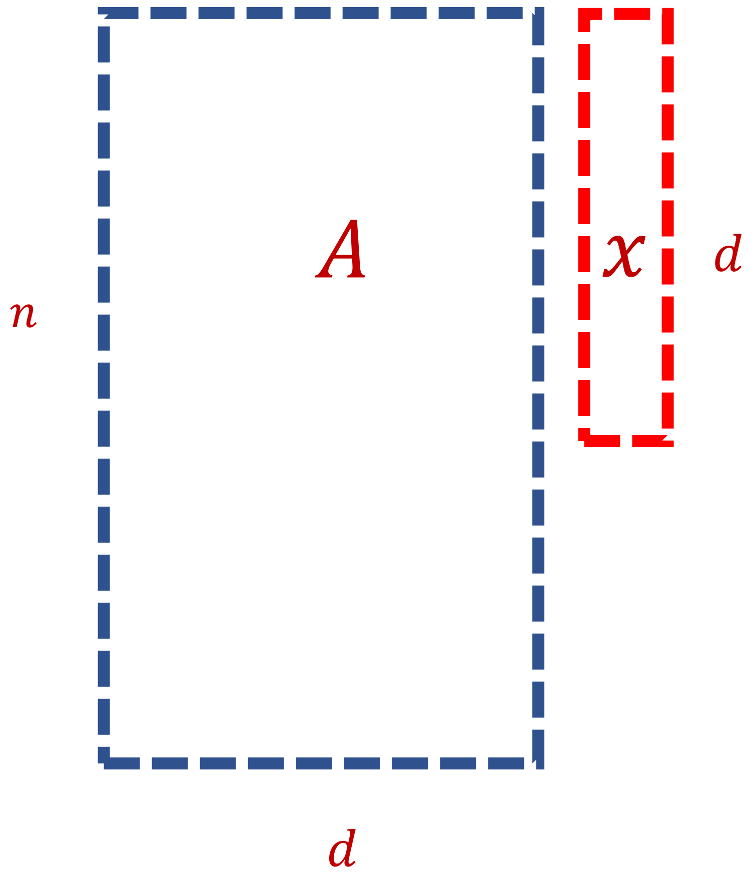
- The *total sensitivity* of X is $\sum_{x \in X} s(x)$ and quantifies how many points will be sampled into X' through importance/sensitivity sampling (before the union bound)

Previously: Sensitivity Sampling

- Recall: $\frac{kd}{\varepsilon^2} \cdot \log \frac{n\Delta}{\varepsilon} \cdot \sum_{x \in X} s(x)$ points sampled
- $\sum_{x \in X} s(x) = O_z(k)$

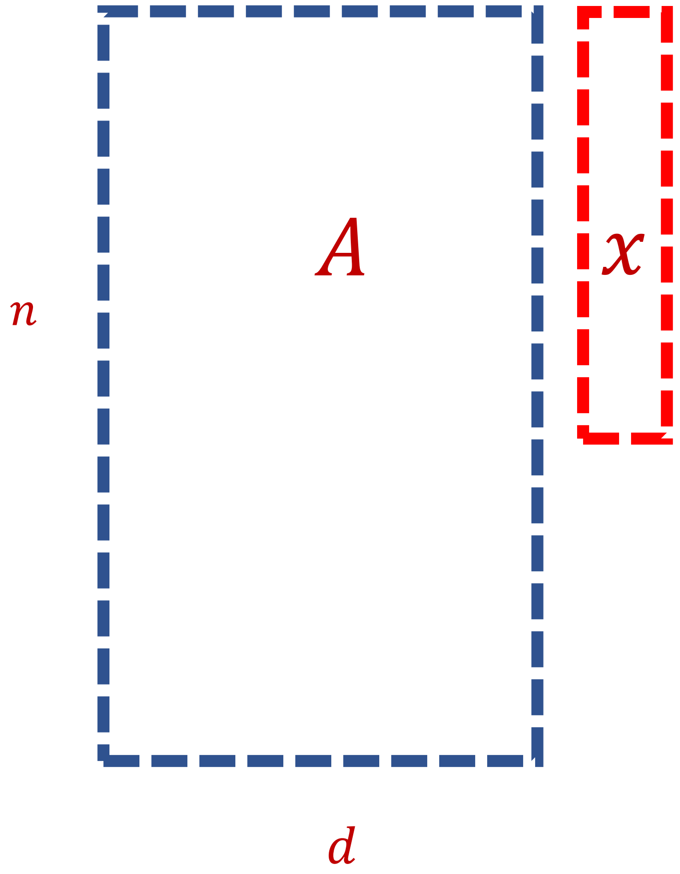
- In total, roughly $\frac{k^2 d}{\varepsilon^2} \cdot \log \frac{n\Delta}{\varepsilon}$ points sampled in expectation

Linear Algebra Review



- For $y = Ax$, we have $y_i = \langle a_i, x \rangle$
- $\|Ax\|_2^2 = \langle a_1, x \rangle^2 + \cdots + \langle a_n, x \rangle^2$

Subspace Embedding

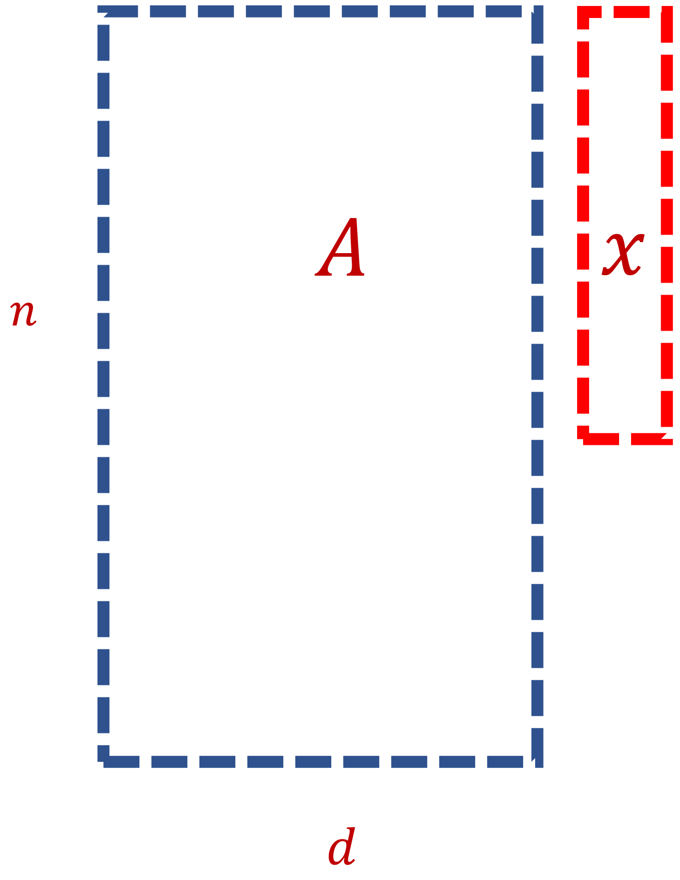


- Subspace embedding: Given $\varepsilon > 0$ and $A \in R^{n\times d}$, find matrix $M \in R^{m\times d}$ with $m \ll n$, such that for *every* $x \in \mathbb{R}^d$,

$$(1-\varepsilon)\|Ax\|_2 \leq \|Mx\|_2 \leq (1+\varepsilon)\|Ax\|_2$$

- Equivalent to $(1-\varepsilon)A^\top A \preccurlyeq M^\top M \preccurlyeq (1+\varepsilon)A^\top A$
- Approximates *all* cuts of a graph when $A^\top A$ is graph Laplacian

Subspace Embedding



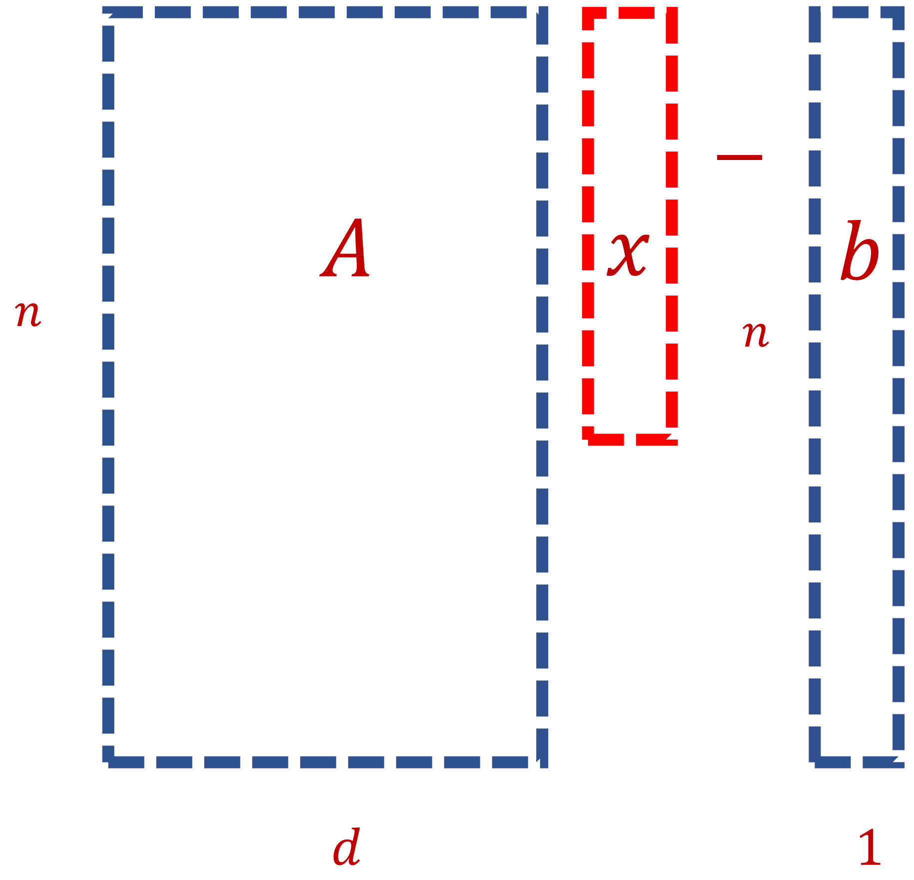
- Subspace embedding: Given $\varepsilon > 0$ and $A \in R^{n\times d}$, find matrix $M \in R^{m\times d}$ with $m \ll n$, such that for *every* $x \in \mathbb{R}^d$,

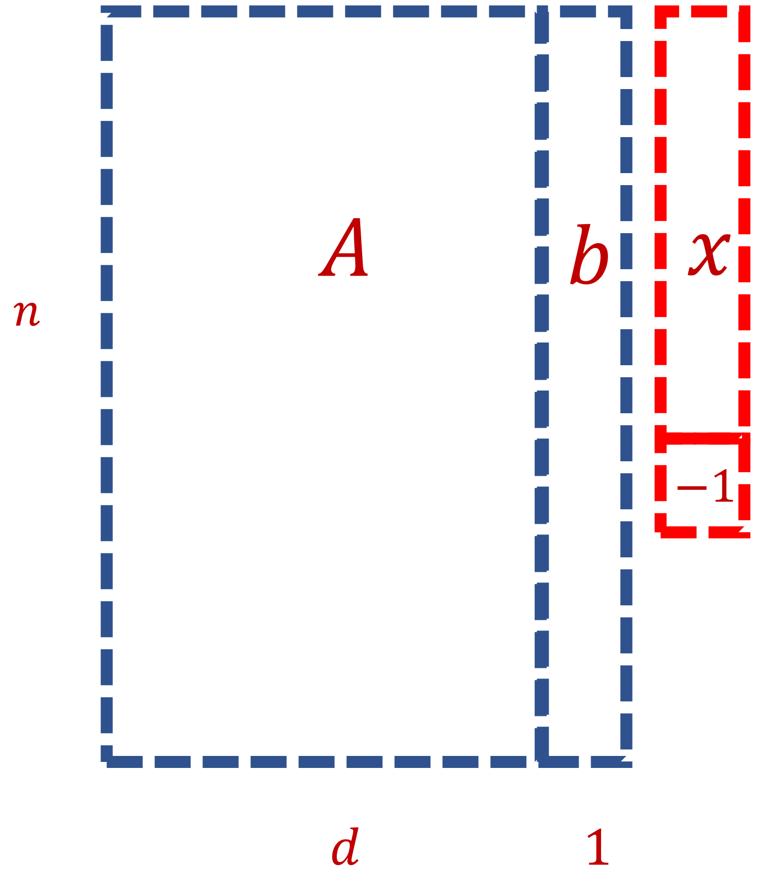
$$(1-\varepsilon)\|Ax\|_2 \leq \|Mx\|_2 \leq (1+\varepsilon)\|Ax\|_2$$

- Claim: A construction of a subspace embedding can be used to approximately solve linear regression

Regression and Subspace Embeddings

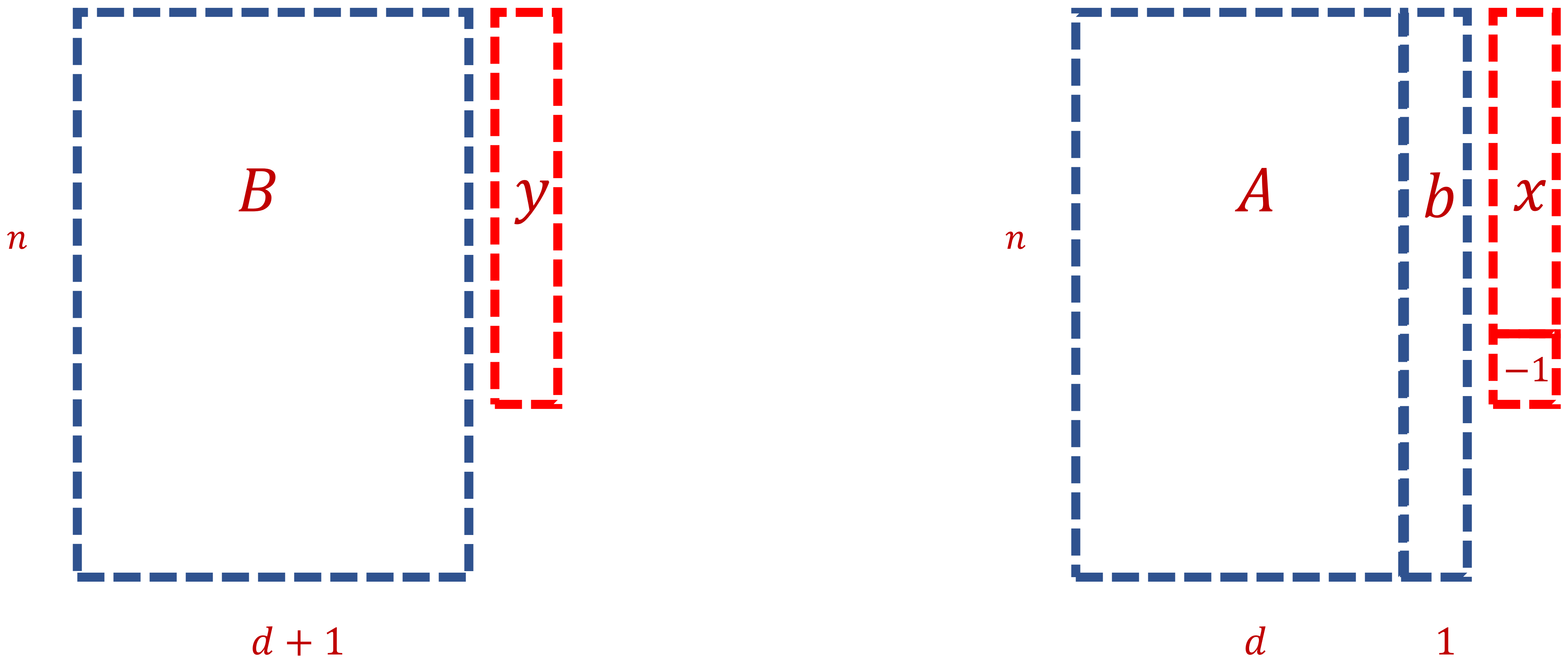
- Recall: Goal is to find x that minimizes $\|Ax - b\|_2$





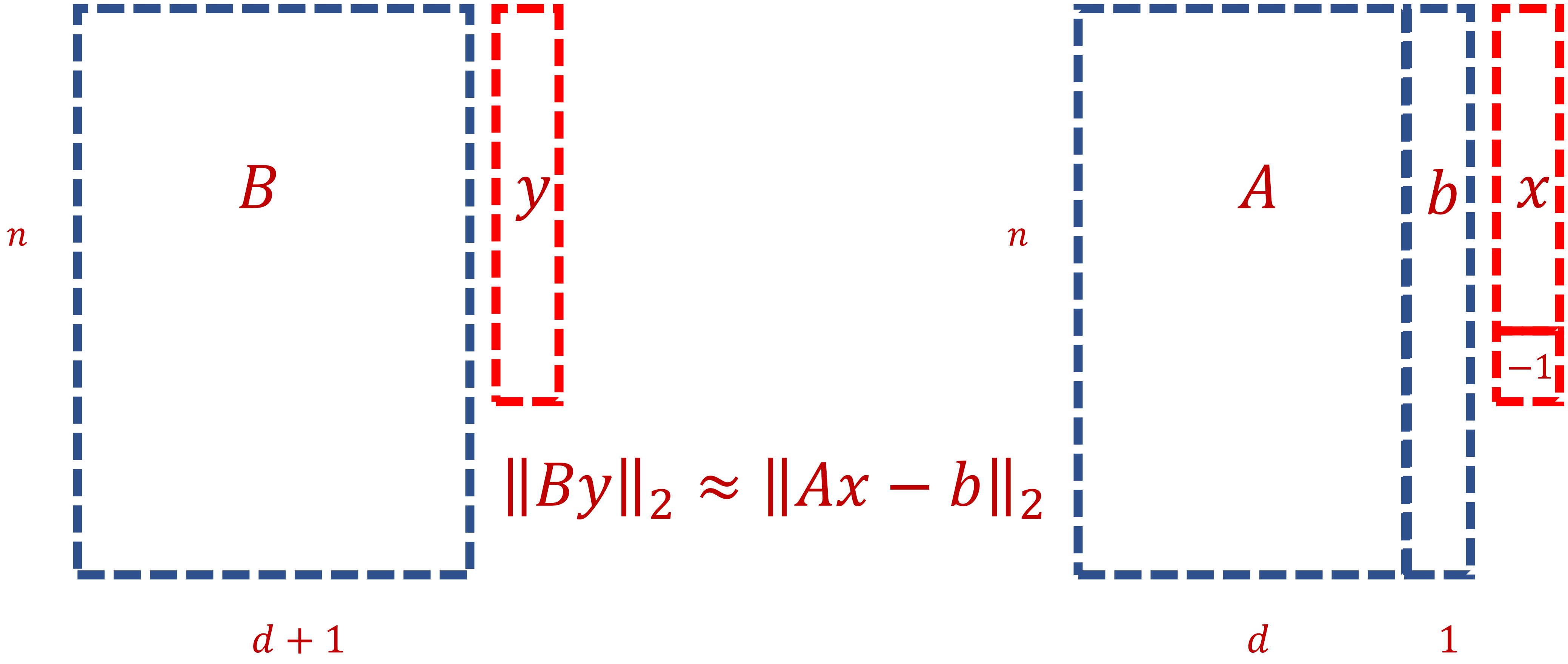
Regression and Subspace Embeddings

- Recall: Goal is to find x that minimizes $\|Ax - b\|_2$



Regression and Subspace Embeddings

- Recall: Goal is to find x that minimizes $\|Ax - b\|_2$



$$\|By\|_2 \approx \|Ax - b\|_2$$

Previously: Coreset Construction and Sampling

- Importance sampling only needs X' to have size $O\left(\frac{1}{\varepsilon^2}\right)$ to achieve $(1+\varepsilon)$-approximation to $\text{Cost}(X, C)$
- To handle all possible sets of k centers:
  - Need to sample each point x with probability $\max_C \frac{\text{Cost}(x,C)}{\text{Cost}(X,C)}$ instead of $\frac{\text{Cost}(x,C)}{\text{Cost}(X,C)}$
  - Need to union bound over a net of all possible sets of k centers

Net with size $\left(\frac{n\Delta}{\varepsilon}\right)^{O(kd)}$

Previously: Sensitivity Sampling

- The quantity $s(x) = \max_C \frac{\mathrm{Cost}(x,C)}{\mathrm{Cost}(X,C)}$ is called the *sensitivity* of x and intuitively measures how "important" the point x is

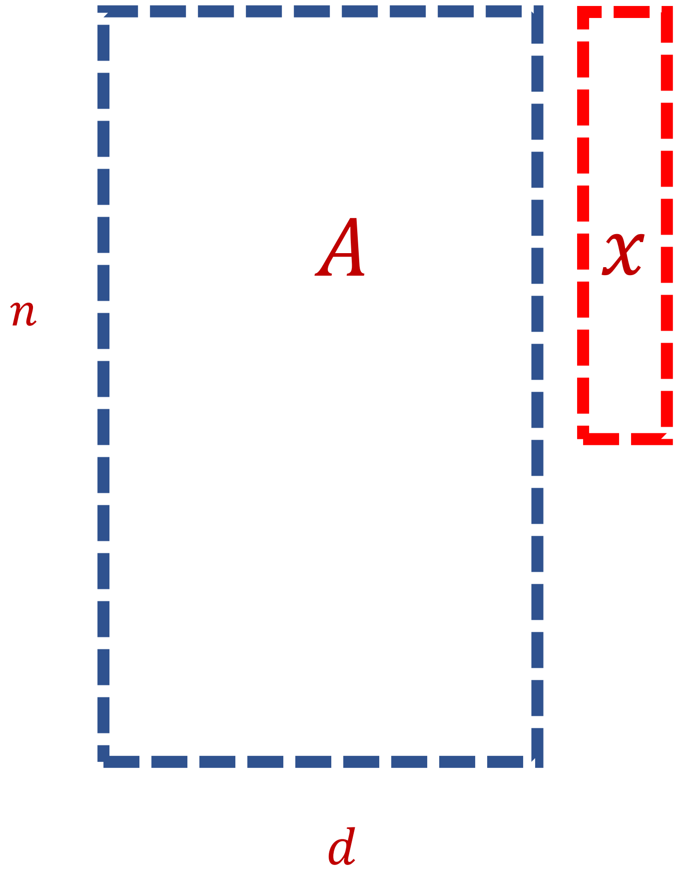
- The *total sensitivity* of X is $\sum_{x \in X} s(x)$ and quantifies how many points will be sampled into X' through importance/sensitivity sampling (before the union bound)

Previously: Sensitivity Sampling

- Recall: $\frac{kd}{\varepsilon^2} \cdot \log \frac{n\Delta}{\varepsilon} \cdot \sum_{x \in X} s(x)$ points sampled
- $\sum_{x \in X} s(x) = O_z(k)$

- In total, roughly $\frac{k^2 d}{\varepsilon^2} \cdot \log \frac{n\Delta}{\varepsilon}$ points sampled in expectation

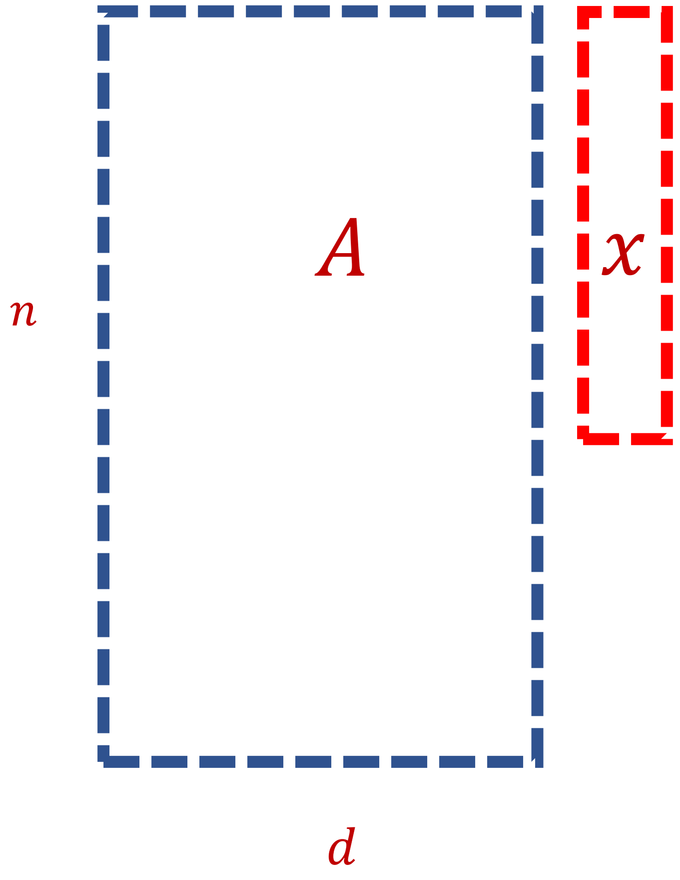
Subspace Embedding



- Subspace embedding: Given $\varepsilon > 0$ and $A \in \mathbb{R}^{n \times d}$, find matrix $M \in \mathbb{R}^{m \times d}$ with $m \ll n$, such that for *every* $x \in \mathbb{R}^d$,

$$(1-\varepsilon)\|Ax\|_2 \leq \|Mx\|_2 \leq (1+\varepsilon)\|Ax\|_2$$

Subspace Embedding



- Subspace embedding: Given $\varepsilon > 0$ and $A \in \mathbb{R}^{n \times d}$, find matrix $M \in \mathbb{R}^{m \times d}$ with $m \ll n$, such that for *every* $x \in \mathbb{R}^d$,

$$(1-\varepsilon)\|Ax\|_2^2 \le \|Mx\|_2^2 \le (1+\varepsilon)\|Ax\|_2^2$$

- Recall: $\|Ax\|_2^2 = \langle a_1, x\rangle^2 + \cdots + \langle a_n, x\rangle^2$

Subspace Embedding

- Subspace embedding: Given $\varepsilon > 0$ and $A \in \mathbb{R}^{n \times d}$, find matrix $M \in \mathbb{R}^{m \times d}$ with $m \ll n$, such that for *every* $x \in \mathbb{R}^d$,

$$(1-\varepsilon)\|Ax\|_2^2 \leq \|Mx\|_2^2 \leq (1+\varepsilon)\|Ax\|_2^2$$

- Question: For a *fixed* $x \in \mathbb{R}^d$, how would we produce a matrix M such that $\|Mx\|_2^2 \approx \|Ax\|_2^2$?

Subspace Embedding

- Question: For a *fixed* $x \in \mathbb{R}^d$, how would we produce a matrix M such that $\|Mx\|_2^2 \approx \|Ax\|_2^2$?

- Recall that $\|Ax\|_2^2 = \langle a_1, x\rangle^2 + \cdots + \langle a_n, x\rangle^2$

- Hint #1: What if M is a weighted subset of rows of A, i.e., a coreset?

Subspace Embedding

- Question: For a *fixed* $x \in \mathbb{R}^d$, how would we produce a matrix M such that $\|Mx\|_2^2 \approx \|Ax\|_2^2$?

- Recall that $\|Ax\|_2^2 = \langle a_1, x\rangle^2 + \cdots + \langle a_n, x\rangle^2$

- Hint #2: What if $\langle a_1, x\rangle^2 = \cdots = \langle a_n, x\rangle^2$?

Bernstein's Inequality

- Bernstein's inequality: Let $y_1, \ldots, y_n \in [-M, M]$ be independent random variables and let $y = y_1 + \cdots + y_n$ have mean μ and variance σ^2. Then for any $t \geq 0$:

$$\Pr[|y - \mu| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2 + \frac{4}{3}Mt}}$$

Coreset Construction and Uniform Sampling

- Consider a fixed $x \in \mathbb{R}^d$, which induces "cost" $\|Ax\|_2^2$
- Suppose all rows have the same cost, $\langle a_1, x\rangle^2 = \cdots = \langle a_n, x\rangle^2$
- Can get a 2-approximation to $\|Ax\|_2^2$ even for $p = \Theta\left(\frac{1}{n}\right)$
- How many samples do we expect? $np = \Theta(1)$

Coreset Construction and Uniform Sampling

- Consider a fixed $x \in \mathbb{R}^d$, which induces "cost" $\|Ax\|_2^2$
- Suppose all rows have cost between 1 and n
- Suppose $p_i = p$ for all $i \in [n]$
- How many rows do I need to sample to approximate $\|Ax\|_2^2$ within a $(1 + \varepsilon)$-factor?

Uniform Sampling for Subspace Embedding

- Consider a fixed $x \in \mathbb{R}^d$, which induces "cost" $\|Ax\|_2^2$
- Suppose all rows have cost between 1 and n
- Suppose $p_i = p$ for all $i \in [n]$
- For Bernstein's inequality, we require $\frac{2n^2}{p} \approx \left(\frac{\|Ax\|_2^2}{2}\right)^2$ and $\|Ax\|_2^2$ can be as small as n, so we need $p \approx 1$

Coreset Construction and Sampling

- Importance sampling only needs M to have $O\left(\frac{1}{\varepsilon^2}\right)$ rows to achieve $(1+\varepsilon)$-approximation to $\|Ax\|_2^2$
- To handle all possible sets of k centers:
  - Need to sample each row a_i with probability $\max_{x \in \mathbb{R}^d} \frac{\langle a_1, x\rangle^2}{\|Ax\|_2^2}$ instead of $\frac{\langle a_1, x\rangle^2}{\|Ax\|_2^2}$
  - Need to union bound over a net of all choices of $x \in \mathbb{R}^d$

Leverage Scores

- Intuition: how *unique* a row is (recall importance sampling)
- $\ell_i = \max_{x \in \mathbb{R}^d} \frac{\langle a_1, x \rangle^2}{\|Ax\|_2^2}$ are the *leverage scores* of A (in this case of row a_i)

$\boxed{\begin{matrix} 1 & 0 \\ 1 & 1 \end{matrix}}$

  - Take $x = (1 \quad -1)$ to see that $\ell_1 = 1$
  - Take $x = (0 \quad 1)$ to see that $\ell_2 = 1$

- $\ell_i = a_i (A^\top A)^{-1} a_i^\top, \ \sum \ell_i = d$