CSCE 689: Special Topics in Modern Algorithms for Data Science Fall 2023

Lecture 31 — November 13, 2023

Prof. Samson Zhou *Scribe: Tzu-Shen (Jason), Wang*

1 Review

Linear algebra review. For $y = Ax$, we have $y_i = \langle a_i, x\rangle$, where $A \in R^{n\times d}$ and $x \in R^{d\times 1}$

Recall the following formulation of Bernstein's inequality:

Theorem 1 (Bernstein's inequality)**.** *Let $y_1, ..., y_n \in [-M, M]$ be independent random variables and let $y = y_1 + ... + y_n$ have mean μ and variance σ^2. Then for any $t \geq 0$, we have:*

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

Coreset construction and sampling. Importance sampling only $O\left(\frac{1}{\varepsilon^2}\right)$ samples 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, where Net with size $\left(\frac{n\Delta}{\varepsilon}\right)^{O(kd)}$

2 Subspace Embedding

Definition (Subspace embedding)**.** Given matrix $A \in R^{n\times d}$, a *subspace embedding* is a matrix $M \in R^{m\times d}$, with $m \ll n$, such that for every $x \in \mathbb{R}^d$, we have:

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

Claim 1. Subspace embeddings can be used to approximately solve linear regression

Recall that a regression is to find x that minimize $\|Ax - b\|_2$
We show how to utilize subspace embedding to solve approximate regression. Observe that we can set B to be equal to the matrix A concatenated with the column vector b, and append -1 to the last row of x.

i.e. $B = [A \mid b]$, $y = \begin{bmatrix} x \\ -1 \end{bmatrix}$, so $By = Ax - b$.
By computing a subspace embedding for B, we have $M \in \mathbb{R}^{m \times d}$, where for every y we have

$$(1-\varepsilon)\|By\|_2 \le \|My\|_2 \le (1+\varepsilon)\|By\|_2.$$

Then we use solve approximate regression by solving the regression problem given M and b, so that the answer is a ε-approximation by the guarantee of the subspace sampling problem (for every y we have $(1-\varepsilon)\|By\|_2 \le \|My\|_2 \le (1+\varepsilon)\|By\|_2$).

2.1 Intuition of using subspace embedding to solve approximate regression problem

One can solve the regression problem by computing x as $A^{\dagger}b$. However, since $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^{d \times 1}$, it takes time $\Theta(nd^2)$ using naive matrix multiplication. Whereas subspace embedding gives a matrix $M \in \mathbb{R}^{m \times d}$, where $m \ll n$, the time complexity is reduced to $\Theta(md^2)$.

2.2 Solving subspace embedding:

Consider a fixed $x \in \mathbb{R}^d$, which induces cost $\|Ax\|_2^2$. We intend to find matrix $M \in \mathbb{R}^{m \times d}$, with $m \ll n$, s.t for every $x \in R^d$,

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

Uniform sampling. In a simple case: Suppose all rows induce the same cost $\langle a_1, x\rangle^2, \langle a_2, x\rangle^2, ..., \langle a_n, x\rangle^2$. Then we can use uniform sampling, each row will be sampled by a probability of $p = \Theta\left(\frac{1}{n}\right)$. And the expected number of samples are $np = \Theta(1)$, which is only a constant number of samples. However, if we consider all rows have cost between 1 and n and suppose each row i is still sampled with the same probability, i.e., $p_i = p$, then by Bernstein's equality, we might need $\frac{2n^2}{p} \approx \frac{\|Ax\|_2^2{}^2}{2}$ and $\|Ax\|_2^2$ can be as small as n. Thus we need $p \approx 1$, so we sample approximately $\Theta(n)$ rows.

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 $\|Ax\|_2^2$
However to handle all possible $x \in \mathbb{R}^d$:

- Need to sample row a_i with probability $\max_{x \in \mathbb{R}^d} \frac{\langle a_i, x\rangle^2}{\|Ax\|_2^2}$ instead of just $\frac{\langle a_i, x\rangle^2}{\|Ax\|_2^2}$
- Need to union bound over all $x \in \mathbb{R}^d$

Leverage scores Intuition: how unique a row is $\ell_i = \max_{x \in \mathbb{R}^d} \frac{\langle a_i, x\rangle^2}{\|Ax\|_2^2}$ is the leverage score of row a_i in A.

Example 1. E.g., For $A = \begin{bmatrix} 1 & 0 \\ 1 & 1. \end{bmatrix}$:

- If we take $x = (1, -1)$ then $\ell_1 = 1$ (row 1 contributes all, so we must pick row 1)
- If we take $x = (0, 1)$ then $\ell_2 = 1$

It is known that $\ell_i = a_i(A^\top A)^{-1} a_i^\top$, so that $\Sigma \ell_i = d$, we expect to sample d rows, where $d \ll n$.