| | |
|---|---|
| **CSCE 689: Special Topics in Modern Algorithms for Data Science** | Fall 2023 |
| Lecture 25 — October 27, 2023 | |
| *Prof. Samson Zhou* | *Scribe: Ayesha Qamar* |

# 1 Recall

**Theorem 1** (Bernstein's inequality)**.** *Let $y_1, ..., y_n \in [-M, M]$ be independent random variables and let $y_1, ..., y_n$ have mean $\mu$ and variance $\sigma^2$. Then for any $t \geq 0$ : $Pr[]|y-\mu| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2+\frac{4}{3}Mt}}$.*

# 2 Coreset Construction

Given a fixed set $X$ and a fixed set $C$ of $k$ centers, which induces a fixed cost $\text{Cost}(X, C)$. The goal is to find $X'$ with $\text{Cost}(X', C) \approx \text{Cost}(X, C)$.

## 2.1 Uniform Sampling

Uniformly sample points from $X$ to obtain $X'$.

If all points $x$ have the same cost as $\text{Cost}(x, C) = \frac{\text{Cost}(X,C)}{n}$, then following Theorem 1 to get a 2-approximation, set $M = \frac{1}{p}, t = \frac{1}{2}.\text{Cost}(X, C)$ and $\sigma^2 \approx \frac{n}{p}$ for $x = \text{Cost}(X, C)$, so that

$$\Pr\left[]|y-\mu| \geq \frac{x}{2}\right] \leq 2\exp\left(-\frac{(\frac{x}{2})^2}{2(\frac{4n}{p}) + (\frac{4}{3})(\frac{x}{p})}\right).$$

We need $\frac{8n}{p} \approx (\frac{x}{2})^2$ and $x$ can be as small as $n$, so 2-approximation to $\text{Cost}(X, C)$ is possible for $p = \Theta(1/n)$ and number of samples $np = \Theta(1)$.

Suppose all points have cost $\in [1, 100]$ and $p_i = p$ for all $i \in [n]$, set $M = \frac{100}{p}, t = \frac{1}{2}.\text{Cost}(X, C)$, $\sigma^2 \approx \frac{10000n}{p}$, and $x = \text{Cost}(X, C)$

$$\Pr\left[]|y-\mu| \geq \frac{x}{2}\right] \leq 2\exp\left(-\frac{(\frac{x}{2})^2}{2(\frac{10000n}{p}) + (\frac{4}{3})(\frac{50x}{p})}\right).$$

We need $\frac{20000n}{p} \approx (\frac{x}{2})^2$ and $x$ can be as small as $n$, so we need $p \approx \frac{80000}{n}$.

Now suppose all points have cost between 1 and $n$. To approximate cost within $(1+\varepsilon)$-factor, set $M = \frac{n}{p}, t = \frac{x}{2}$, $\sigma^2 \approx \frac{n^2}{p}$ then

$$\Pr\left[]|y-\mu| \geq \frac{x}{2}\right] \leq 2\exp\left(-\frac{(\frac{x}{2})^2}{2(\frac{n^2}{p}) + (\frac{4}{3})(\frac{nx}{2p})}\right).$$

We need $\frac{2n^2}{p} \approx (\frac{x}{2})^2$ and $x$ can be as small as $n$, so we need $p \approx 1$.

Therefore, uniform sampling needs a lot of samples if there is an outlier present in the data—i.e., if one point affects $\text{Cost}(X, C)$ greatly.

## 2.2 Importance Sampling

Let $y_i$ be the contribution of $x_i$ when it is sampled with probability $p_i$, so that

$$y_i = \begin{cases} 0, & \text{w.p. } 1-p_i \\ \frac{\text{Cost}(x_i,C)}{p_i}, & \text{w.p. } p_i \end{cases} = \frac{\text{Cost}(x_i, C)}{p_i}.$$

Observe that:

- $\text{Var}[y_i] \leq \frac{1}{p_i}.(\text{Cost}(x_i, C))^2 \leq \text{Cost}(x_i, C).\text{Cost}(X, C)$
- $\text{Var}[y] \leq \text{Var}[y_1] + ... + \text{Var}[y_n] \leq (\text{Cost}(x_i, C))^2$

Thus we have:

$$\mathbb{E}[\text{Cost}(y_i, C)^2] = \begin{cases} \frac{\text{Cost}(x_i,C)^2}{p_i^2}, & \text{w.p. } p_i \\ 0, & \text{w.p. } 1-p_i \end{cases} = \frac{\text{Cost}(x_i, C)^2}{p_i} = \text{Cost}(X, C).\text{Cost}(x_i, C)$$

Importance sampling only needs $X'$ to have size $O(\frac{1}{\varepsilon^2})$ in expectation to achieve $(1+\varepsilon)$-approximation to $\text{Cost}(X, C)$

**Definition.** A net $N$ is a set of sets $C$ of $k$ centers such that accuracy on $N$ implies accuracy everywhere

In order to handle all possible $k$ centers, sample each point $x$ with probability $\max_C \frac{\text{Cost}(x,C)}{\text{Cost}(X,C)}$. Need to union bound over a net of all possible sets of $k$ centers with a net size of $(\frac{n\Delta}{\varepsilon})^{O(kd)}$

### 2.2.1 Sensitivity Sampling

The quantity $s(x) = \max_C \frac{\text{Cost}(x,C)}{\text{Cost}(X,C)}$ is called the sensitivity of $x$ and intuitively measures how ïmportantẗhe 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 sensitivity sampling.