Lecture 25 — October 30, 2023

Prof. Samson Zhou *Scribe: Lipai Huang*

1 Recall

1.1 Coreset Construction and Sampling

Consider a fixed set X and a fixed set C of k centers, which includes a fixed cost $\text{Cost}(X, C)$.

- Uniform Sampling needs a lot of samples if there is a single point that greatly contributes to $\text{Cost}(X, C)$.
- Importance Sampling, sample each point $x \in X$ into X' with probability proportional $\text{Cost}(X', C)$, and X' with size $O(\frac{1}{\varepsilon^2})$ achieves $(1+\varepsilon)$-approximation

1.2 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 "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).

2 Coreset Construction and Sensitivity Sampling

Definition. If we sample each point with probability $p(x) = min(\frac{s(x)}{\varepsilon^2} \log \frac{1}{\delta})$, then we get achieve $(1+\varepsilon) - approximation$ to $\text{Cost}(X, C)$ with probability $1 - \delta$.

What should δ be? How many points are sampled?
Recall net with size $(\frac{n\Delta}{\varepsilon})^{O(kd)}$, and correctness on net implies correctness everywhere, so we set $\delta = \frac{1}{100}(\frac{\varepsilon}{n\Delta})^{O(kd)}$ and by a union bound, our algorithm succeeds with probability 0.99. Also $p(x) := \min\left(\frac{s(x)}{\varepsilon^2} \log \frac{1}{\delta}, 1\right)$, so we sample $\sum_{x \in X} p(x)$ points in expectation. In addition, at most $\frac{1}{\varepsilon^2} \log \frac{1}{\delta} \sum_{x \in X} s(x)$ points in total, since $\log \frac{1}{\delta} = kd \cdot \log \frac{n\Delta}{\varepsilon}$, then we can sample at most $\frac{kd}{\varepsilon^2} \cdot \log \frac{n\Delta}{\varepsilon} \cdot \sum_{x \in X} s(x)$ points and $\sum_{x \in X} s(x)$ is total sensitivity.

Total Sensitivity

Total sensitivity = Sum of the sensitivities can be at least k (imagine a set of $k+1$ distinct points, which can each have sensitivity 1 when the k centers are placed at the other k points)

$$s(x_t) = \max_{C:|C|\leq k} \frac{\mathrm{Cost}(x_t, C)}{\mathrm{Cost}(X, C)} = \max_{C:|C|\leq k} \frac{\mathrm{Cost}(x_t, C)}{\sum_{i=1}^{n} \mathrm{Cost}(x_i, C)}$$

Intuition: The sum of sensitivities in each cluster induced by **OPT** is at most 1. Since there are k clusters, the sum of the sensitivities is $O_z(k)$.
We have $\sum_{x \in X} s(x) = O_z(k)$.
To sum up, roughly $\frac{k^2 d}{\varepsilon^2} \cdot \log \frac{n\Delta}{\varepsilon}$ points sampled in expectation.

How to compute Sensitivities?

- Estimations to sensitivities suffice
- Bicriteria algorithms