Randomized Algorithms 2019A – Lecture 8
Coresets via Uniform and Importance Sampling*

Robert Krauthgamer

1 Concentration bounds

Chernoff-Hoeffding bound: Let $X = \sum_{i\in[n]} X_i$ where $X_i \in [0,1]$ for $i \in [n]$ are independently distributed random variables. Then

$$\begin{aligned}
\forall t > 0, &\qquad \Pr[|X - \mathbb{E}[X]| \ge t] \le 2e^{-2t^2/n}. \\
\forall 0 < \varepsilon \le 1, &\qquad \Pr[X \le (1-\varepsilon)\,\mathbb{E}[X]] \le e^{-\varepsilon^2\,\mathbb{E}[X]/2}. \\
\forall 0 < \varepsilon \le 1, &\qquad \Pr[X \ge (1+\varepsilon)\,\mathbb{E}[X]] \le e^{-\varepsilon^2\,\mathbb{E}[X]/3}. \\
\forall t \ge 2e\,\mathbb{E}[X], &\qquad \Pr[X \ge t] \le 2^{-t}.
\end{aligned}$$

Exer: Let X be binomial $B(n, 1/3)$. What is the probability that X deviates from its expectation additively by $r > 1$ standard deviations? Think of r being $10, \log n, \sqrt{n}$.

Exer: Let $a_1, \ldots, a_n$ be an array of numbers in the range $[0,1]$. Design a randomized algorithm that estimates their average within $\pm\varepsilon$ (i.e., additive error ε) by reading only $O(1/\varepsilon^2)$ elements. The algorithm should succeed with probability at least 90%.

Exer: Let $S_1, \ldots, S_n$ be subsets of $[n]$. Design an algorithm for 2-coloring the elements $[n]$, such that in every set S_i the balance, defined as $|\#\text{black} - \#\text{white}|$, is at most $O(\sqrt{n \log n})$.

2 Weak Coresets via Uniform Sampling

We study henceforth the case $k = 1$, for which uniform sampling works (although it is rare).

Geometric median: The *geometric median* of n data points $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ is

$$m_X := \operatorname*{argmin}_{m\in\mathbb{R}^d} f(X, \{m\}) = \operatorname*{argmin}_{m} \sum_{x\in X} \|x - m\|.$$

*These notes summarize the material covered in class, usually skipping proofs, details, examples and so forth, and possibly adding some remarks, or pointers. The exercises are for self-practice and need not be handed in. In the interest of brevity, most references and credits were omitted.

Remark: It is easy to see that the minimum is not unique (although it is anyway not important for us).

Theorem 6 (weak coreset): Let X be a set of n points in $\mathbb{R}^d$ and let $\varepsilon \in (0, 1/2)$. Consider a multiset S constructed by sampling independently $|S| \geq L d \varepsilon^{-2} \log \frac{d}{\varepsilon}$ points, each point is chosen uniformly from X, where $L > 0$ is a suitable constant. Then with (constant) high probability,

$$\sum_{x \in X} \|x - m_S\| \leq (1 + \varepsilon) \sum_{x \in X} \|x - m_X\|.$$

Remark: The other direction $\sum_{x \in X} \|x - m_S\| \geq \sum_{x \in X} \|x - m_X\|$ is obvious.

We will need the lemma below, which intuitively shows that if a potential center point b is "not good" for X (compared to the optimum m_X), then most likely it will be "not good" also for a sample S.

Lemma 7: Let X, $\varepsilon' = \varepsilon/5$, and S be as above, and denote $\mathrm{OPT} := \sum_{x \in X} \|x - m_X\|$. If $b \in \mathbb{R}^d$ satisfies

$$\sum_{x \in X} \|x - b\| \geq (1 + 4\varepsilon')\mathrm{OPT},$$

then

$$\Pr\Big[\sum_{x \in S} \|x - b\| \leq \sum_{x \in S} \|x - m_X\| + \varepsilon' |S| \cdot \mathrm{OPT}/n\Big] \leq e^{-\varepsilon'^2 |S|/6}.$$

Proof of Theorem 6: Was seen in class, using the ball-cover lemma to discretize $B(m_X, 4|S| \cdot \mathrm{OPT}/n$, and applying Lemma 7 to the resulting set of points.

Proof of Lemma 7 (sketch): A sketch Was seen in class, using Chernoff bounds.

Exer: Show that uniform sampling does not produce (with high probability) a strong coreset for 1-median.

Hint: place two "extreme" points

3 Strong Coresets via Importance Sampling

Definition: The *sensitivity* of a point $x \in X$ is

$$s(x) := \sup_{c \in \mathbb{R}^d} \frac{\|x - c\|}{\sum_{z \in X} \|z - c\|},$$

and the *total sensitivity* of X is $S(X) = \sum_{x \in X} s(x)$.

Observe that for a given $c \in \mathbb{R}^d$ (i.e., without the supremum) the above ratio is the "desired" sampling probability in Importance Sampling.

Importance Sampling approach: Suppose we sample one point, where each $x \in X$ is picked with probability $q(x) := \frac{s(x)}{S(X)}$. We then give it weight $\frac{1}{q(x)}$. Of course, we should repeat a few times to reduce variance.

Lemma 8: $S(X) \leq 6$.

Lemma 9: Let Y be a multiset of $m \geq 24/\varepsilon^2$ points, each sampled iid from X according to $q(\cdot)$. Then

$$\forall c \in \mathbb{R}^d, \qquad \Pr\left[\tfrac{1}{m}\sum_{y \in Y} \frac{\|y-c\|}{q(y)} \in (1 \pm \varepsilon) \sum_{x \in X} \|x-c\|\right] \geq 3/4.$$

This does not give a strong coreset, but it is an important step in that direction.

Proof of Lemma 8: Was seen in class by bounding each $s(x) \leq \frac{4}{n} + \frac{\|x-c^*\|}{\mathrm{OPT}/2}$.

Proof of Lemma 9: Was seen in class by applying the Importance Sampling Theorem seen in the previous class for each $y \in Y$.

Amplifying the probability: We would like to improve the success probability in Lemma 9 to $1-\delta$. Using Chebyshev's inequality, this would require increasing m by a factor of $\frac{1}{\delta}$.

Using Chernoff-Hoeffding concentration bounds would be better and require increasing m only by a factor of $O(\log \frac{1}{\delta})$. But for this, we need that no one sample $y \in Y$ ever contributes too much, which indeed holds in our setting.

Lemma 10: $\hat{Z} \leq S(X) \cdot \mathbb{E}\,\hat{Z}$ with probability 1.

Proof of Lemma 10: Was seen in class.

Lemma 11: The success probability in Lemma 9 can be improved $1-\delta$ by using $m \geq L\varepsilon^{-2} \log \frac{1}{\delta}$ for a suitable constant $L > 0$.

Exer: Prove this lemma.

Strong Coreset: To obtain a strong coreset, we must consider any $c \in \mathbb{R}^d$. If there were only a few potential centers, then we could apply Lemma 11 to each of them together with a union bound.

The idea is then to discretize the space of potential centers using the ε-ball cover lemma, and show that it suffices to consider only these centers. Then it would suffice to apply Lemma 4 and a union bound.

Theorem 12: Let Y be a multiset of $m \geq L' d \varepsilon^{-2} \log \frac{1}{\varepsilon}$ points from X, each sampled iid according to distribution $q(.)$ and reweighted by $w(x) = \frac{1}{mq(x)}$, for a suitable constant $L' > 0$. Then with high probability, Y is a strong coreset for the geometric median of X.

Due to time constraints, we saw in class only an outline of the proof, which is based on the lemmas below.

One potential obstacle is the total weight of Y. It need not be n, but with high probability should be close.

Lemma 13: Under the conditions of Lemma 11, i.e., $m \geq L\varepsilon^{-2} \log \frac{1}{\delta}$,

$$\Pr[w(Y) \in (1 \pm \varepsilon)n] \geq 1 - \delta.$$

Exer: Prove this lemma using concentration bounds.

Hint: Write $w(Y) = \frac{1}{m} \sum_{y \in Y} \frac{1}{q(y)}$, show a bound $\frac{1}{q(x)} \leq O(n)$ (with probability 1), and then use concentration bound.