# Randomized Algorithms 2025A – Lecture 13*
# Coresets via Importance Sampling

Robert Krauthgamer

## 1 Concentration Bounds (for reference; was skipped in class)

**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

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

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 $\varepsilon$) by reading only $O(1/\varepsilon^2)$ elements. The algorithm should succeed with probability at least 90%.

## 2 Strong Coresets for 1-Median 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 the sampled $x$ new weight $\frac{1}{q(x)}$. Of course, we should average a few repetitions to reduce variance.

*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.

**Lemma 1:** $S(X) \leq 6$.

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

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

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

**Proof of Lemma 1:** Was seen in class by bounding each $s(x) \leq \frac{4}{n} + \frac{\|x-c^*\|}{\mathrm{OPT}/2}$.

**Proof of Lemma 2:** Was seen in class by applying the Importance Sampling Theorem seen in the previous class for each sample $y \in Y$.

**Amplifying the probability:** We would like to improve the success probability in Lemma 2 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 3:** $\hat{Z} \leq S(X)\cdot\mathbb{E}[\hat{Z}]$ with probability 1.

**Proof of Lemma 3:** Was seen in class by direct calculation.

**Lemma 4:** The success probability in Lemma 2 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 using concentration bounds.

**Strong Coreset:** To obtain a strong coreset, we need a bound for all $c \in \mathbb{R}^d$ simultaneously. If there were only a few potential centers, then we could apply Lemma 4 to each of them together with a union bound.

The idea is then to discretize the space of potential centers using the $\varepsilon$-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 5:** 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 1-median of $X$.

Remark: This is not in the scope of the course but the theorem rextends to $k$-median with roughly linear dependence on $k$, and moreover the dependence on $d$ is not really needed.

The following lemma will be needed later. It can also be viewed as a sanity check for the total weight of $Y$, which need not be exactly $n$, but with high probability should be close.

**Lemma 6:** Under the conditions of Lemma 4, 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.

**Proof of Theorem 5:** Was seen in class, by a discretization $N$ of the possible centers (using the method of a ball cover, seen in an earlier class), applying Lemma 4 to the each center $c' \in N$ and taking a union bound. Finally, we show that the bound"extends" from the centers $c' \in N$ to all centers $c \in \mathbb{R}^d$.