CSCE 689: Special Topics in Modern Algorithms for Data Science

Lecture 24

Samson Zhou

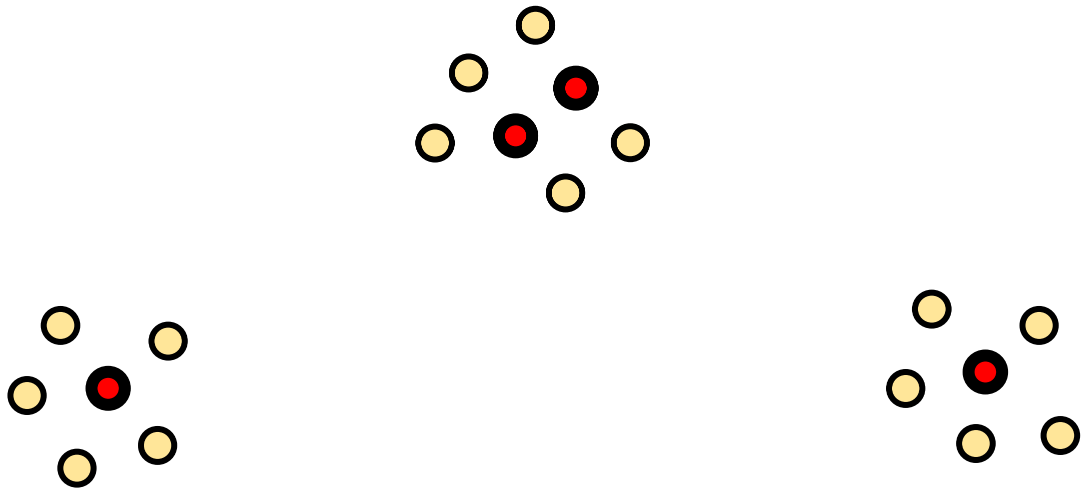
Presentation Schedule

- November 27: Chunkai, Jung, Galaxy AI
- November 29: STMI, Anmol, Jason
- December 1: Bokun, Ayesha, Dawei, Lipai

Previously: Coreset

- Subset X' of representative points of X for a specific clustering objective

- $\mathrm{Cost}(X, C) \approx \mathrm{Cost}(X', C)$ for all sets C with $|C| = k$



Previously: Coreset

- Given a set X and an accuracy parameter $\varepsilon > 0$, we say a set X' with weight function w is an $(1+\varepsilon)$-*multiplicative coreset* for a cost function Cost, if for all queries C with $|C| \le k$, we have

$$(1-\varepsilon)\mathrm{Cost}(X, C) \le \mathrm{Cost}(X', C, w) \le (1+\varepsilon)\mathrm{Cost}(X, C)$$

↑

(k, z)-clustering: $\mathrm{Cost}(X', C, w) = \sum_{x \in X'} w(x) \cdot \left(\mathrm{dist}(x, C)\right)^z$

Previously: Bernstein's Inequality

- Bernstein's inequality: Let $X_1, \ldots, X_n \in [-M, M]$ be independent random variables and let $X = X_1 + \cdots + X_n$ have mean μ and variance σ^2. Then for any $t \geq 0$:

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

- Example: Suppose $M = 1$ and let $t = k\sigma$. Then

$$\Pr[|X - \mu| \geq k\sigma] \leq 2\exp\left(-\frac{k^2}{4}\right)$$

Last Time: Sampling for Sum Estimation

- Consider a fixed set $X = \{x_1, \ldots, x_n\}$ of n numbers

- Suppose we sample each point x_i with some probability p_i and rescale by $\frac{1}{p_i}$

- What is the expected sum? $\mathrm{E}[y_1 + \cdots + y_n] = x_1 + \ \ldots + x_n$
- What can we say about concentration?

Last Time: Uniform Sampling for Sum Estimation

- Suppose $x_1 = \cdots = x_n = 1$
- Suppose $p_i = p$ for all $i \in [n]$

- What can we say about concentration?
- Can get a 2-approximation even for $p = \Theta\left(\frac{1}{n}\right)$
- How many samples do we expect? $np = \Theta(1)$

Last Time: Uniform Sampling for Sum Estimation

- Suppose $x_1, \dots, x_n \in [1,100]$
- Suppose $p_i = p$ for all $i \in [n]$

- What can we say about concentration?
- Can get a 2-approximation even for $p \approx \frac{80000}{n}$
- How many samples do we expect? np is now WAY larger

Last Time: Uniform Sampling for Sum Estimation

- Suppose $x_1, \dots, x_n \in [1, n]$
- Suppose $p_i = p$ for all $i \in [n]$

- What can we say about concentration?
- Can get a 2-approximation for $p \approx 1$
- How many samples do we expect? np is now n

Uniform Sampling for Sum Estimation

- Suppose $x_1, \ldots, x_n \in [1, n]$
- Suppose $p_i = p$ for all $i \in [n]$

- Do we really need p to be a constant? YES!

$$1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ n\ n$$

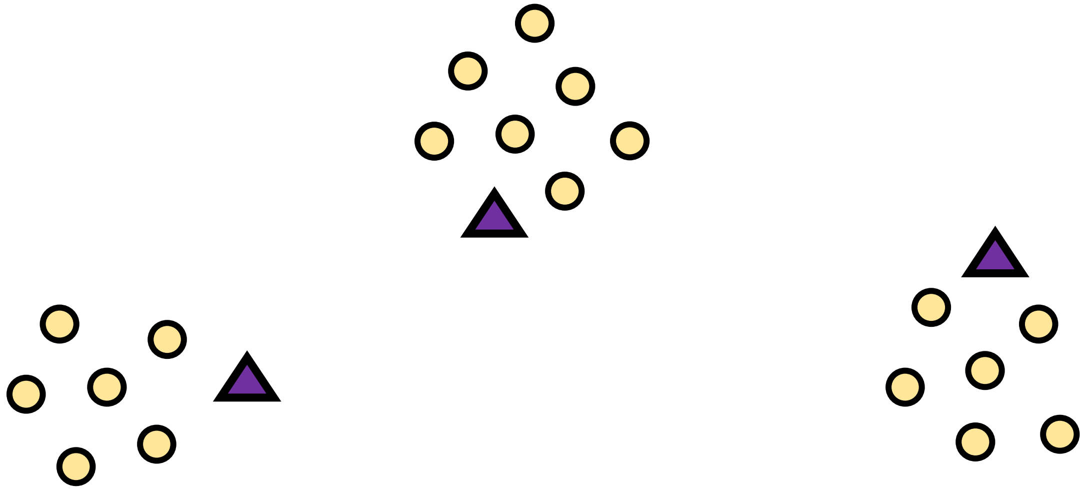
Last Time: Importance Sampling for Sum Estimation

- Suppose $x_1, \ldots, x_n \in [1, n]$
- Suppose $p_i = \frac{x_i}{x}$ for all $i \in [n]$

- Can get a 2-approximation for importance sampling
- How many samples do we expect? $\frac{x_1}{x} + \cdots + \frac{x_n}{x} = 1$, so just a constant number of samples!

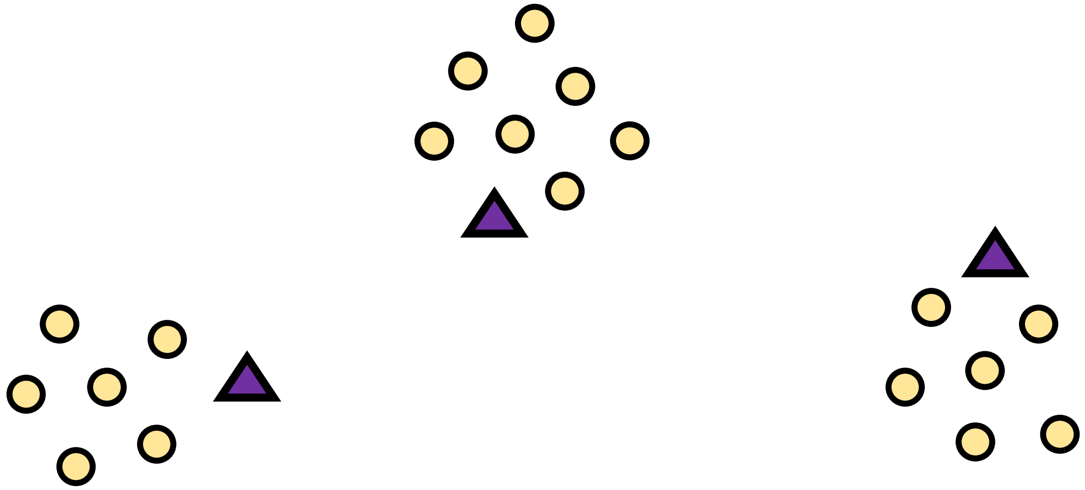
Coreset Construction and Sampling

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



Coreset Construction and Sampling

- Consider a fixed set X and a fixed set C of k centers, which induces a fixed cost $\mathrm{Cost}(X, C)$
- A simple way to obtain X' with $\mathrm{Cost}(X', C) \approx \mathrm{Cost}(X, C)$ is to uniformly sample points of X into X'



Coreset Construction and Uniform Sampling

- Consider a fixed set X and a fixed set C of k centers, which induces a fixed cost $\text{Cost}(X, C)$
- Suppose all points have the same cost, $\text{Cost}(x, C) = \frac{\text{Cost}(X,C)}{n}$
- How many points do I need to sample to approximate $\text{Cost}(X, C)$ within a 2-factor?

Bernstein's Inequality

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

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

Bernstein's Inequality

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

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

- Set $M = \frac{1}{p}$, $t = \frac{1}{2} \cdot \mathrm{Cost}(X, C)$, and $\sigma^2 \approx \frac{n}{p}$. Then for $x = \mathrm{Cost}(X, C)$,

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

Coreset Construction and Uniform Sampling

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

- Suppose all points have the same cost, $\text{Cost}(x, C) = \frac{\text{Cost}(X,C)}{n}$
- Can get a 2-approximation to $\text{Cost}(X, C)$ even for $p = \Theta\left(\frac{1}{n}\right)$
- How many samples do we expect? $np = \Theta(1)$

Coreset Construction and Uniform Sampling

- Consider a fixed set X and a fixed set C of k centers, which induces a fixed cost $\text{Cost}(X, C)$
- Suppose all points have cost between 1 and 100
- Suppose $p_i = p$ for all $i \in [n]$

Bernstein's Inequality

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

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

Bernstein's Inequality

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

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

- Set $M = \frac{100}{p}$, $t = \frac{1}{2} \cdot \mathrm{Cost}(X, C)$, and $\sigma^2 \approx \frac{10000n}{p}$. Then for $x = \mathrm{Cost}(X, C)$,

$$\Pr\left[|y-\mu| \geq \frac{x}{2}\right] \leq 2\exp\left(-\frac{(x/2)^2}{2(100n/p)+(4/3)(50x/p)}\right)$$

Coreset Construction and Uniform Sampling

- Suppose $x_1, \ldots, x_n \in [1,100]$
- Suppose $p_i = p$ for all $i \in [n]$

- For $\Pr\left[|y - \mu| \geq \frac{x}{2}\right] \leq 2\exp\left(-\frac{(x/2)^2}{2(10000n/p)+(4/3)(100x/p)}\right)$, we require $\frac{20000n}{p} \approx \left(\frac{x}{2}\right)^2$ and x can be as small as n, so we need $p \approx \frac{80000}{n}$

Coreset Construction and Uniform Sampling

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

- Suppose all points have cost between 1 and 100
- Can get a 2-approximation even for $p \approx \frac{80000}{n}$
- How many samples do we expect? np is now WAY larger

Coreset Construction and Uniform Sampling

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

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

- Set $M = \frac{n}{p}$, $t = \frac{1}{2} \cdot \text{Cost}(X, C)$, and $\sigma^2 \approx \frac{n^3}{p}$. Then for $x = \text{Cost}(X, C)$,

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

Coreset Construction and Uniform Sampling

- Consider a fixed set X and a fixed set C of k centers, which induces a fixed cost $\text{Cost}(X, C)$
- Suppose all points have cost between 1 and 100
- Suppose $p_i = p$ for all $i \in [n]$
- For $\Pr\left[|y - \mu| \geq \frac{x}{2}\right] \leq 2\exp\left(-\frac{(x/2)^2}{2(10000n/p)+(4/3)(100x/p)}\right)$, we require $\frac{20000n}{p} \approx \left(\frac{x}{2}\right)^2$ and x can be as small as n, so $p \approx \frac{80000}{n}$

Coreset Construction and Uniform Sampling

- Suppose $p_i = p$ for all $i \in [n]$

- What can we say about concentration?
- Can get a 2-approximation even for $p \approx \frac{80000}{n}$
- How many samples do we expect? np is now WAY larger

Coreset Construction and Uniform Sampling

- Consider a fixed set X and a fixed set C of k centers, which induces a fixed cost $\text{Cost}(X, C)$
- Suppose all points have cost between 1 and n
- How many points do I need to sample to approximate $\text{Cost}(X, C)$ within a $(1 + \varepsilon)$-factor?

Coreset Construction and Uniform Sampling

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

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

- Set $M = \frac{n}{p}$, $t = \frac{x}{2}$, and $\sigma^2 \approx \frac{n^2}{p}$. Then

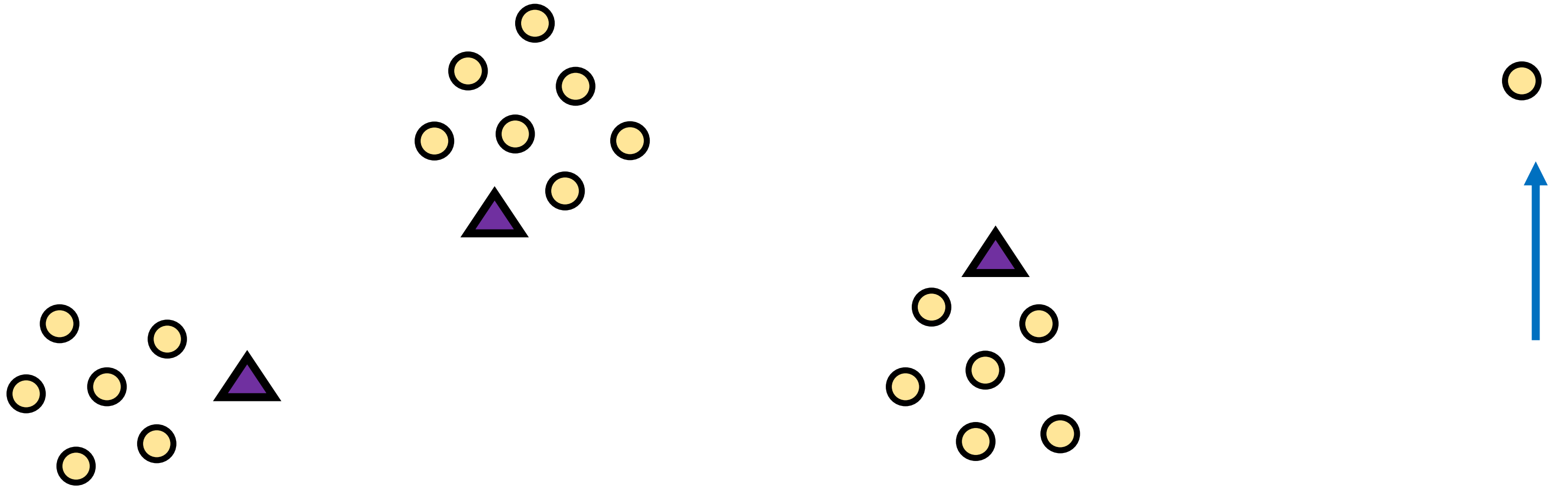
$$\Pr\left[|y - \mu| \geq \frac{x}{2}\right] \leq 2\exp\left(-\frac{(x/2)^2}{2(n^2/p) + (4/3)(nx/2p)}\right)$$

Uniform Sampling for Sum Estimation

- Consider a fixed set X and a fixed set C of k centers, which induces a fixed cost $\text{Cost}(X, C)$
- Suppose all points have cost between 1 and n
- Suppose $p_i = p$ for all $i \in [n]$
- For $\Pr\left[|y - \mu| \geq \frac{x}{2}\right] \leq 2\exp\left(-\frac{(x/2)^2}{2(n^2/p)+(4/3)(nx/2p)}\right)$, we require $\frac{2n^2}{p} \approx \left(\frac{x}{2}\right)^2$ and x can be as small as n, so we need $p \approx 1$

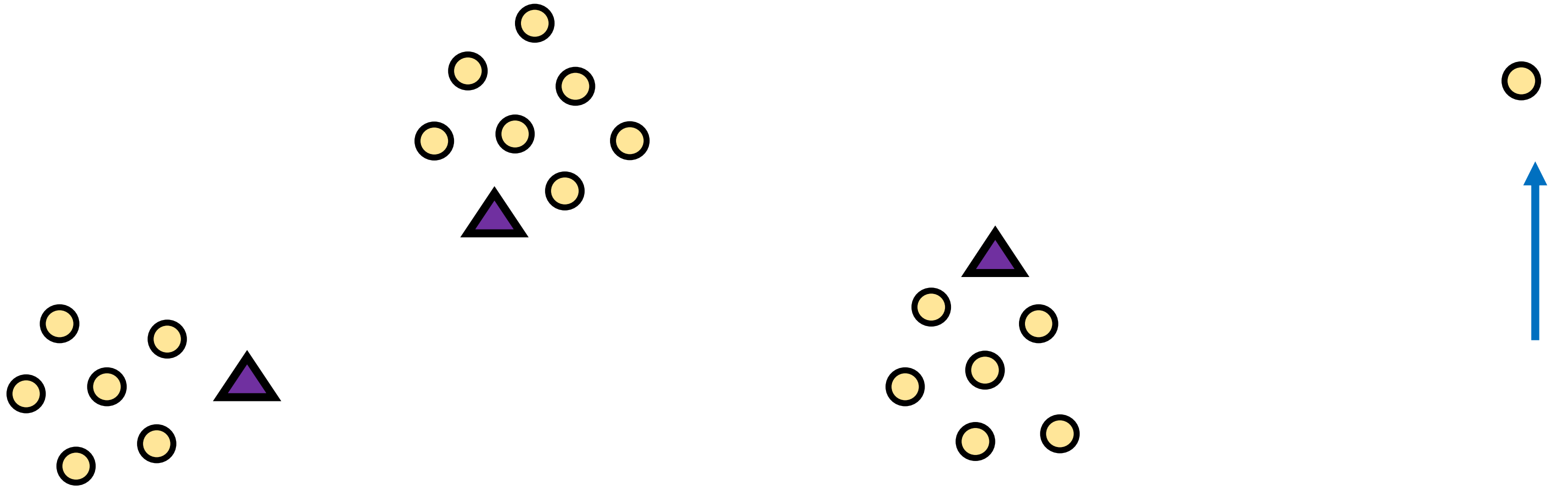
Coreset Construction and Sampling

- Consider a fixed set X and a fixed set C of k centers, which induces a fixed cost $\mathrm{Cost}(X, C)$
- Uniform sampling needs a lot of samples if there is a single point that greatly contributes to $\mathrm{Cost}(X, C)$



Coreset Construction and Sampling

- Fix: Importance sampling, sample each point $x \in X$ into X' with probability proportional $\mathrm{Cost}(x, C)$, i.e., $\mathrm{Cost}(x, C)/\mathrm{Cost}(X, C)$



Coreset Construction and Sampling

- Fix: Importance sampling, sample each point $x \in X$ into X' with probability proportional $\mathrm{Cost}(x, C)$, i.e., $\mathrm{Cost}(x, C)/\mathrm{Cost}(X, C)$

Importance Sampling for Coreset Construction

- What is the variance for each y_i?

- $\mathrm{Var}[y_i] \leq \frac{1}{p_i} \cdot \left(\mathrm{Cost}(x_i, C)\right)^2 \leq \mathrm{Cost}(x_i, C) \cdot \mathrm{Cost}(X, C)$
- $\mathrm{Var}[y] = \mathrm{Var}[y_1] + \cdots + \mathrm{Var}[y_n] \leq \left(\mathrm{Cost}(X, C)\right)^2$

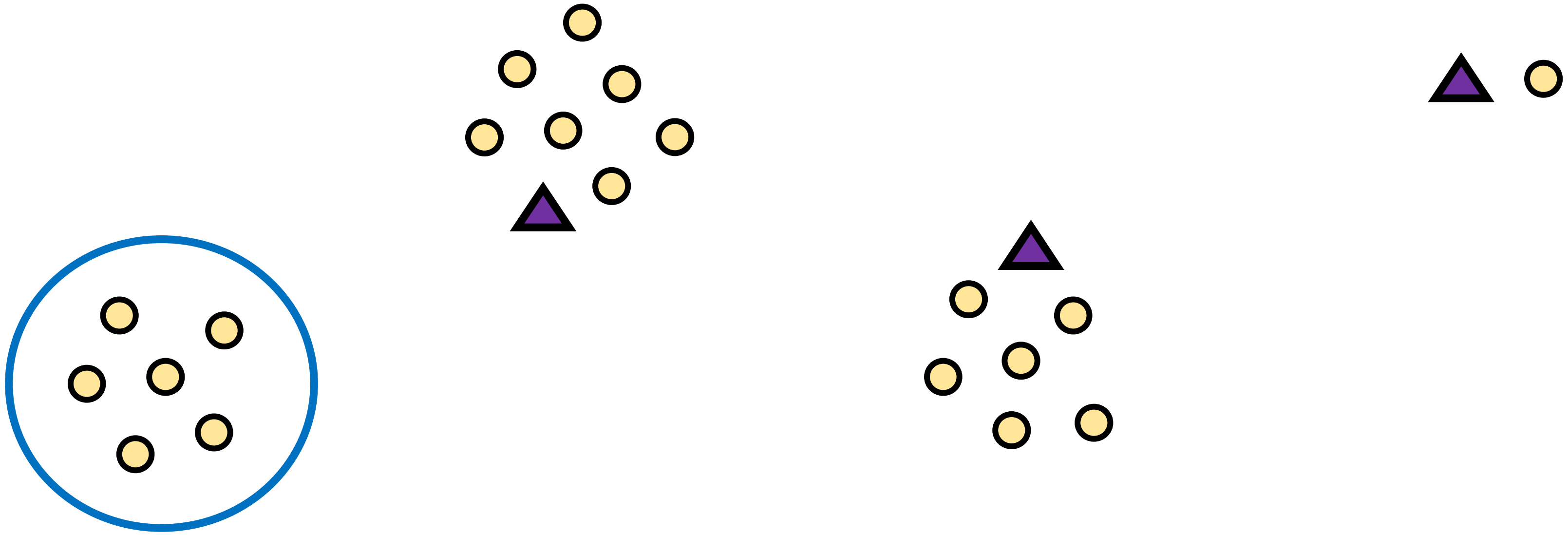
Coreset Construction and Sampling

- Fix: Importance sampling, sample each point $x \in X$ into X' with probability proportional $\mathrm{Cost}(x, C)$, i.e., $\mathrm{Cost}(x, C)/\mathrm{Cost}(X, C)$

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

Coreset Construction and Sampling

- Importance sampling only needs X' to have size $O\left(\frac{1}{\varepsilon^2}\right)$ to achieve $(1+\varepsilon)$-approximation to $\mathrm{Cost}(X, C)$
- What about a different choice C of k centers?



Coreset Construction and Sampling

- Importance sampling only needs X' to have size $O\left(\frac{1}{\varepsilon^2}\right)$ 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

Nets

- A net N is a set of sets C of k centers such that accuracy on N implies accuracy everywhere

Coreset Construction and Sampling

- Importance sampling only needs X' to have size $O\left(\frac{1}{\varepsilon^2}\right)$ 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

Net with size $\left(\frac{n\Delta}{\varepsilon}\right)^{O(kd)}$

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)