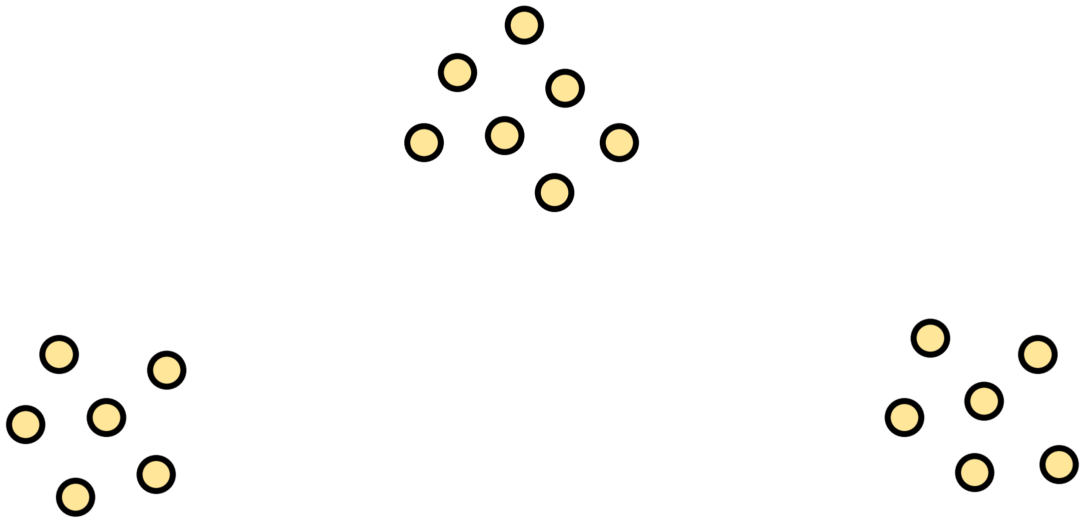
CSCE 658: Randomized Algorithms

Lecture 12

Samson Zhou

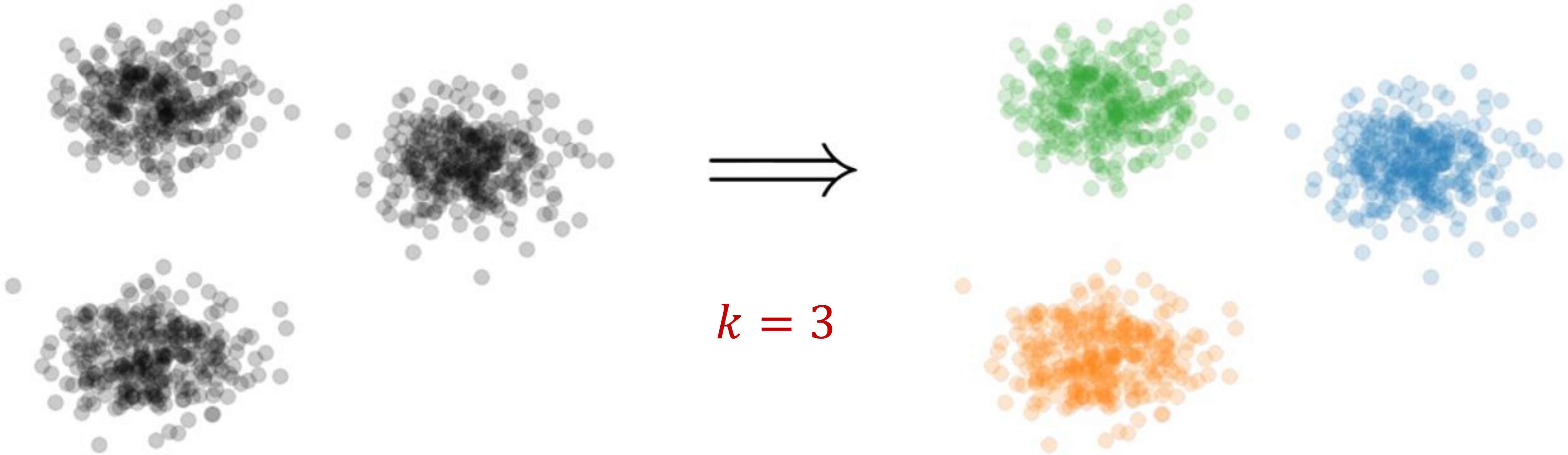
Class Logistics

- March 5: Lecture canceled, i.e., do NOT show up to HRBB 126 (unless you want to see an empty classroom)



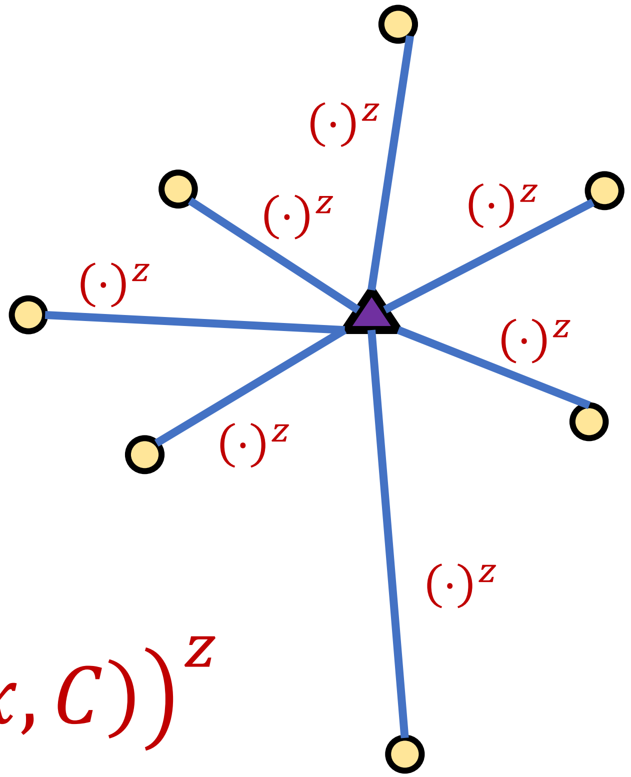
Last Time: k-Clustering

- Goal: Given input dataset X, partition X so that “similar” points are in the same cluster and “different” points are in different clusters
- There can be at most k different clusters



Last Time: k-Clustering

- Define clustering cost $\mathrm{Cost}(X, C)$ to be a function of $\{\mathrm{dist}(x, C)\}_{x \in C}$
- k-center: $\mathrm{Cost}(X, C) = \max_{x \in X} \mathrm{dist}(x, C)$
- k-median: $\mathrm{Cost}(X, C) = \sum_{x \in X} \mathrm{dist}(x, C)$
- k-means: $\mathrm{Cost}(X, C) = \sum_{x \in X} \left(\mathrm{dist}(x, C)\right)^2$
- (k, z)-clustering: $\mathrm{Cost}(X, C) = \sum_{x \in X} \left(\mathrm{dist}(x, C)\right)^z$



Last Time: (k, z)-Clustering in the Streaming Model

- Merge-and-reduce framework

- Suppose there exists a $(1+\varepsilon)$-coreset construction for (k, z)-clustering that uses $f\left(k, \frac{1}{\varepsilon}\right)$ weighted input points

$\boxed{\tilde{O}\left(\frac{k^2}{\varepsilon^2}\right)}$

- Partition the stream into blocks containing $f\left(k, \frac{\log n}{\varepsilon}\right)$ points

Last Time: (k,z)-Clustering in the Streaming Model

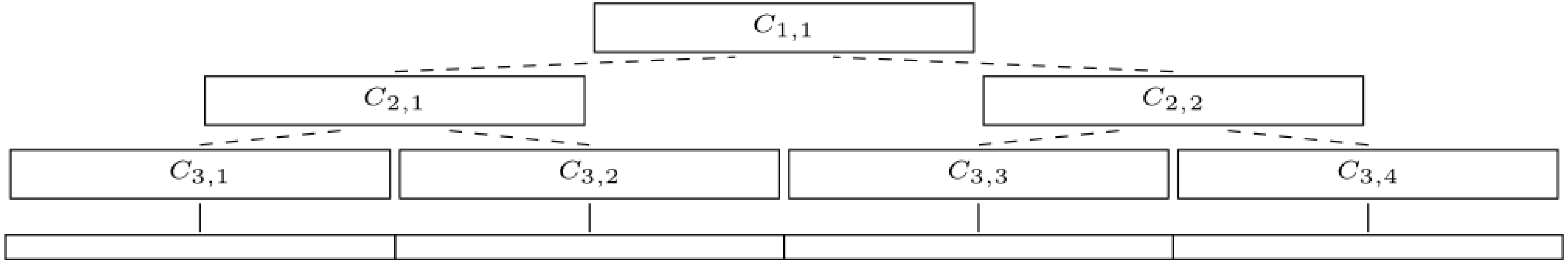
- Partition the stream into blocks containing $f\left(k, \frac{\log n}{\varepsilon}\right)$ points
- Create a $\left(1+\frac{\varepsilon}{\log n}\right)$-coreset for each block
- Create a $\left(1+\frac{\varepsilon}{\log n}\right)$-coreset for the set of points formed by the union of two coresets for each block

Reduce

Merge

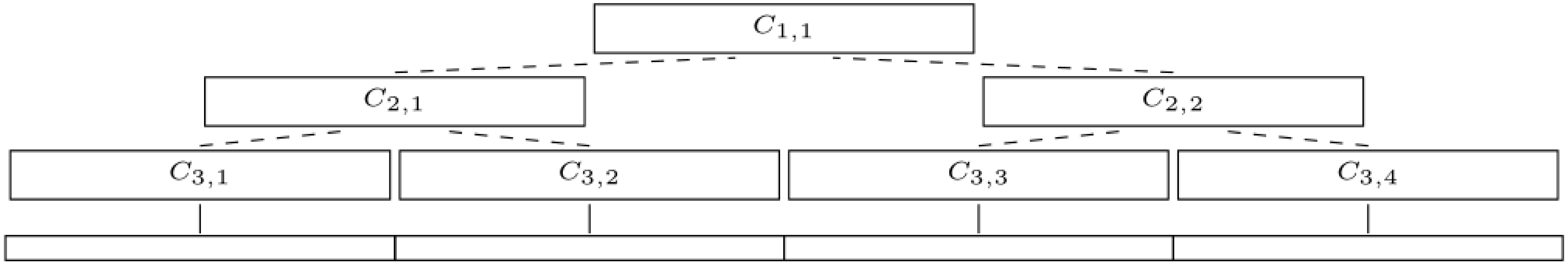
Last Time: (k, z)-Clustering in the Streaming Model

- Partition the stream into blocks containing $f\left(k, \frac{\log n}{\varepsilon}\right)$ points
- Create a $\left(1 + \frac{\varepsilon}{\log n}\right)$-coreset for each block
- Create a $\left(1 + \frac{\varepsilon}{\log n}\right)$-coreset for the set of points formed by the union of two coresets for each block



Last Time: (k, z)-Clustering in the Streaming Model

- There are $O(\log n)$ levels
- Each coreset is a $\left(1 + \frac{\varepsilon}{\log n}\right)$-coreset of two coresets
- Total approximation is $\left(1 + \frac{\varepsilon}{\log n}\right)^{\log n} = (1 + O(\varepsilon))$



Previously: Bernstein's Inequality

- Bernstein's inequality: Let $X_1, \dots, X_n \in [-M, M]$ 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

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

Sampling for Sum Estimation

- Consider a fixed set $X = \{x_1, \dots, x_n\}$ 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?

Sampling for Sum Estimation

- Let y_i be the contribution of the sample corresponding to x_i

- $y_i = 0$ with probability $1 - p_i$
- $y_i = \frac{1}{p_i} \cdot x_i$ with probability p_i
- $\mathrm{E}[y_i] = x_i$
- $\mathrm{E}[y_1 + \cdots + y_n] = x_1 + \ \ldots + x_n$

Sampling for Sum Estimation

- Consider a fixed set $X = \{x_1, \dots, x_n\}$ 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? $\mathrm{E}[y_1 + \cdots + y_n] = x_1 + \ \dots + x_n$

Sampling for Sum Estimation

- Consider a fixed set $X = \{x_1, \dots, x_n\}$ 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?

Uniform 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 $\frac{1}{p_i}$
- Suppose $p_i = p$ for all $i \in [n]$
- What can we say about concentration?

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 we get a 2-approximation with high probability?

Uniform Sampling for Sum Estimation

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

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

Uniform Sampling for Sum Estimation

- 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 $M = \frac{1}{p}$, $t = \frac{n}{2}$, and $\sigma^2 = \frac{n}{p}$. Then

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

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 $p = \Theta\left(\frac{1}{n}\right)$

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?

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? $np = \Theta(1)$

Uniform Sampling for Sum Estimation

- Suppose $x_1, \ldots, x_n \in [1,2]$
- Suppose $p_i = p$ for all $i \in [n]$

- Can we get a 2-approximation with high probability?

Uniform Sampling for Sum Estimation

- 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 $M = \frac{2}{p}$, $t = \frac{x}{2}$, and $\sigma^2 \approx \frac{4n}{p}$. Then

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

Uniform Sampling for Sum Estimation

- Suppose $x_1, \ldots, x_n \in [1,2]$
- Suppose $p_i = p$ for all $i \in [n]$

- For $\Pr\left[|y - \mu| \geq \frac{x}{2}\right] \leq 2\exp\left(-\frac{(x/2)^2}{2(4n/p)+(4/3)(x/p)}\right)$, we require $\frac{8n}{p} \approx \left(\frac{x}{2}\right)^2$ and x can be as small as n, so $p \approx \frac{2}{n}$

Uniform Sampling for Sum Estimation

- Suppose $x_1, \dots, x_n \in [1,2]$
- Suppose $p_i = p$ for all $i \in [n]$

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

Uniform Sampling for Sum Estimation

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

- Can we get a 2-approximation with high probability?

Uniform Sampling for Sum Estimation

- Bernstein's inequality: Let $y_1, \dots, y_n \in [-M, M]$ 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 $M = \frac{100}{p}$, $t = \frac{x}{2}$, and $\sigma^2 \approx \frac{10000n}{p}$. Then

$$\Pr\left[|y - \mu| \geq \frac{x}{2}\right] \leq 2\exp\left(-\frac{(x/2)^2}{2(10000n/p) + (4/3)(100x/p)}\right)$$

Uniform Sampling for Sum Estimation

- Suppose $x_1, \dots, x_n \in [1,100]$
- 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 $p \approx \frac{80000}{n}$

Uniform Sampling for Sum Estimation

- Suppose $x_1, \ldots, x_n \in [1,100]$
- 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

Uniform Sampling for Sum Estimation

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

- Can we get a 2-approximation with high probability?

Uniform Sampling for Sum Estimation

- 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

- Suppose $x_1, \ldots, x_n \in [1, n]$
- 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 $p \approx 1$

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?

Uniform Sampling for Sum Estimation

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

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

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

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?

Sampling for Sum Estimation

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

- What if we choose the probability p_i different for each x_i?

Sampling for Sum Estimation

- Consider a fixed set $X = \{x_1, \ldots, x_n\}$ of n numbers
- What if we choose the probability p_i different for each x_i?
- Choose p_i proportional to x_i

Importance Sampling for Sum Estimation

- Consider a fixed set $X = \{x_1, \dots, x_n\}$ of n numbers
- What if we choose the probability p_i different for each x_i?
- Choose p_i proportional to x_i
- Let $x = x_1 + \cdots + x_n$, set $p_i = \frac{x_i}{x}$

Importance Sampling for Sum Estimation

- Bernstein's inequality: Let $y_1, \dots, y_n \in [-M, M]$ 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 $t = \frac{x}{2}$. What about M and σ^2?

Importance Sampling for Sum Estimation

- $y_i \leq \frac{1}{p} \cdot x_i = \frac{x}{x_i} \cdot x_i = x$
- Can set $M = x$ in Bernstein's inequality

Importance Sampling for Sum Estimation

- What is the variance for each y_i?

- $\mathrm{Var}[y_i] \leq \frac{1}{p_i} \cdot x_i^2 \leq x_i \cdot x$
- $\mathrm{Var}[y] = \mathrm{Var}[y_1] + \cdots + \mathrm{Var}[y_n] \leq x \cdot (x_1 + \cdots + x_n) = x^2$
- What is the variance for y under uniform sampling? $\frac{nx_i^2}{p}$
- What is the variance for each y_i under uniform sampling? $\frac{x_i^2}{p}$

Importance Sampling for Sum Estimation

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

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

Importance Sampling for Sum Estimation

- Suppose $x_1, \dots, x_n \in [1, n]$
- 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?

Importance Sampling for Sum Estimation

- How many samples do we expect?
- Let S_i be the indicator random variable for whether we sampled x_i (which we do with probability p_i)

- $S = S_1 + \cdots + S_n$ is the total number of samples
- $\mathrm{E}[S] = \mathrm{E}[S_1] + \cdots + \mathrm{E}[S_n]$ by linearity of expectation
- $\mathrm{E}[S_i] = p_i = \frac{x_i}{x}$

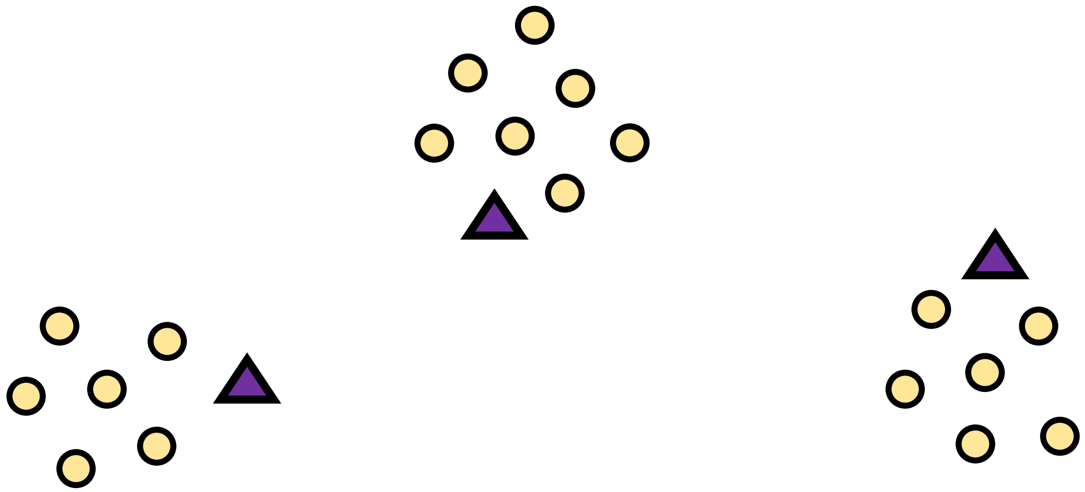
Importance Sampling for Sum Estimation

- Suppose $x_1, \dots, x_n \in [1, n]$
- 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? $\mathrm{E}[S] = \frac{x_1}{x} + \cdots + \frac{x_n}{x} = 1$, so just a constant number of samples!

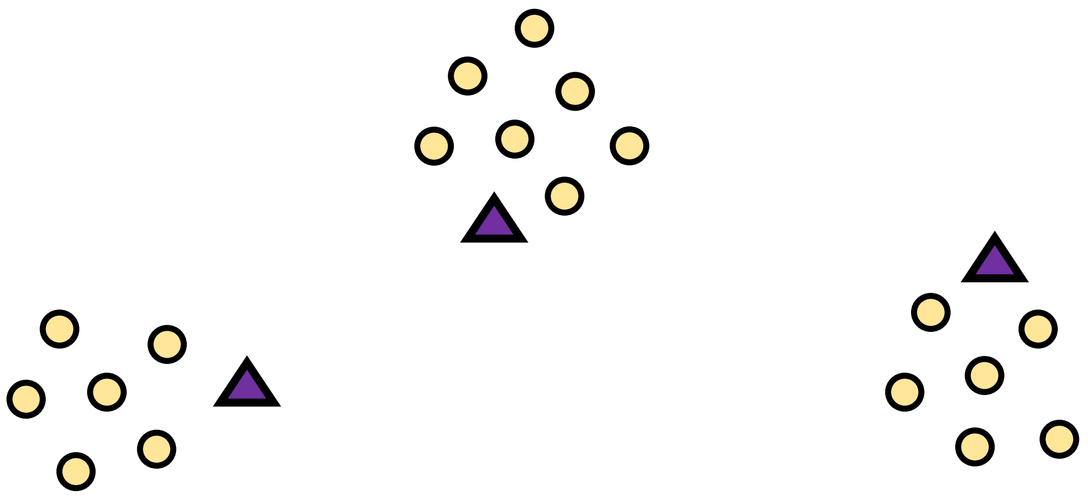
Coreset Construction and Sampling

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



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 $\mathrm{Cost}(X, C)$

- Suppose all points have the same cost, $\mathrm{Cost}(x, C) = \frac{\mathrm{Cost}(X,C)}{n}$

- How many points do I need to sample to approximate $\mathrm{Cost}(X, C)$ 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 $y_1, \dots, y_n \in [-M, M]$ 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 $\mathrm{Cost}(X, C)$

- Suppose all points have the same cost, $\mathrm{Cost}(x, C) = \frac{\mathrm{Cost}(X,C)}{n}$
- Can get a 2-approximation to $\mathrm{Cost}(X, C)$ 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 $\mathrm{Cost}(X, C)$
- 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 $M = \frac{n}{p}$, $t = \frac{1}{2} \cdot \mathrm{Cost}(X, C)$, and $\sigma^2 \approx \frac{n^3}{p}$. Then for $x = \mathrm{Cost}(X, C)$,

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

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

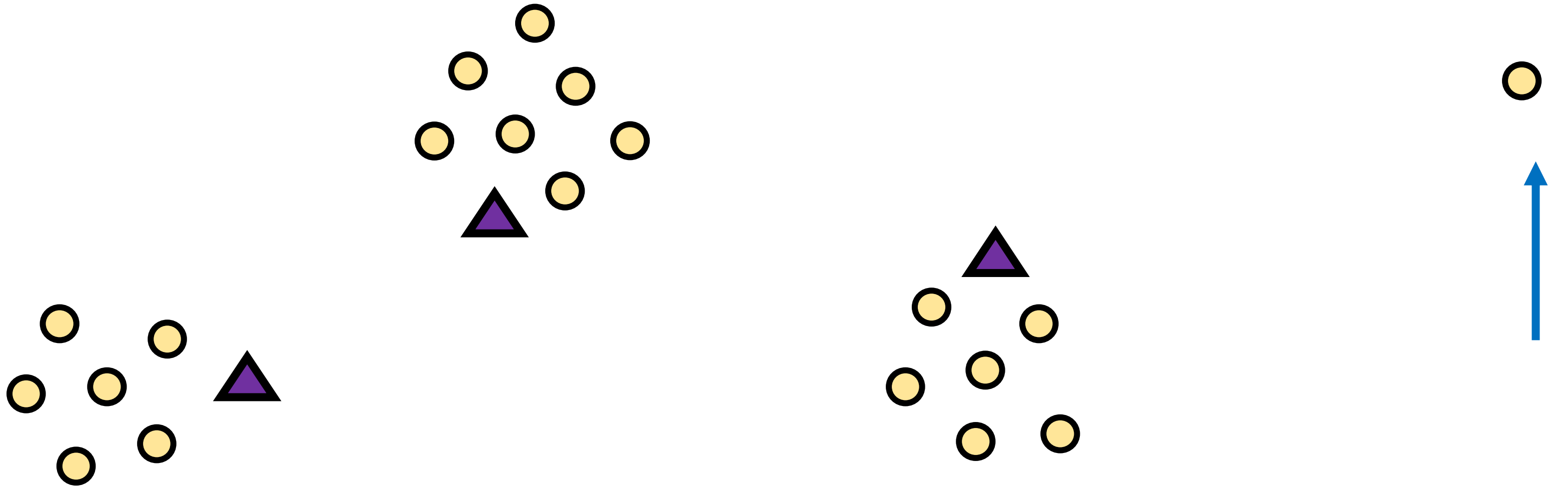
$$\Pr\left[|y - \mu| \geq \frac{x}{2}\right] \leq 2\exp\left(-\frac{(x/2)^2}{2(n^2/p) + (4/3)(nx/2p)}\right)$$

Uniform Sampling for Sum Estimation

- Consider a fixed set X and a fixed set C of k centers, which induces a fixed cost $\mathrm{Cost}(X, C)$
- 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 $p \approx 1$

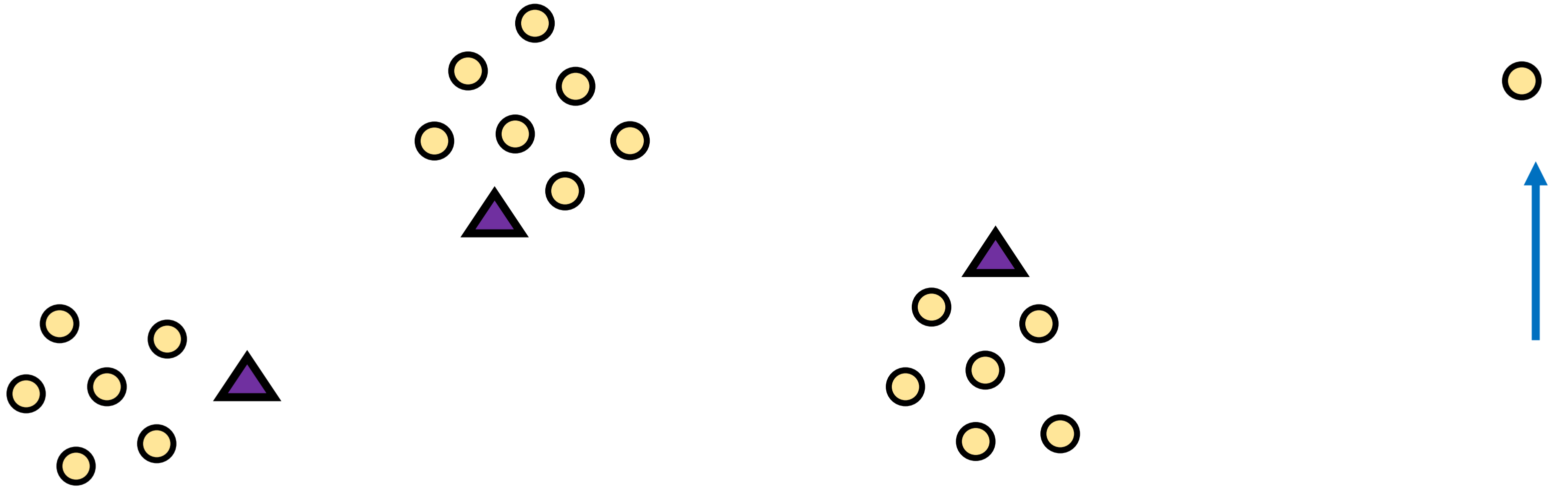
Coreset Construction and Sampling

- 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?

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

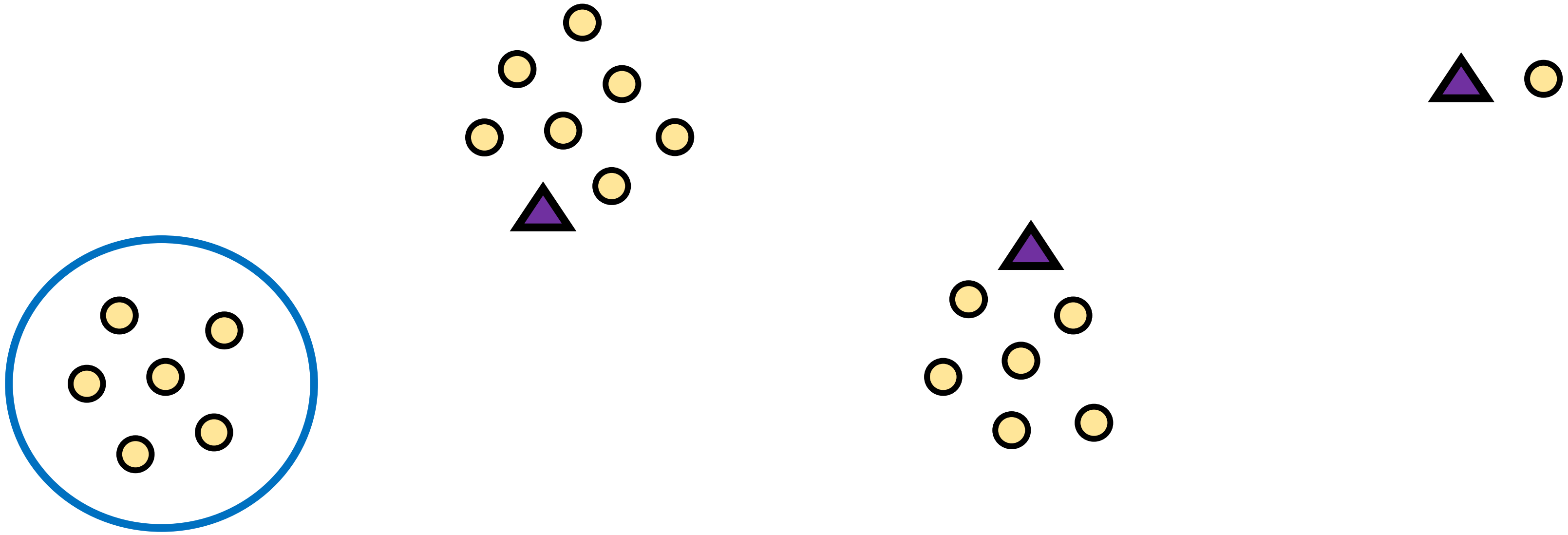
Coreset Construction and Sampling

- 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)