CSCE 689: Special Topics in Modern Algorithms for Data Science Fall 2023

Lecture 23 — October 25, 2023

Prof. Samson Zhou *Scribe: Lipai Huang*

Recall

Theorem 1 (Bernstein's Inequality)**.** *Let $y_1, ..., y_n \in [-M, M]$ be independent random variables and let $y = y_1 + ... + y_n$ have mean μ and variance σ^2. Then for any $t \geq 0$:*

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

Sampling for Sum Estimation

Goal: Approximate the original sum by the sum of the rescaled sampled numbers.
Consider a fixed set $X = \{x_1, ..., x_n\}$ of n numbers, and suppose we sample each point x_i with some probability p_i and rescale by $\frac{1}{p_i}$, also we let y_i be the contribution of the sample corresponding to x_i, we have:

- $y_i = 0$ with probability $1 - p_i$
- $y_i = \frac{1}{p_i} \cdot x_i$ with probability p_i
- $\mathbb{E}[y_i] = x_i$
- Expected sum: $\mathbb{E}[y_1 + ... + y_n] = x_1 + ... + x_n$

Uniform Sampling

Suppose $p_i = p$ for all $i \in [n]$, given Theorem 1, if $x_1 = ... = X_n = 1$, set $M = \frac{1}{p}$, $t = \frac{n}{2}$, and $\sigma^2 = \frac{n}{p}$, then:

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

We require $2(\frac{n}{p}) \approx (\frac{n}{2})^2$, so we need $p = \Theta(\frac{1}{n})$.
If $x_1, ..., x_n \in [1, 2]$, set $M = \frac{2}{p}$, $t = \frac{x}{2}$ and $\sigma^2 \approx \frac{4n}{p}$, then:

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

We require $2(\frac{4n}{p}) \approx \frac{x}{2}^2$ and x can be as small as n, so $p \approx \frac{2}{n}$.
If $x_1, ..., x_n \in [1, 100]$, set $M = \frac{100}{p}$, $t = \frac{x}{2}$ and $\sigma^2 \approx \frac{10000n}{p}$, then:

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

We require $2(\frac{10000n}{p}) \approx \frac{x}{2}^2$ and x can be as small as n, so $p \approx \frac{80000}{n}$.
If $x_1, ..., x_n \in [1, n]$, set $M = \frac{n}{p}$, $t = \frac{x}{2}$ and $\sigma^2 \approx \frac{n^2}{p}$, then:

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

We require $2(\frac{n^2}{p}) \approx \frac{x}{2}^2$ and x can be as small as n, so we need $p \approx 1$.
To sum up:

Domain of $x_1, ..., x_n$	Conditions of p for 2-approximation	Expected samples amount
$x_1 = ... = x_n = 1$	$p = \Theta(\frac{1}{n})$	$np = \Theta(1)$
$x_1, ..., x_n \in [1, 2]$	$p \approx \frac{2}{n}$	slightly larger np
$x_1, ..., x_n \in [1, 100]$	$p \approx \frac{80000}{n}$	way larger np
$x_1, ..., x_n \in [1, n]$	$p \approx 1$	n

Table 1: Uniform Sampling Summary

Importance Sampling

Suppose $x = x_1 + ... + x_n$, since p_i is chosen proportionally to x_i, let $p_i = \frac{x_i}{x}$, we have $y_i \leq \frac{1}{p} \cdot x_i = \frac{x}{x_i} \cdot x_i = x$, thus:

- $\mathrm{Var}[y_i] \leq \frac{1}{p_i} x_i^2 \leq x_i \cdot x$
- $\mathrm{Var}[y] = \mathrm{Var}[y_1] + ... + \mathrm{Var}[y_n] \leq x \cdot (x_1 + ... + x_n) = x^2$

Given Theorem 1, set $M = x$, $t = \frac{x}{2}$, and $\sigma^2 \approx x^2$. Then

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

Suppose $x_1, ..., x_n \in [1, n]$, we can get a 2-approximation and based on $\frac{x_1}{x} + ... + \frac{x_n}{x} = 1$, we expect a constant number of samples.