Simple Analysis of Priority Sampling

New York University, Majid Daliri

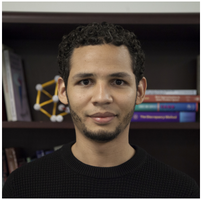
January 9, 2024

Collaboration

This project was a joint effort by the following individuals:

Prof. Juliana Freire
New York University

Prof. Christopher Musco
New York University



Aécio Santos
New York University

Haoxiang Zhang
New York University

Motivation: Priority Sampling Problem

Problem Overview

Consider a set of items, labeled from 1 to n, with each item i having an associated **positive** weight w_i.

Specific Query

Given a subset $\mathbf{Q}$ of $\{1, 2, ..., n\}$, the query asks: *"What is the total sum of weights in* $\mathbf{w}$ *corresponding to the elements in* $\mathbf{Q}$*?"*

$$\sum_{i=1}^{n} w_i \cdot \mathbb{1}\left[i \in Q\right]$$

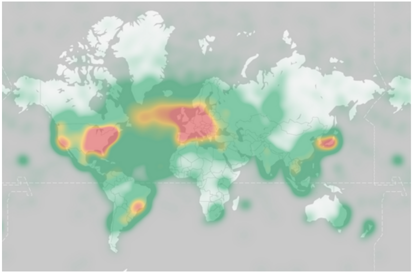
Significance

This problem is fundamental in data analysis, where efficient and accurate estimations of such sums are crucial, especially in large datasets.

Motivating Example: Website Traffic Analysis

Example:

Consider a scenario where we want to analyze the website traffic, specifically counting the number of user count accessing the site from a certain region.



Motivating Example: Website Traffic Analysis

Example:

Consider a scenario where we aim to analyze website traffic, specifically by counting the number of users accessing the site from a certain region.

Analysis Goals

- Users are labeled as items 1 to n.
- Let w_i represent the number of visits by user i.
- The goal is to compute the total number of visits to the website from users in Washington D.C.
- Define Q as the set of users i who are located in Washington D.C.
- $\sum_{i=1}^{n} w_i \cdot \mathbb{1}[i \in Q]$

Sampling

Objective

Because we are dealing with large-scale datasets, our aim is to store only a limited number ($k << n$) of items from the data structure.

Sampling

Objective

Because we are dealing with large-scale datasets, our aim is to store only a limited number ($k << n$) of items from the data structure.

Sampling Strategy

This approach assigns each item a sampling ratio of p_i and samples each item i with probability p_i.

$$\sum_{i=1}^{n} p_i \approx k$$

Sampling

Objective

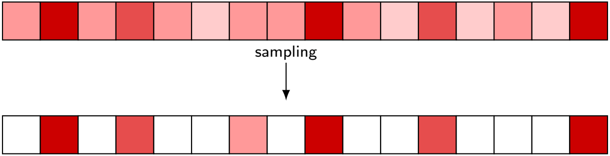
Because we are dealing with large-scale datasets, our aim is to store only a limited number ($k << n$) of items from the data structure.

Sampling Strategy

This approach assigns each item a sampling ratio of p_i and samples each item i with probability p_i.

$$\sum_{i=1}^{n} p_i \approx k$$

To achieve more accurate sampling, items with higher weight should be sampled with a higher probability.



Formal Objectives for Effective Sampling

$$\hat{w}_i = \begin{cases} \frac{w_i}{p_i} & \text{sampled with probability } p_i, \\ 0 & \text{otherwise} \end{cases}.$$

Formal Objectives for Effective Sampling

$$\hat{w}_i = \begin{cases} \frac{w_i}{p_i} & \text{sampled with probability } p_i, \\ 0 & \text{otherwise} \end{cases}.$$

Then, for any distribution of p_i, the estimator

$$\mathbb{E}\left[\sum_{i=1}^{n} \hat{w}_i \cdot \mathbb{1}[i \in Q]\right] = \sum_{i=1}^{n} w_i \cdot \mathbb{1}[i \in Q]$$

Formal Objectives for Effective Sampling

$$\hat{w}_i = \begin{cases} \frac{w_i}{p_i} & \text{sampled with probability } p_i, \\ 0 & \text{otherwise} \end{cases}.$$

Then, for any distribution of p_i, the estimator

$$\mathbb{E}\left[\sum_{i=1}^{n} \hat{w}_i \cdot \mathbb{1}[i \in Q]\right] = \sum_{i=1}^{n} w_i \cdot \mathbb{1}[i \in Q]$$

In our ideal case, to minimize the variance of the estimator when sampling k items, if $\sum p_i = k$, we aim to minimize

$$\min_{p_1,\ldots,p_n} \text{Var}\left[\sum_{i=1}^{n} \hat{w}_i \cdot \mathbb{1}[i \in Q]\right]$$

Formal Objectives for Effective Sampling

$$\hat{w}_i = \begin{cases} \frac{w_i}{p_i} & \text{sampled with probability } p_i, \\ 0 & \text{otherwise} \end{cases}.$$

Then, for any distribution of p_i, the estimator

$$\mathbb{E}\left[\sum_{i=1}^{n} \hat{w}_i \cdot \mathbb{1}[i \in Q]\right] = \sum_{i=1}^{n} w_i \cdot \mathbb{1}[i \in Q]$$

In our ideal case, to minimize the variance of the estimator when sampling k items, if $\sum p_i = k$, we aim to minimize

$$\min_{p_1,\ldots,p_n} \sum_{i=1}^{n} \mathrm{Var}\left[\hat{w}_i \cdot \mathbb{1}[i \in Q]\right]$$

Formal Objectives for Effective Sampling

$$\hat{w}_i = \begin{cases} \frac{w_i}{p_i} & \text{sampled with probability } p_i, \\ 0 & \text{otherwise} \end{cases}.$$

Then, for any distribution of p_i, the estimator

$$\mathbb{E}\left[\sum_{i=1}^{n} \hat{w}_i \cdot \mathbb{1}[i \in Q]\right] = \sum_{i=1}^{n} w_i \cdot \mathbb{1}[i \in Q]$$

In our ideal case, to minimize the variance of the estimator when sampling k items, if $\sum p_i = k$, we aim to minimize

$$\min_{p_1,\ldots,p_n} \sum_{i=1}^{n} \mathrm{Var}\left[\hat{w}_i \cdot \mathbb{1}[i \in Q]\right]$$

Given that Q is unknown at the time of building the data structure and we cannot speculate about the query, our goal is to minimize

$$\min_{p_1,\ldots,p_n} \sum_{i=1}^{n} \mathrm{Var}[\hat{w}_i]$$

Threshold Sampling: Sampling WOR

Sampling Criterion

Consider a hashing function $h : \{1, 2, \ldots, n\} \to (0, 1]$.
An item i is sampled if:

$$\frac{h_i}{w_i} < \tau \Rightarrow i \in \mathcal{S}$$

where h_i is the hash value of item i, w_i is its weight, and τ is a predefined threshold.

Threshold Sampling: Sampling WOR

Sampling Criterion

Consider a hashing function $h : \{1, 2, \ldots, n\} \to (0, 1]$.
An item i is sampled if:

$$h_i < w_i \cdot \tau \Rightarrow i \in \mathcal{S}$$

where h_i is the hash value of item i, w_i is its weight, and τ is a predefined threshold.

Threshold Sampling: Sampling WOR

Sampling Criterion

Consider a hashing function $h : \{1, 2, \ldots, n\} \to (0, 1]$.
An item i is sampled if:

$$h_i < w_i \cdot \tau \Rightarrow i \in \mathcal{S}$$

where h_i is the hash value of item i, w_i is its weight, and τ is a predefined threshold.

Probability of Sampling an item

The probability of sampling an item in this context is determined by the condition:

$$p_i = \min(1, \tau w_i)$$

Threshold Sampling: Sampling WOR

Sampling Criterion

Consider a hashing function $h : \{1, 2, \ldots, n\} \to (0, 1]$.
An item i is sampled if:

$$\frac{h_i}{w_i} < \tau \Rightarrow i \in \mathcal{S}$$

where h_i is the hash value of item i, w_i is its weight, and τ is a predefined threshold.

Setting the Threshold

Setting the threshold τ as $\frac{k}{\sum w_i}$, where k is a constant, results of sampling k items in the **expectation**.

$$\sum_{i=1}^{n} p_i = k$$

Threshold Sampling: Answer Queries

Definition of Estimated Weight

We define

$$\hat{w}_i = \begin{cases} \frac{w_i}{\min(1, w_i \tau)} & i \in \mathcal{S} \\ 0 & i \notin \mathcal{S} \end{cases}.$$

Threshold Sampling: Answer Queries

Definition of Estimated Weight

We define

$$\hat{w}_i = \begin{cases} \frac{w_i}{\min(1, w_i \tau)} & i \in \mathcal{S} \\ 0 & i \notin \mathcal{S} \end{cases}.$$

Query Answer

$$\sum_{i=1}^{n} \hat{w}_i \cdot \mathbb{1}\left[i \in Q\right]$$

Threshold Sampling: Variance Bound

Theorem (Variance Bound)

$$\mathrm{Var}\left[\hat{w}_i\right] \leq \frac{w_i}{\tau} = w_i \cdot \frac{W}{k}$$

Where $W = \sum_{i=1}^{n} w_i$

Threshold Sampling: Variance Bound

Theorem (Variance Bound)

$$\mathrm{Var}\left[\hat{w}_i\right] \leq \frac{w_i}{\tau} = w_i \cdot \frac{W}{k}$$

Where $W = \sum_{i=1}^{n} w_i$

Estimator Variance Objective

$$\sum_{i=1}^{n} \mathrm{Var}[\hat{w}_i] \leq \frac{W^2}{k}$$

Threshold Sampling: Variance Bound

Theorem (Variance Bound)

$$\mathrm{Var}\left[\hat{w}_i\right] \leq \frac{w_i}{\tau} = w_i \cdot \frac{W}{k}$$

Where $W = \sum_{i=1}^{n} w_i$

Estimator Variance Objective

$$\sum_{i=1}^{n} \mathrm{Var}[\hat{w}_i] \leq \frac{W^2}{k}$$

This bound is optimal among all choices of probabilities.

Threshold Sampling: Sampling WOR

Sampling Criterion

Consider a hashing function $h : \{1, 2, \ldots, n\} \rightarrow (0, 1]$.
An item i is sampled if:

$$\frac{h_i}{w_i} < \tau \Rightarrow i \in \mathcal{S}$$

where h_i is the hash value of item i, w_i is its weight, and τ is a predefined threshold.

Main Disadvantage

There is no **deterministic** upper limit on the number of items that are sampled.

Priority Sampling: Fixed-Size Sampling WOR

Challenge

The key challenge in sampling without replacement lies in consistently achieving a fixed number of samples

Priority Sampling : Fixed-Size Sampling WOR

Literature

Significant contributions in this area include:

- Introduction of Priority Sampling [Duffield, Lund, and Thorup, SIGMETRICS 2004]
- Upper Bound on Variance [Alon, Duffield, Lund, and Thorup, PODS 2005]
- Tight Upper Bound on Variance [Szegedy, STOC 2006]

Tight Variance Bound

$$\mathrm{Var}\left[\hat{w}_i\right] \leq w_i \cdot \frac{W}{k-1}$$

Where $W = \sum_{i=1}^{n} w_i$

Our Contribution

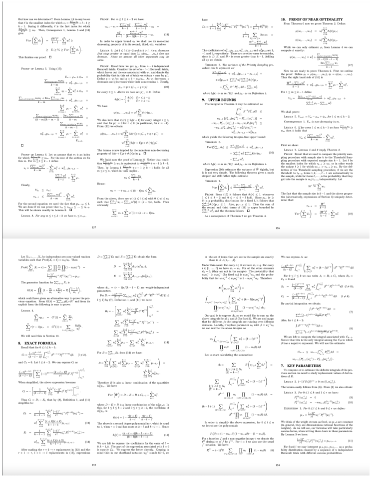


Figure: "The DLT priority sampling is essentially optimal", Szegedy, STOC 2006

Figure: Our Paper, SOSA 2024

Priority Sampling

Sampling Criterion

Consider a hashing function $h : \{1, 2, \ldots, n\} \to (0, 1]$.
define τ as $(k+1)^{\text{th}}$ smallest $\frac{h_i}{w_i}$.
An item i is sampled if:

$$\frac{h_i}{w_i} < \tau \Rightarrow i \in \mathcal{S}$$

where h_i is the hash value of item i, w_i is its weight.

Key Difference from Threshold Sampling

Priority Sampling is similar to Threshold Sampling, but with a crucial difference: the threshold τ is adaptively chosen as the $(k+1)^{\text{st}}$ smallest $\frac{h_i}{w_i}$.

Priority Sampling: Fixed-Size Sampling WOR

Sampling Criterion

Consider a hashing function $h : \{1, 2, \ldots, n\} \rightarrow (0, 1]$.
define τ as $(k+1)^{\text{th}}$ smallest $\frac{h_i}{w_i}$.
An item i is sampled if:

$$\frac{h_i}{w_i} < \tau \Rightarrow i \in \mathcal{S}$$

where h_i is the hash value of item i, w_i is its weight.

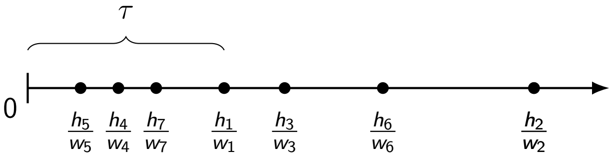


Figure: Illustration of Selecting τ for $k = 3$ in a Set of $n = 7$ Elements

Priority Sampling

Sampling Criterion

Consider a hashing function $h : \{1, 2, \ldots, n\} \rightarrow (0, 1]$.
define τ as $(k+1)^{\text{th}}$ smallest $\frac{h_i}{w_i}$.
An item i is sampled if:

$$\frac{h_i}{w_i} < \tau \Rightarrow i \in \mathcal{S}$$

where h_i is the hash value of item i, w_i is its weight.

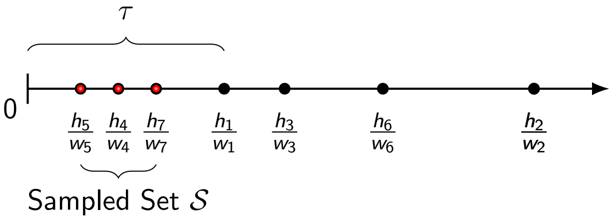


Figure: Illustration of the Sampling Process for Selected Elements

Priority Sampling: Answer Queries

Definition of Estimated Weight

We define

$$\hat{w}_i = \begin{cases} \frac{w_i}{\min(1, w_i \tau)} & i \in \mathcal{S} \\ 0 & i \notin \mathcal{S} \end{cases}.$$

Priority Sampling: Answer Queries

Definition of Estimated Weight

We define

$$\hat{w}_i = \begin{cases} \frac{w_i}{\min(1, w_i \tau)} & i \in \mathcal{S} \\ 0 & i \notin \mathcal{S} \end{cases}.$$

Fact (Expected Value)

$$\mathbb{E}[\hat{w}_i] = w_i$$

Fact (Pairwise Uncorrelated)

$$\mathbb{E}[\hat{w}_i \cdot \hat{w}_j] = w_i \cdot w_j$$

Priority Sampling: Answer Queries

Definition of Estimated Weight

We define

$$\hat{w}_i = \begin{cases} \frac{w_i}{\min(1, w_i \tau)} & i \in \mathcal{S} \\ 0 & i \notin \mathcal{S} \end{cases}.$$

Fact (Expected Value)

$$\mathbb{E}[\hat{w}_i] = w_i$$

Fact (Pairwise Uncorrelated)

$$\mathbb{E}[\hat{w}_i \cdot \hat{w}_j] = w_i \cdot w_j$$

Query Answer

$$\sum_{i=1}^{n} \hat{w}_i \cdot \mathbb{1}\left[i \in Q\right]$$

Priority Sampling: Variance Bound

Theorem (Variance Bound)

$$\mathrm{Var}\left[\hat{w}_i\right] \leq w_i \cdot \frac{W}{k-1}$$

Where $W = \sum_{i=1}^{n} w_i$

Priority Sampling: Variance Bound

Theorem (Variance Bound)

$$\mathrm{Var}\left[\hat{w}_i\right] \leq w_i \cdot \frac{W}{k-1}$$

Where $W = \sum_{i=1}^{n} w_i$

Estimator Variance Objective

$$\sum_{i=1}^{n} \mathrm{Var}[\hat{w}_i] \leq \frac{W^2}{k-1}$$

Priority Sampling: Variance Bound Proof Structure

Definition

Define τ_i for each i as k^{st} smallest $\frac{h_j}{w_j}$ for $j \in \{1, 2, \ldots, n\} - \{i\}$.

Priority Sampling: Variance Bound Proof Structure

Definition

Define τ_i for each i as k^{st} smallest $\frac{h_j}{w_j}$ for $j \in \{1, 2, \ldots, n\} - \{i\}$.

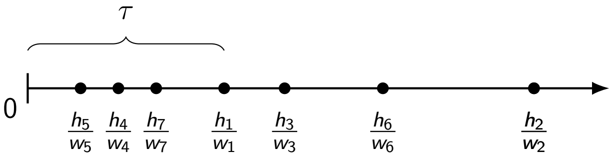


Figure: Illustration of Selecting τ for $k = 3$ in a Set of $n = 7$ Elements

Priority Sampling: Variance Bound Proof Structure

Definition

Define τ_i for each i as k^{st} smallest $\frac{h_j}{w_j}$ for $j \in \{1, 2, \ldots, n\} - \{i\}$.

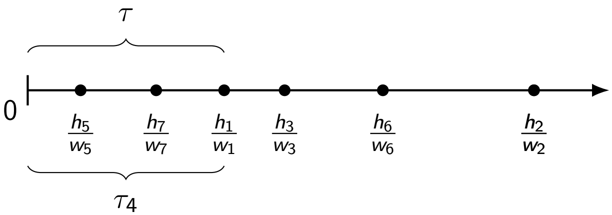


Figure: Illustration of Selecting τ, τ_4 for $k = 3$ in a Set of $n = 7$ Elements

Priority Sampling: Variance Bound Proof Structure

Definition

Define τ_i for each i as k^{st} smallest $\frac{h_j}{w_j}$ for $j \in \{1, 2, \ldots, n\} - \{i\}$.

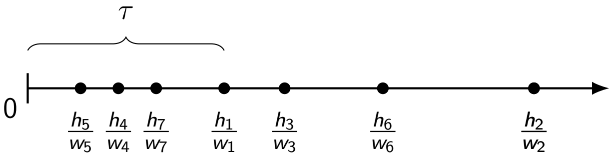


Figure: Illustration of Selecting τ for $k = 3$ in a Set of $n = 7$ Elements

Priority Sampling: Variance Bound Proof Structure

Definition

Define τ_i for each i as k^{st} smallest $\frac{h_j}{w_j}$ for $j \in \{1, 2, \ldots, n\} - \{i\}$.

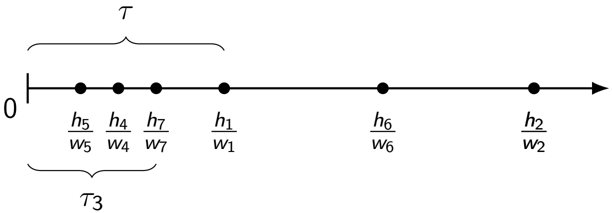


Figure: Illustration of Selecting τ, τ_3 for $k = 3$ in a Set of $n = 7$ Elements

Priority Sampling: Variance Bound Proof Structure

Definition

Define τ_i for each i as k^{st} smallest $\frac{h_j}{w_j}$ for $j \in \{1, 2, \ldots, n\} - \{i\}$.

Lemma (Variance Bound of Estimated Weight)

$$\text{Var}\left[\hat{w}_i\right] \leq w_i \cdot \mathbb{E}\left[\frac{1}{\tau_i}\right]$$

Priority Sampling: Variance Bound Proof Structure

Definition

Define τ_i for each i as k^{st} smallest $\frac{h_j}{w_j}$ for $j \in \{1, 2, \dots, n\} - \{i\}$.

Lemma (Variance Bound of Estimated Weight)

$$\text{Var}\left[\hat{w}_i\right] \leq w_i \cdot \mathbb{E}\left[\frac{1}{\tau_i}\right]$$

Lemma (Expected Inverse Threshold)

$$\mathbb{E}\left[\frac{1}{\tau_i}\right] \leq \frac{W}{k-1}$$

Where $W = \sum_{i=1}^{n} w_i$

Priority Sampling: Proof

Lemma (Expected Inverse Threshold)

$$\mathbb{E}\left[\frac{1}{\tau_i}\right] \leq \frac{W}{k-1}$$

Where $W = \sum_{i=1}^{n} w_i$

Priority Sampling: Proof

Lemma (Expected Inverse Threshold)

$$\mathbb{E}\left[\frac{1}{\tau_i}\right] \leq \frac{W}{k-1}$$

Where $W = \sum_{i=1}^{n} w_i$

$$\hat{W} = \sum_{i=0}^{n} \hat{w}_i$$

Priority Sampling: Proof

Lemma (Expected Inverse Threshold)

$$\mathbb{E}\left[\frac{1}{\tau_i}\right] \leq \frac{W}{k-1}$$

Where $W = \sum_{i=1}^{n} w_i$

$$\hat{W} = \sum_{i=0}^{n} \hat{w}_i = \sum_{i \in S} \frac{w_i}{\min(1, \tau w_i)}$$

Priority Sampling: Proof

Lemma (Expected Inverse Threshold)

$$\mathbb{E}\left[\frac{1}{\tau_i}\right] \leq \frac{W}{k-1}$$

Where $W = \sum_{i=1}^{n} w_i$

$$\hat{W} = \sum_{i=0}^{n} \hat{w}_i = \sum_{i \in S} \frac{w_i}{\min(1, \tau w_i)} = \sum_{i \in S} \max\left(w_i, \frac{1}{\tau}\right).$$

Priority Sampling: Proof

Lemma (Expected Inverse Threshold)

$$\mathbb{E}\left[\frac{1}{\tau_i}\right] \leq \frac{W}{k-1}$$

Where $W = \sum_{i=1}^{n} w_i$

$$\hat{W} = \sum_{i=0}^{n} \hat{w}_i = \sum_{i \in S} \frac{w_i}{\min(1, \tau w_i)} = \sum_{i \in S} \max\left(w_i, \frac{1}{\tau}\right).$$

$$\Rightarrow \hat{W} \geq \sum_{i \in S} \frac{1}{\tau} = \frac{k}{\tau}$$

Priority Sampling: Proof

Lemma (Expected Inverse Threshold)

$$\mathbb{E}\left[\frac{1}{\tau_i}\right] \leq \frac{W}{k-1}$$

Where $W = \sum_{i=1}^{n} w_i$

$$\hat{W} = \sum_{i=0}^{n} \hat{w}_i = \sum_{i \in S} \frac{w_i}{\min(1, \tau w_i)} = \sum_{i \in S} \max\left(w_i, \frac{1}{\tau}\right).$$

$$\Rightarrow \hat{W} \geq \sum_{i \in S} \frac{1}{\tau} = \frac{k}{\tau}$$

$$\Rightarrow \mathbb{E}[\hat{W}] \geq k\,\mathbb{E}\left[\frac{1}{\tau}\right]$$

Priority Sampling: Proof

Lemma (Expected Inverse Threshold)

$$\mathbb{E}\left[\frac{1}{\tau_i}\right] \leq \frac{W}{k-1}$$

Where $W = \sum_{i=1}^{n} w_i$

$$\hat{W} = \sum_{i=0}^{n} \hat{w}_i = \sum_{i \in S} \frac{w_i}{\min\left(1, \tau w_i\right)} = \sum_{i \in S} \max\left(w_i, \frac{1}{\tau}\right).$$

$$\Rightarrow \hat{W} \geq \sum_{i \in S} \frac{1}{\tau} = \frac{k}{\tau}$$

$$\Rightarrow \mathbb{E}[\hat{W}] \geq k\,\mathbb{E}\left[\frac{1}{\tau}\right]$$

$$\Rightarrow W \geq k\,\mathbb{E}\left[\frac{1}{\tau}\right] \Rightarrow \mathbb{E}\left[\frac{1}{\tau}\right] \leq \frac{W}{k}$$

Priority Sampling: Proof

Lemma (Expected Inverse Threshold)

$$\mathbb{E}\left[\frac{1}{\tau_i}\right] \leq \frac{W}{k-1}$$

Where $W = \sum_{i=1}^{n} w_i$

$$\mathbb{E}\left[\frac{1}{\tau}\right] \leq \frac{W}{k}$$

τ: $(k+1)^{\text{st}}$ smallest $\frac{h_j}{w_j}$ for $j \in \{1, 2, \ldots, n\}$.
τ_i: k^{st} smallest $\frac{h_j}{w_j}$ for $j \in \{1, 2, \ldots, n\} - \{i\}$.

Priority Sampling: Proof

Lemma (Expected Inverse Threshold)

$$\mathbb{E}\left[\frac{1}{\tau_i}\right] \leq \frac{W}{k-1}$$

Where $W = \sum_{i=1}^{n} w_i$

$$\mathbb{E}\left[\frac{1}{\tau}\right] \leq \frac{W}{k}$$

τ: $(k+1)^{\text{st}}$ smallest $\frac{h_j}{w_j}$ for $j \in \{1, 2, \ldots, n\}$.
τ_i: k^{st} smallest $\frac{h_j}{w_j}$ for $j \in \{1, 2, \ldots, n\} - \{i\}$.

$$\Rightarrow \mathbb{E}\left[\frac{1}{\tau_i}\right] \leq \frac{W - w_i}{k-1} \leq \frac{W}{k-1}$$

Priority Sampling: Application

Inner Product Sketch

Our study demonstrates that straightforward proof enables extending our method to inner product sketching. Our analysis shows priority sampling outperforms the Johnson-Lindenstrauss (JL) transform in reducing estimation error.

Reference Paper :

Title: "Sampling Methods for Inner Product Sketching"
Authors: Majid Daliri, Juliana Freire, Christopher Musco, Aécio Santos, Haoxiang Zhang

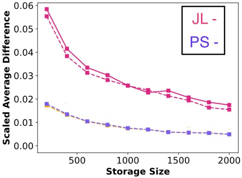


Figure: Experimental Results of JL vs Priority sampling

Questions?

Any Questions?