Perfect Sampling in Turnstile Streams Beyond Small Moments

David P. Woodruff[1], Shenghao Xie[2], Samson Zhou[2]
Carnegie Mellon University[1], Texas A&M University[2]

To appear in PODS 2025

Model

Frequency vector: Given a set S of m elements from $[n]$, let f_i be the frequency of element i.

$$1\ 1\ 2\ 1\ 3\ 1\ 2\ 3 \rightarrow [4,\ 2,\ 2,\ 0] := f$$

Streaming model: Elements of the data set S arrives sequentially in a data stream.

Turnstile stream: Updates can increase and decrease the coordinates of f.

Goal: Evaluation of a given function on f, using sublinear space in the size m of input S.

Problem

Approximate L_p sampler: Given $\varepsilon > 0$, sample $i \in [n]$ with probability $(1+\varepsilon)\frac{|f_i|^p}{\|f\|_p^p} + \frac{1}{\text{poly}(n)}$, where $\|f\|_p^p := f_1^p + f_2^p + \ldots + f_n^p$.

Perfect L_p sampler: Sample $i \in [n]$ with probability $\frac{|f_i|^p}{\|f\|_p^p} + \frac{1}{\text{poly}(n)}$, where $\|f\|_p^p := f_1^p + f_2^p + \ldots + f_n^p$.

Motivation: Minimal bias; privacy protection.

Application: DDoS attack detection; database management; distributed computing.

Serve as subroutine in essential problems: F_p moment estimation, finding heavy-hitter, finding duplicates.

Previous Results: $p \le 2$

Space UB	Remark
$O\left(\frac{1}{\varepsilon^{\max(1,p)}}\log^2 n\right)$, [JST11]	$p < 2$, approximate
$O\left(\frac{1}{\varepsilon^2}\log^3 n\right)$, [JST11]	$p = 2$, approximate
$O(\log^2 n)$, [JW18]	$p < 2$, perfect
$O(\log^3 n)$, [JW18]	$p = 2$, perfect
Space LB	**Remark**
$\Omega(\log^2 n)$, [JST11]	$p < 2$, approximate

Perfect Sampler for $p \ge 2$

Why $p \ge 2$ matters? Prioritize elements with larger contributions. Have applications to sparse signal recovery, outlier detection, and high-dimensional data analysis.

Theorem 1: Given $p \ge 2$, there exists a perfect L_p sampler on a turnstile stream that uses $n^{1-2/p}\text{polylog}(n)$ bits of space. Moreover, it obtains a $(1+\varepsilon)$-estimation to the sampled item using $\frac{1}{\varepsilon^2}n^{1-2/p}\text{polylog}(n)$ bits of space.

Rejection Sampling: Use perfect L_2 sampler to sample i w.p. $\frac{|f_i|^2}{\|f\|_2^2}$. Reject each sample w.p. $p_i = |f_i|^{p-2} \cdot \frac{\|f\|_2^2}{n^{1-2/p}\cdot\|f\|_p^p}$. Use unbiased estimates of each term in the actual implementation.

Rejection probability is well-defined: $0 < p_i < 1$.

In expectation, returning each index i w.p. $\frac{|f_i|^p}{\|f\|_p^p} + \frac{1}{\text{poly}(n)}$.

Sketching dimension lower bound: $\Omega(n^{1-2/p}\log n)$ for L_p sampler using linear sketch.

Rejection Sampling Framework

Perfect G sampler: Given a non-negative function G, sample $i \in [n]$ with probability $\frac{G(f_i)}{\sum_{j=1}^n G(f_j)} + \frac{1}{\text{poly}(n)}$.

L_0 sampler: Sample $i \in [n]$ with probability $\frac{1}{\|f\|_0} + \frac{1}{\text{poly}(n)}$.

Framework: Suppose that $H > G(z)$. Obtain a L_0 sample, then reject with probability $\frac{G(f_i)}{H}$.

Function	Space
$G(z) = \log(1+\|z\|)$	$O(\log^3 n)$
$G(z) = \min(T, \|z\|^p)$	$O(T\log^2 n)$
$G(z) = \sum_{d=1}^{D} a_d\|z\|^d$	$n^{\max(0,1-2/p)}\text{polylog}(n)$

Approximate Sampler for $p \ge 2$

Theorem 2: Given $p \ge 2$, there exists an approximate L_p sampler that uses $n^{1-2/p}\log^2(n)\log(\frac{1}{\varepsilon})\text{polyloglog}(n)$ bits of space and has update time $\frac{1}{\varepsilon}\text{polylog}(\frac{1}{\varepsilon}, n)$.

Exponential scalings: Draw n i.i.d. exponential random variables $(e_1, \ldots, e_n)$, obtain vector $z \in \mathbb{R}^n$ by $z_i = \frac{f_i}{e_i^{1/p}}$.

$Pr[D(1) = i] = \frac{|f_i|^p}{\|f\|_p^p}$, $z_{D(i)}$ is the i-th largest coordinate.

Statistical test: Use CountSketch to estimate z. Reject If $z_{D(1)}$ and $z_{D(2)}$ is close. → Cannot detect the max.

Dependency: The failure probability depends on which index achieves the max, leading to incorrect distribution.

Duplication: [JW18] duplicates each coordinate n^c times and scale with different exponentials.

$$O(\log^2(n^c)) \text{ for } p < 2,\ O(n^{c^{1-2/p}}) \text{ for } p > 2$$

Two-stage Countsketch: Maintain CountSketch1 for the vector w consisted of the maximum of duplications of each entry: $w_i = \max_j f_i/e_{i,j}^{1/p}$, select the largest $\log(\frac{1}{\varepsilon})$ entry. Maintain CountSketch2 for the total vector z with w zeroed out, only record the first $\log(\frac{1}{\varepsilon})$ entry.

Application

Norm estimation of post-processing subsets: Given a post-processing subset Q, there is an algorithm that gives a $(1+\varepsilon)$-estimation to $\|f_Q\|_p^p$. For $\|f_Q\|_p^p < \alpha\cdot\|f_Q\|_p^p$, the algorithm uses $\frac{1}{\alpha\varepsilon^2}n^{1-2/p}\text{polylog}(n)$ bits of space.

References

[JST11] Hossein Jowhari, Mert Saglam, Gábor Tardos. Tight Bounds for Lp Samplers, Finding Duplicates in Streams, and Related Problems. PODS 2011.
[JW18] Jayaram Rajesh, David P.Woodruff. Perfect Lp Sampling in a Data Stream. FOCS 2018.