Optimal Dynamic Parameterized Subset Sampling

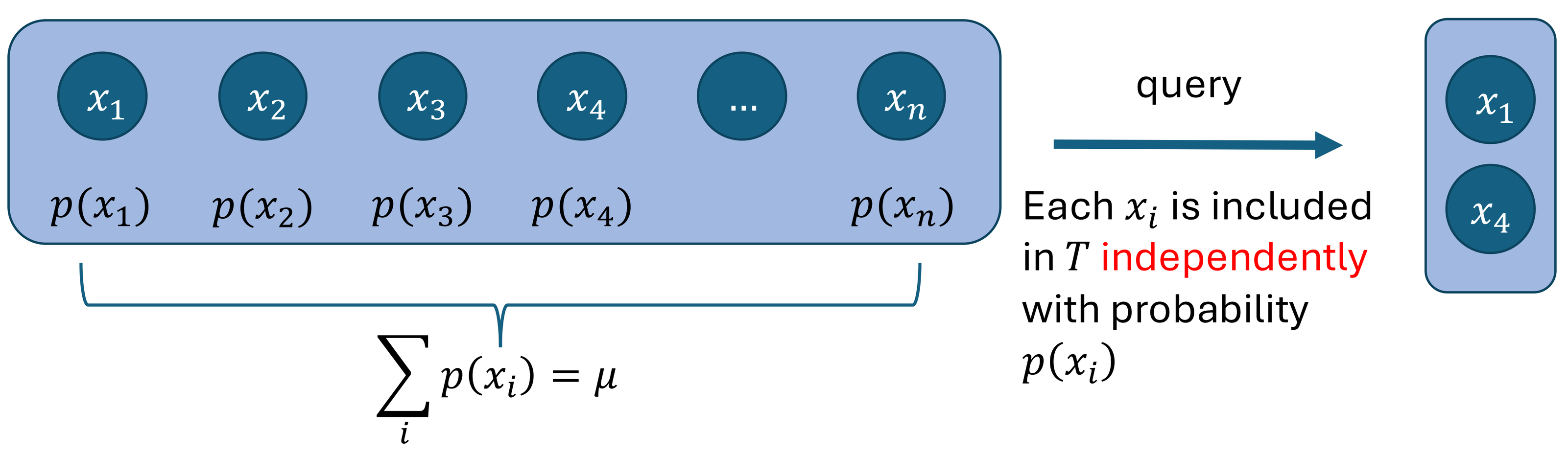
Junhao Gan[1], Seeun William Umboh[1,4], Hanzhi Wang[3],
Anthony Wirth[2], and **Zhuo Zhang[1]**

[1] The University of Melbourne, [2] The University of Sydney, [3] BARC, University of Copenhagen,
[4] ARC Training Centre in Optimisation Technologies, Integrated Methodologies, and Applications (OPTIMA)

Problem Overview

Subset Sampling

Given a set of n distinct items $S = \{x_1, \dots, x_n\}$, in which each item x_i has an associated probability $p(x_i)$, a query for the subset sampling problem returns a subset $T \subseteq S$, such that every x_i is independently included in T w.p. $p(x_i)$.

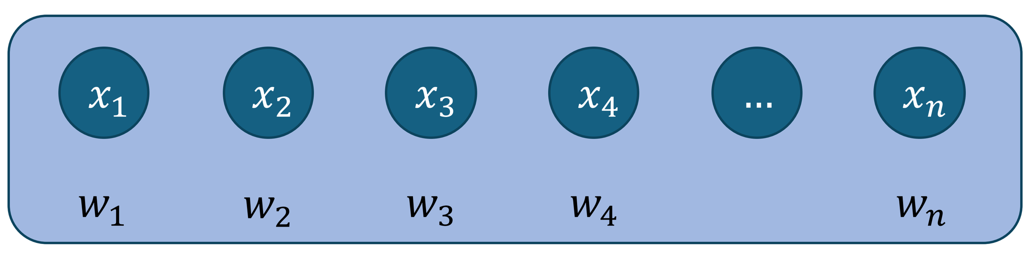


Best-known Result:

- query time: $O(1 + \mu)$
- update time: $O(1)$ (insert/deletion events)
- space complexity: O(n)

static setting [Bringmann & Panagiotou, 2017]
[Bhattacharya, Kiss, Sidford, Wajc, 2024]
[Yi, Wang, Wei, 2023]

Parameterized Subset Sampling



$p(x_i) = \min\left\{1, \frac{w_i}{W_S(\alpha,\beta)}\right\}$,
$W_S(\alpha, \beta) = \alpha \sum_{i=1}^{n} w_i + \beta$

➢ w_i: non-negative integer weights

α, β are input parameters, and can be given on-the-fly.

Our contributions in the Word RAM model:

- query time: $O(1 + \mu)$ in expectation
- update time (insert/delete): $O(1)$ worst case
- space complexity: $O(n)$ words for worst-case at all times
- preprocessing time: $O(n)$ worst-case

Other Results

➢ **Generating geometric variates:** Let $p \in (0,1)$ be a rational number which can be represented by a $O(1)$-word integer nominator and $O(1)$-word integer denominator. We can generate the following random variates in **$O(1)$ expected time with $O(n)$ worst-case space.**

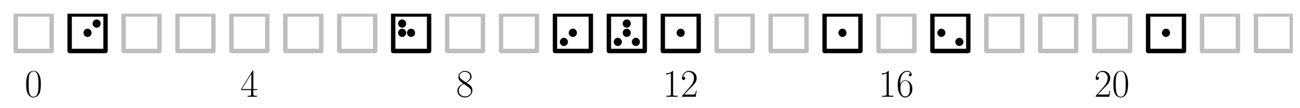
- Bernoulli(p) [Bringmann & Friedrich, 2013]
- Bernoulli$\left(\frac{1-(1-p)^n}{pn}\right)$ (new by us)
- Bound Geometric BG(p, n): $\Pr\{\mathrm{BG}(p,n) = i\} = \begin{cases} p(1-p)^{i-1}, & i \in [1, n) \\ (1-p)^{n-1}, & i = n \end{cases}$ [Bringmann & Friedrich, 2013]
- Truncated Geometric TG(p, n) (new by us): $\Pr\{\mathrm{TG}(p,n) = i\} = \frac{p(1-p)^{i-1}}{1-(1-p)^n}$

➢ **Hardness result:**

- Integer Sorting can be reduced to the deletion-only Dynamic Parameterized Subset Sampling (DPSS) with float weights.
- Optimal deletion-only DPSS with float weights implies a $O(N)$-expected-time algorithm for sorting N integers in the word RAM model with $\Omega(\log N)$ bit length. The latter remains an open problem.

Techniques

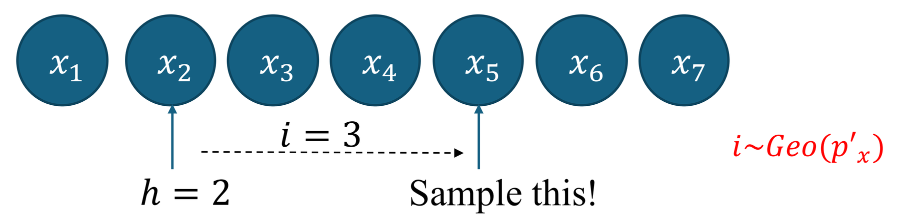
Bucketing-Based Algorithm: A Warm-Up



➢ Organize items into **power-of-two buckets**:
- ◆ Bucket $B(i)$ contains items with weights in $[2^i, 2^{i+1})$

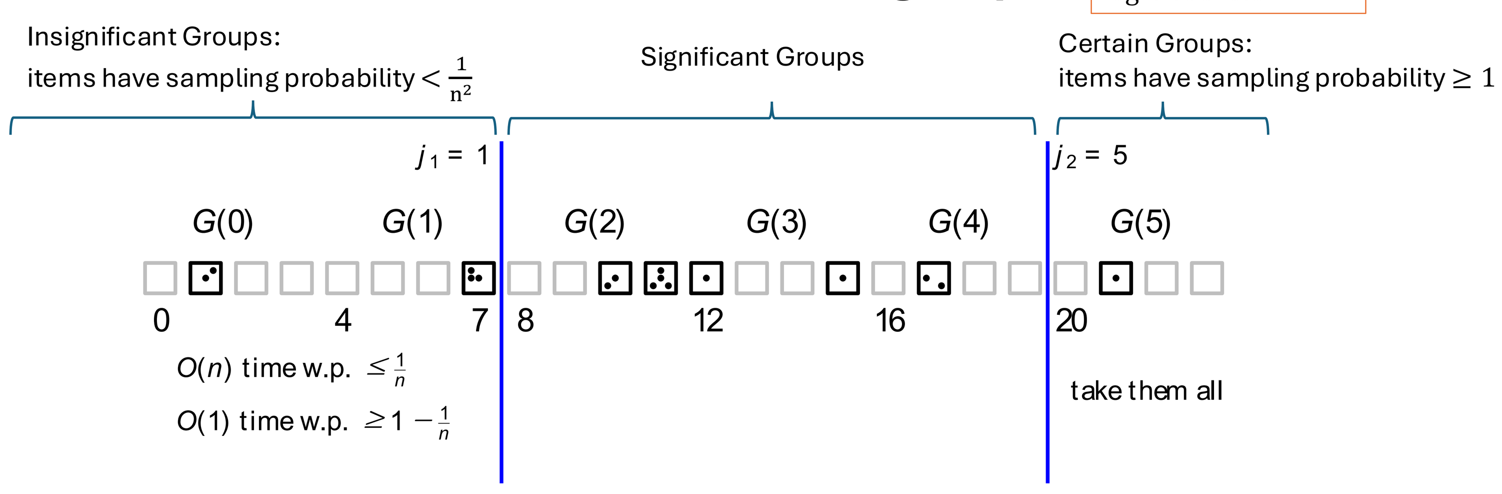
➢ For each non-empty bucket $B(i)$:
- ◆ Sample potential items using upper-bound probability $p'_x = \min\left\{1, \frac{2^{i+1}}{W_S(\alpha,\beta)}\right\}$
- ◆ Accept each potential item x with probability $\frac{p_x}{p'_x}$

➢ Query time: $O(b + \mu)$, where b is the number of non-empty buckets



Key idea: Avoid touching all buckets!

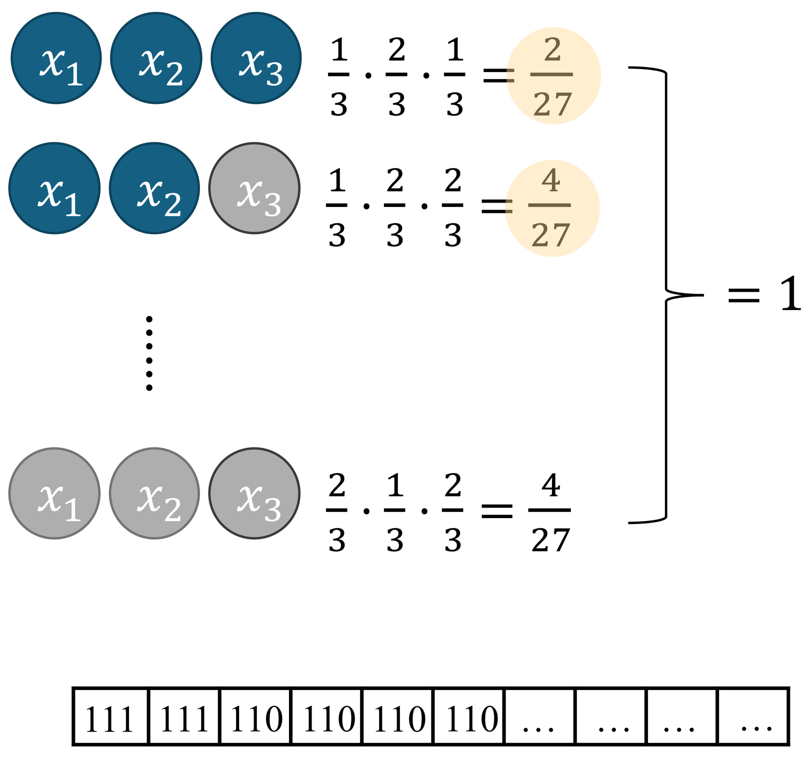
➢ Our solution: Partition the buckets into **groups**! (Each group contain $\log n$ buckets.)

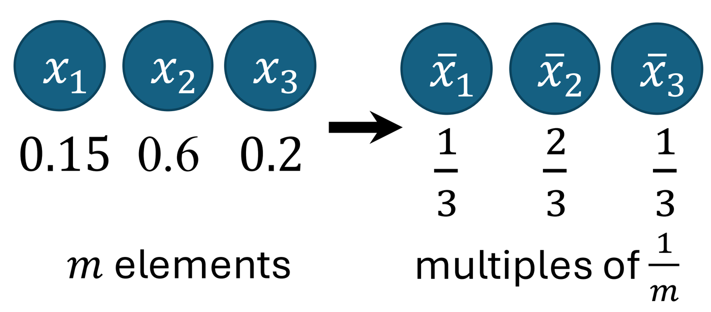


➢ **Handling Significant Groups**
- ◆ Step 1: Find the potential buckets, those containing at least one potential item:
  - This is another subset sampling problem! Recursion!
  - After three times recursion, the problem size is $O(\log\log\log n)$.
  - Small enough to be solved in a pre-computed look-up table.
- ◆ Step 2: Sample from the potential buckets:
  - How to find first potential item index k from a potential bucket?
  - This is a conditional probability problem!
  - It follows the Truncated Geometric Distribution.

Look-up Table Trick Overview

➢ The sampling probabilities of subsets





Deal with non-multiples of $1/m$

➢ $\bar{p}(x_i) = \frac{\lceil m p(x_i) \rceil}{m} \in \left\{\frac{1}{m}, \dots, \frac{m}{m}\right\}$

➢ Accept the event with $\bar{p}(x_i)/p(x_i)$

➢ Obtain a row for sampling!

➢ Just uniformly select an entry and return the subset as the sample result

- [Bringmann & Panagiotou, 2012] Bringmann K, Panagiotou K. Efficient sampling methods for discrete distributions // ICALP 2012
- [Bringmann & Friedrich, 2013] Bringmann K, Friedrich T. Exact and efficient generation of geometric random variates and random graphs //ICALP 2013
- [Bhattacharya, Kiss, Sidford, Wajc] Bhattacharya S, Kiss P, Sidford A, Wajc D. Near-optimal dynamic rounding of fractional matchings in bipartite graphs //STOC 2024
- [Yi, Wang, Wei, 2023] Yi L, Wang H, Wei Z. Optimal dynamic subset sampling: Theory and applications //KDD 2023