Online Thinning to Reduce Discrepancy

Raaz Dwivedi, Ohad Feldheim, Ori Gurel-Gurevich and Aaditya Ramdas
raaz.rsk@eecs.berkeley.edu, ohad@huji.edu, ori@huji.edu, aramdas@berkeley.edu
Berkeley UNIVERSITY OF CALIFORNIA
האוניברסיטה העברית בירושלים THE HEBREW UNIVERSITY OF JERUSALEM

Set-up

For a sequence $X^\infty = \{X_1, X_2, \ldots\}$ in $[0,1)^d$, let X^N denote the set $\{X_1, \ldots, X_N\}$. Define the discrepancy sequence as

$$\mathrm{Dis}(X^N) = \sup_{R\in\mathcal{R}} \left|\frac{1}{N}\nu_N(R) - \mathrm{vol}(R)\right|,$$

where $\mathcal{R}$ is the set of all axis aligned hyper-rectangles in $[0,1)^d$, $\mathrm{vol}(R)$ denotes the Lebesgue measure of R and

$$\nu_N(S) := |S \cap X^N| \quad \text{for any } S \subset [0,1)^d.$$

Objectives

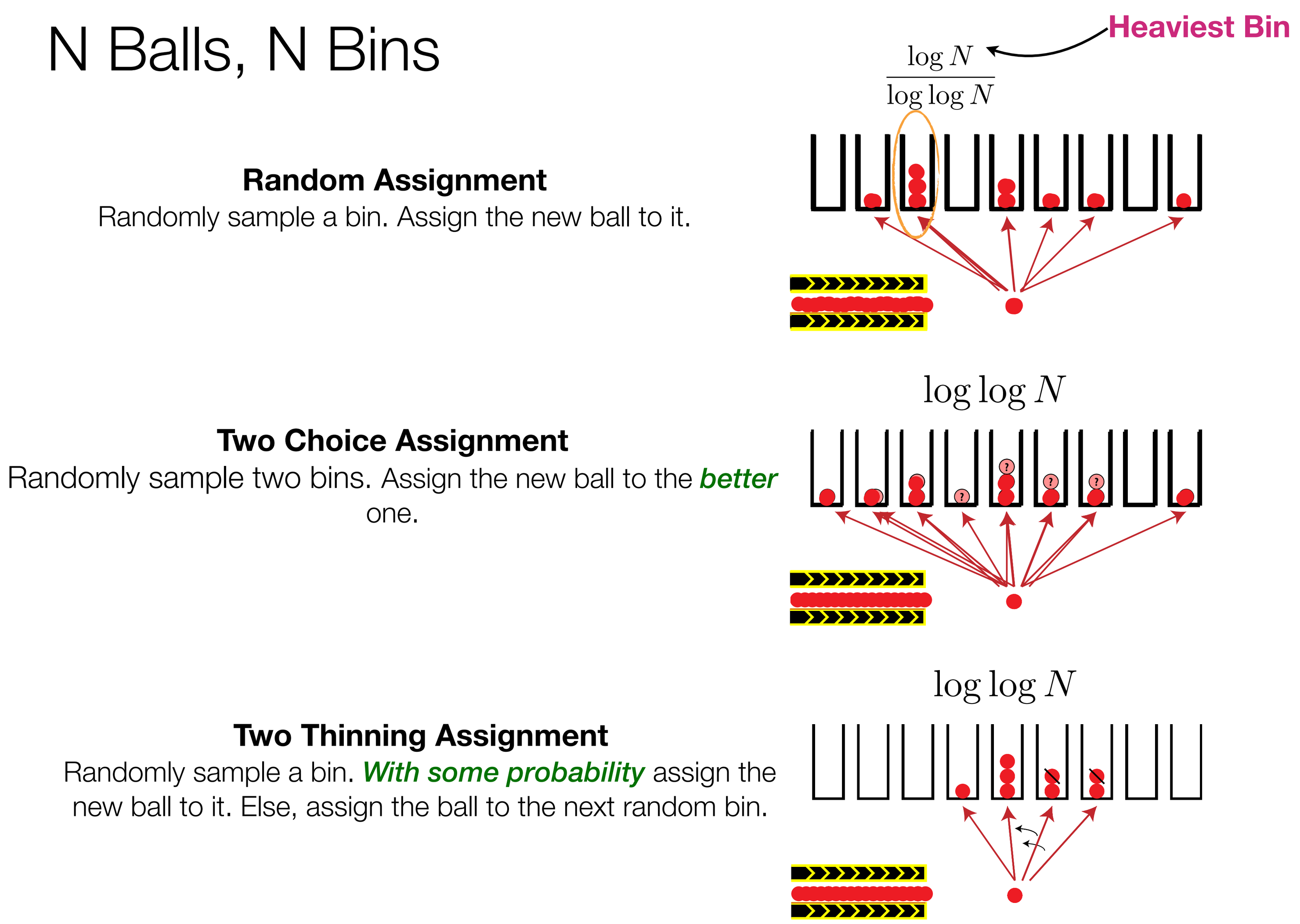
Fact: For independent and identically distributed (i.i.d.) sequence such that $X_i \sim U[0,1]^d, i = 1, 2, \ldots$

$$\mathrm{Dis}(X^N) = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right).$$

Goal: To select a *dense* subset $Z^\infty = \{Z_1, Z_2, \ldots\}$ of a streaming sequence X^∞ with $X_i \sim U[0,1]^d$, using an *online* and *time and space efficient* strategy such that

$$\mathrm{Dis}(Z^N) = \mathcal{O}\left(\frac{\mathrm{polylog}(N)}{N}\right).$$

Motivation

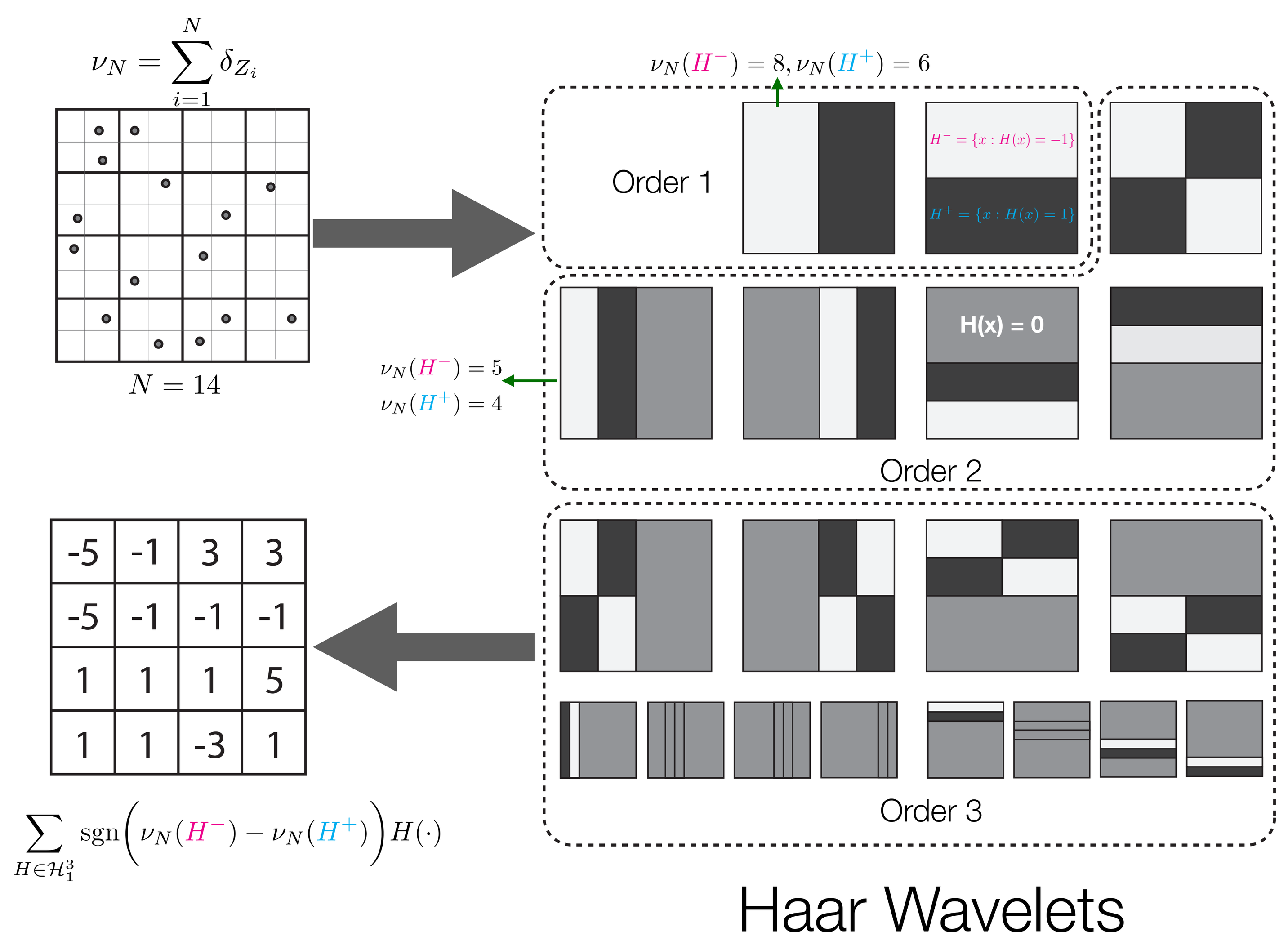


Extend the two thinning strategy to our set-up of axis alligned hyper-rectangles. To overcome, the cardinality of the set and heavy dependence across the elements of the set.
Key idea: Decouple them using Haar wavelets –

Haar 2-Thinning Strategy

Given $Z_1, \ldots, Z_N$, consider Haar wavelets of order up-to $\log_2 N$, denoted by $\mathcal{H}_N$. Using the *tilt* of each Haar, determine the net tilt of $\mathcal{H}_N$, and compute ***rate function of thinning*** as

$$\mathbb{P}(\text{accept } x) = 0.5 + \frac{1}{\sum_{H\in\mathcal{H}_N}|H(x)|}\overbrace{\sum_{H\in\mathcal{H}_N}\mathrm{sgn}(\nu_N(H^-) - \nu_N(H^+))\cdot H(x)}^{\text{net tilt of } \mathcal{H}_N \text{ for } x}$$



Haar Wavelets

Main Result

For a streaming sequence X^∞ of i.i.d. $U[0,1)^d$ random variables, Haar 2-Thinning strategy outputs a streaming sequence Z^∞ such that for any $\Delta > 1$

$$\mathbb{P}\left(\exists N \in \mathbb{N} : \mathrm{Dis}(Z^N) > \Delta \cdot \frac{d\log^{2d+1}(N)}{N}\right) \le 10^7 \exp(-2\Delta).$$

The strategy takes $\mathcal{O}(d\log^d N)$ space and $\mathcal{O}(d\log^d N)$ computation time to output N points and *accepts at least half of the points* of X^N for any N.

Haar Greedy-Thinning Strategy

Conjecture: A simplified greedy strategy with $\mathcal{O}(\log^d N)$ space-time complexity outputs a thinned sequence Z^∞ improves over Haar 2-Thinning by a factor of $\log^d N$, that is

$$\mathrm{Dis}(Z^N) = \mathcal{O}(d\log^d N/N).$$

Greedy Strategy: Accept a point x in the sequence, if the net tilt of $\mathcal{H}_N$ at x is positive, otherwise accept the next point of the sequence.

Numerical Experiments

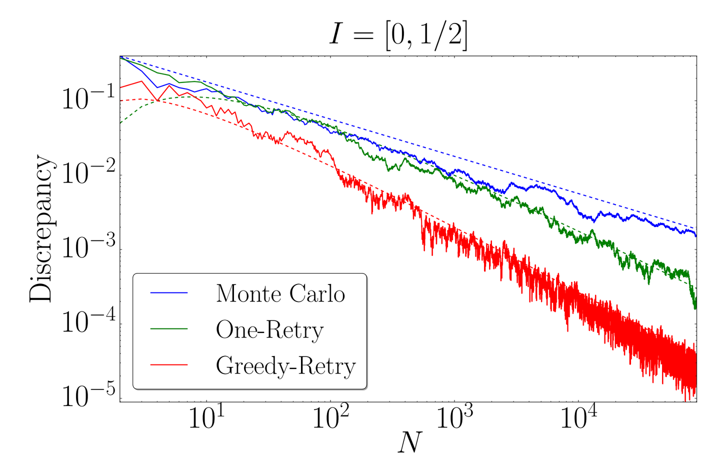


Figure 1:

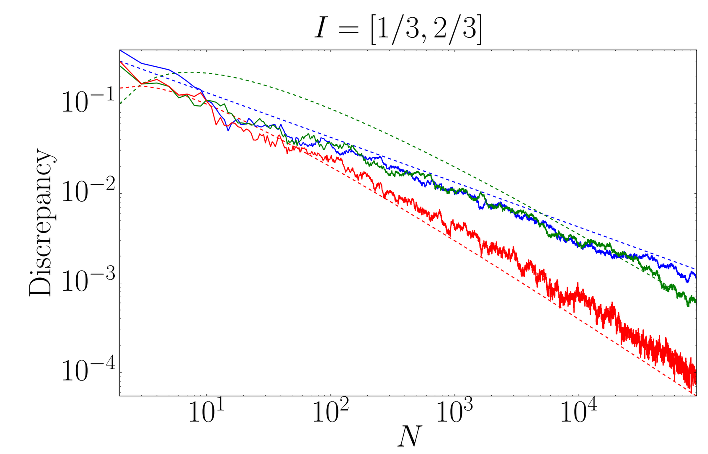


Figure 2:

Proof Techniques

- Write arbitrary rectangles in terms of "diadic rectangles"
- Write diadic rectangles in terms of "Haar wavelets"
- Control the balance of points in Haar wavelets upto some resolution

References