Minimax Optimal Kernel Two-sample Testing in Sub-quadratic Time

Ikjun Choi, Shourya Pandey, Purnamrita Sarkar

The University of Texas at Austin

Problem: Two-sample testing

- Given mutually independent samples

$$\mathbf{X} = (X_1, \ldots, X_n) \overset{iid}{\sim} P$$

$$\mathbf{Y} = (Y_1, \ldots, Y_n) \overset{iid}{\sim} Q$$

- A test $\varphi(X, Y)$ decides between

$$H_0 : P = Q \quad \text{v.s.} \quad H_1 : d(P, Q) > \epsilon$$

where d is the metric of interest.

- Type I error: $\mathbb{P}(\varphi \text{ rejects } H_0 \,|\, P = Q) \leq \alpha \quad (\text{level-}\alpha)$
- Type II error: $\mathbb{P}(\varphi \text{ accepts } H_0 \,|\, d(P, Q) > \epsilon) \leq \beta \quad (\text{power}(\varphi) \geq 1 - \beta)$

Criterion: Minimax Framework

- Uniform separation rate $\rho_n(\varphi)$: for any (P, Q),

  $\text{power}(\varphi) \geq 1 - \beta$ is uniformly guaranteed when $d(P, Q) > \rho_n$.

  $\boxed{\text{Lower uniform separation rate} \rightarrow \text{better test performance in terms of uniform power guarantee}}$

- Minimax separation rate ϵ_n^* : lowest $\rho_n(\varphi)$ among all level-α tests.

  $\boxed{\text{Lowest uniform separation rate} = \text{best test in terms of uniform power guarantee}}$

Criterion: Minimax Framework

- Uniform separation rate $\rho_n(\varphi)$: for any (P, Q),

  $\mathrm{power}(\varphi) \geq 1 - \beta$ is uniformly guaranteed when $d(P, Q) > \rho_n$.

  Lower uniform separation rate $\rightarrow$ better test performance in terms of uniform power guarantee

- Minimax separation rate ϵ_n^* : lowest $\rho_n(\varphi)$ among all level-α tests.

  Lowest uniform separation rate $=$ best test in terms of uniform power guarantee

Goal: Define a test achieving performance comparable to the best test

If $\rho_n(\varphi) = O(\epsilon_n^*)$, a test φ is minimax (rate) optimal.

Background: Maximum Mean Discrepancy

- A distance between distributions parameterized by a kernel k

$$\mathrm{MMD}^2(P, Q\,; k) := \mathbb{E}_{X,X'\sim P}\left[k(X, X')\right] - 2\mathbb{E}_{X\sim P, Y\sim Q}\left[k(X, Y)\right] + \mathbb{E}_{Y,Y'\sim Q}\left[k(Y, Y')\right].$$

- Quadratic-time $O(n^2 d)$ unbiased estimator:

$$\widehat{\mathrm{MMD}}^2(\mathbf{X}, \mathbf{Y}\,; k) := \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} k\left(X_i, X_j\right) - \frac{2}{nm} \sum_{i=1}^{n}\sum_{j=1}^{m} k\left(X_i, Y_j\right) + \frac{1}{m(m-1)} \sum_{1\le i\ne j\le m} k\left(Y_i, Y_j\right).$$

- **Standard MMD test**:

$$\varphi(\mathbf{X}, \mathbf{Y}) = \mathbb{1}\{\widehat{\mathrm{MMD}}^2(\mathbf{X}, \mathbf{Y}\,; k) > q_\alpha\}$$

  - threshold q_α can be selected using permutations.

Motivation

Standard MMD test

- minimax optimal power
- quadratic time

Scalable approx. MMD test

- sub-optimal power
- sub-quadratic time

[GFH+09, ZGB13, YWT+19, SKG+22,...]

$\boxed{\text{Can we simultaneously achieve both optimal power and sub-quadratic time?}}$

Motivation

Standard MMD test

- minimax optimal power
- quadratic time

Scalable approx. MMD test

- sub-optimal power
- sub-quadratic time

[GFH+09, ZGB13, YWT+19, SKG+22,...]

Can we simultaneously achieve both optimal power and sub-quadratic time?

Prior works: yes, under favorable tail assumptions on distributions.

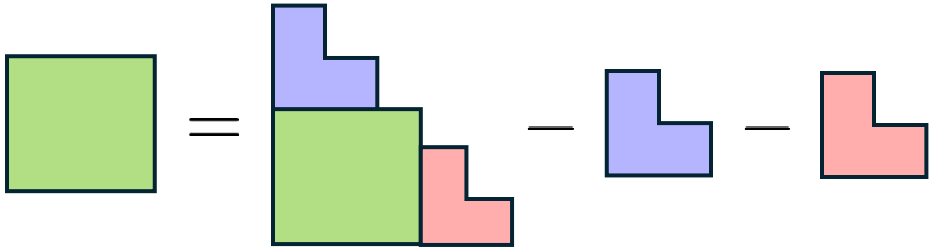
- Compress Then Test: Powerful Kernel Testing in Near-linear Time — Domingo-Enrich, Dwivedi, Mackey (2023)
  - Comp.-Stat. Trade-off in Kernel Two-Sample Testing with Random Fourier Features — Choi, Kim (2024)

Our contribution: yes, without such assumptions.

Insight: $O(n^2) \to O(n \log n)$ at $d = 1$

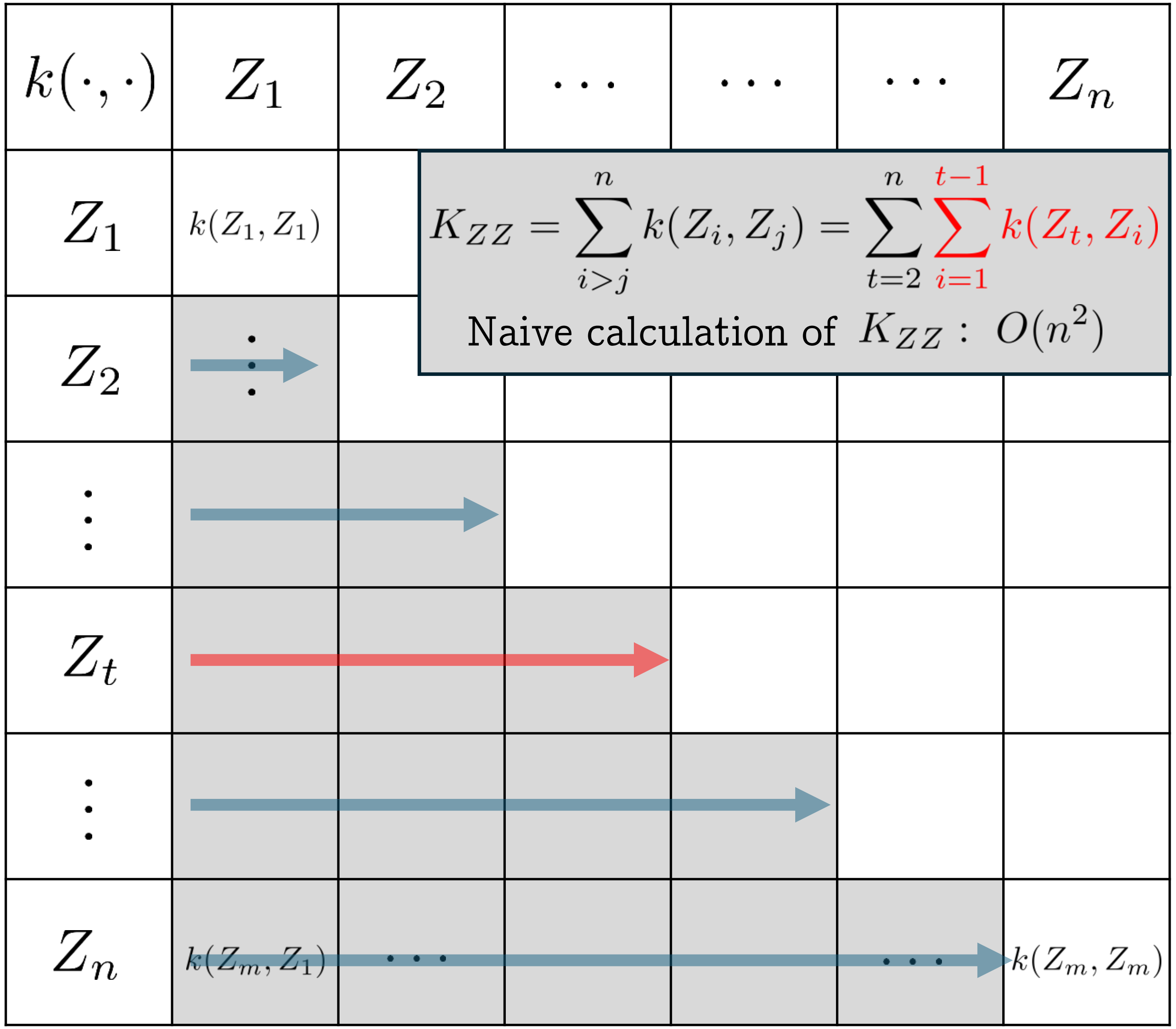
$k(\cdot,\cdot)$	X_1	$\cdots$	X_n	Y_1	$\cdots$	Y_m
X_1	$k(X_1, X_1)$					
$\vdots$	$\vdots$					
X_n	$k(X_n, X_1)$	$\cdots$	$k(X_n, X_n)$			
Y_1	$k(Y_1, X_1)$	$\cdots$	$k(Y_1, X_n)$	$k(Y_1, Y_1)$		
$\vdots$	$\vdots$		$\vdots$	$\vdots$		
Y_m	$k(Y_m, X_1)$	$\cdots$	$k(Y_m, X_n)$	$k(Y_n, Y_1)$	$\cdots$	$k(Y_n, Y_n)$

$$\widehat{\mathrm{MMD}}^2(\mathbf{X}, \mathbf{Y}; k) := \frac{2}{n(n-1)} K_{XX} - \frac{2}{n^2} K_{XY} + \frac{2}{n(n-1)} K_{YY}$$



MMD calculation $\longrightarrow$ lower triangular sum

Insight: $O(n^2) \to O(n \log n)$ at $d = 1$



Insight: $O(n^2) \to O(n \log n)$ at $d = 1$

$k(\cdot,\cdot)$	Z_1	Z_2	$\cdots$	$\cdots$	$\cdots$	Z_n
Z_1	$k(Z_1, Z_1)$					
Z_2	$\vdots$					
$\vdots$						
Z_t						
$\vdots$						
Z_n	$k(Z_m, Z_1)$	$\cdots$			$\cdots$	$k(Z_m, Z_m)$

$$K_{ZZ} = \sum_{i>j}^{n} k(Z_i, Z_j) = \sum_{t=2}^{n} \sum_{i=1}^{t-1} k(Z_t, Z_i)$$

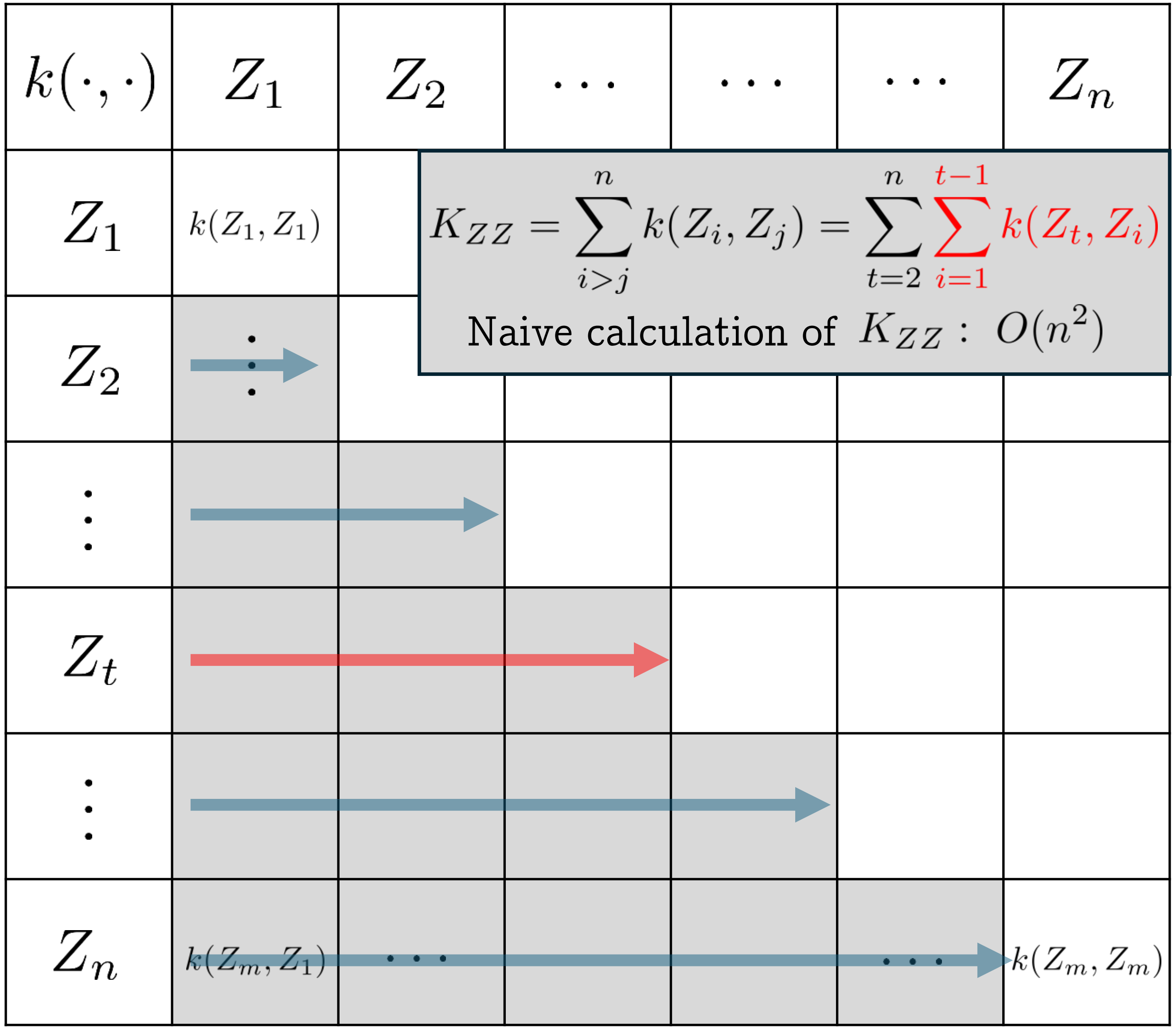
Naive calculation of K_{ZZ} : $O(n^2)$

Laplace kernel

$$k(x, y) = \exp(-|x - y|)$$
$$= \mathbb{1}\{x \geq y\} e^{-x} e^{y} + \mathbb{1}\{x < y\} e^{x} e^{-y}$$

$$\sum_{i=1}^{t-1} k(Z_t, Z_i) = \sum_{i:Z_t > Z_i}^{t} k(Z_t, Z_i) + \sum_{i:Z_t < Z_i}^{t} k(Z_t, Z_i)$$
$$= \sum_{i:Z_t > Z_i}^{t} e^{-Z_t} e^{Z_i} + \sum_{i:Z_t < Z_i}^{t} e^{Z_t} e^{-Z_i}$$
$$= e^{-Z_t} \boxed{\sum_{i:Z_t > Z_i}^{t} e^{Z_i}} + e^{Z_t} \boxed{\sum_{i:Z_t < Z_i}^{t} e^{-Z_i}}$$

Insight: $O(n^2) \to O(n \log n)$ at $d = 1$



$$K_{ZZ} = \sum_{i>j}^{n} k(Z_i, Z_j) = \sum_{t=2}^{n} \sum_{i=1}^{t-1} k(Z_t, Z_i)$$

Naive calculation of K_{ZZ} : $O(n^2)$

Laplace kernel

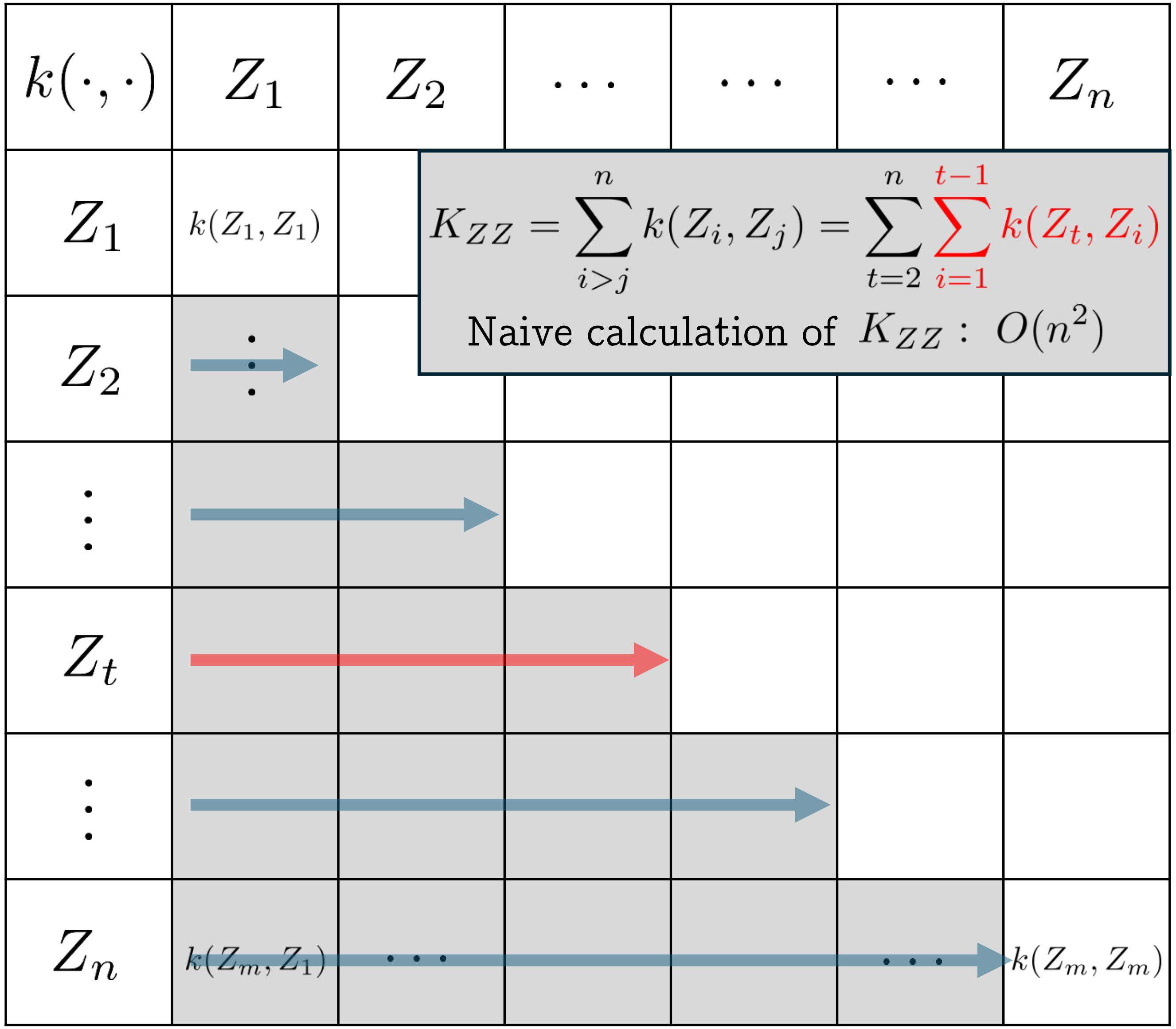
$$k(x,y) = \exp(-|x-y|)$$
$$= \mathbb{1}\{x \geq y\} e^{-x} e^{y} + \mathbb{1}\{x < y\} e^{x} e^{-y}$$

$$\sum_{i=1}^{t-1} k(Z_t, Z_i) = \sum_{i:Z_t > Z_i}^{t} k(Z_t, Z_i) + \sum_{i:Z_t < Z_i}^{t} k(Z_t, Z_i)$$
$$= e^{-Z_t} \boxed{\sum_{i:Z_t > Z_i}^{t} e^{Z_i}} + e^{Z_t} \boxed{\sum_{i:Z_t < Z_i}^{t} e^{-Z_i}}$$

Goal: calculate $\boxed{\text{sums}}$ efficiently

euMMD — Bodenham, Kawahara (2023): Running sum

Insight: $O(n^2) \to O(n \log n)$ at $d = 1$



Laplace kernel

$$k(x, y) = \exp(-|x - y|)$$
$$= \mathbb{1}\{x \geq y\} e^{-x} e^{y} + \mathbb{1}\{x < y\} e^{x} e^{-y}$$

$$\sum_{i=1}^{t-1} k(Z_t, Z_i) = \sum_{i:Z_t > Z_i}^{t} k(Z_t, Z_i) + \sum_{i:Z_t < Z_i}^{t} k(Z_t, Z_i)$$
$$= e^{-Z_t} \boxed{\sum_{i:Z_t > Z_i}^{t} e^{Z_i}} + e^{Z_t} \boxed{\sum_{i:Z_t < Z_i}^{t} e^{-Z_i}}$$

Goal: calculate $\boxed{\text{sums}}$ efficiently

euMMD — Bodenham, Kawahara (2023): Running sum

Our approach: Fenwick structure

$\boxed{\text{sums}}$ **in** $O(\log n)$ ➡ K_{ZZ} **in** $O(n \log n)$

Main Result: Beyond $d = 1$ Laplace

$d = 1$ Laplace MMD can be run in $O(n \log n)$. Same idea can be applied to:

Separable Kernel: compatible to Fenwick

$$k(x, y) = \mathbb{1}\{x \geq y\} f_-(x) g_-(y) + \mathbb{1}\{x < y\} f_+(x) g_+(y)$$

- Product of separable kernels $\Longrightarrow$ multidimensional Fenwick: $O(\log^d n)$ time for sums

Main Result: Beyond $d = 1$ Laplace

$d = 1$ Laplace MMD can be run in $O(n \log n)$. Same idea can be applied to:

Separable Kernel: compatible to Fenwick

$$k(x, y) = \mathbb{1}\{x \geq y\} f_-(x) g_-(y) + \mathbb{1}\{x < y\} f_+(x) g_+(y)$$

- Product of separable kernels ➡ multidimensional Fenwick: $O(\log^d n)$ time for sums

d-dimensional MMD test with $k(x, y) = \prod_{i=1}^d k_i(x_i, y_i)$

- Represent each k_i as an exact/approximate sum of separable kernels

➡ Represent k as a sum of product of separable kernels: $\prod_{i=1}^{d} \left(\sum \text{sep.} \right) = \prod_{i=1}^{d} \text{sep.} + \cdots + \prod_{i=1}^{d} \text{sep.}$

Main Result: Beyond $d = 1$ Laplace

d-dimensional MMD test with $k(x, y) = \prod_{i=1}^{d} k_i(x_i, y_i)$

- For minimax optimality:

k_i : **Kernel type**	R_i :# **Sep. kernels**	**Sep. representation**
Laplace	1	Exact
Matérn $(\nu = p + 1/2)$	$p + 1$	Exact
Gaussian	$\log n$	Approximate
Inverse-multiquadric	$\log^3 n$	Approximate

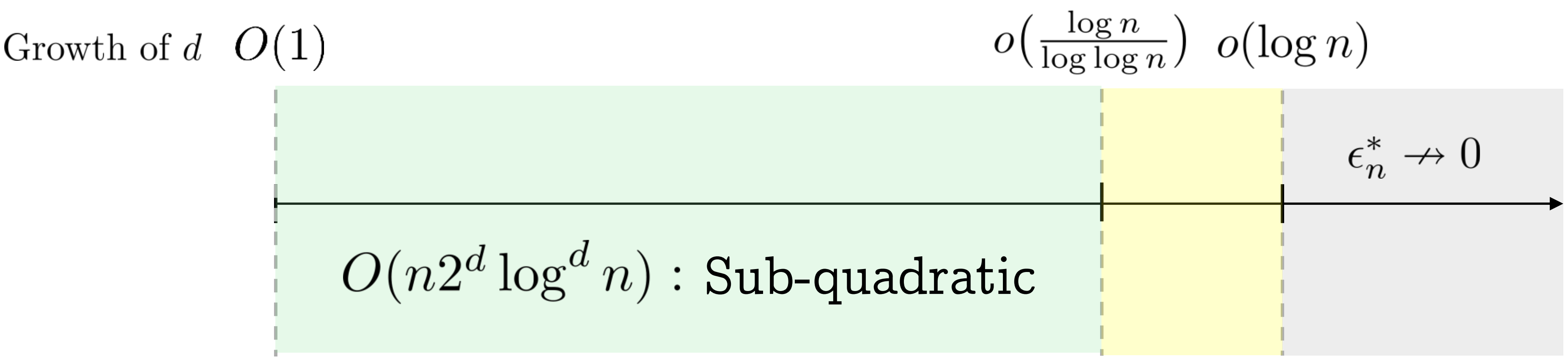
- Near-linear time: $O(n^2 d) \longrightarrow \prod_{i=1}^{d} R_i \cdot O(n 2^d \log^d n)$
- Minimax optimal for both L_2-separation over Sobolev ball: $\epsilon_n^* \asymp n^{-2s/(4s+d)}$

  MMD-separation: $\epsilon_n^* \asymp n^{-1/2}$

Main Result: Beyond $d = 1$ Laplace

Extension. Minimax analysis in growing d regime (new!)

$$L_2\text{-distance, over Sobolev ball: } \epsilon_n^* \asymp C^{d/(4s+d)} n^{-2s/(4s+d)}$$



(Laplace, Gaussian) minimax optimal Fenwick-MMD test time.

MMD with PCA for low-dim. latent structure

- $O(n2^d \log^d n) \longrightarrow$ limited scalability due to exponential dependence on d.
- High-dimensional data often exhibit a low-dimensional latent structure.
- Suggestion: 1. split the data.
  2. fit PCA on one half, project the other half onto the principal components.
  3. run the Fenwick-MMD test.

MMD with PCA for low-dim. latent structure

- $O(n2^d \log^d n) \longrightarrow$ limited scalability due to exponential dependence on d .
- High-dimensional data often exhibit a low-dimensional latent structure.
- Suggestion: 1. split the data.
  2. fit PCA on one half, project the other half onto the principal components.
  3. run the Fenwick-MMD test.

Latent low-dimensional signal model

Ambient dimension d, latent dimension r, $d > r$

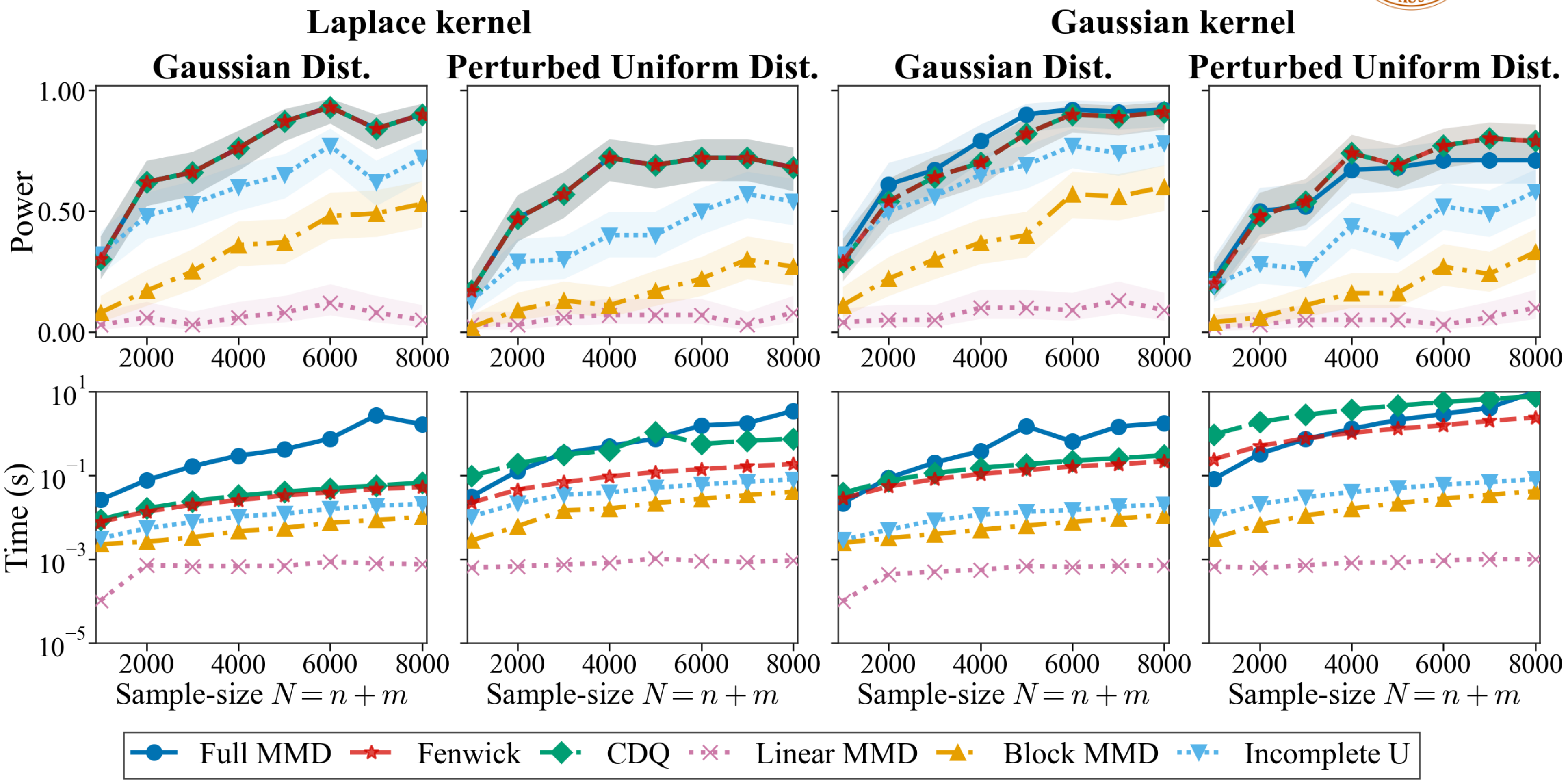
$$X = \mathbf{M}U + \xi, \quad Y = \mathbf{M}V + \xi'$$

$X, Y \in \mathbb{R}^d$, $U, V \in \mathbb{R}^r$, $\mathbf{M} \in \mathbb{R}^{d \times r}$, $\mathbf{M}^\top \mathbf{M} = \mathbf{I}_r$

U, V : arbitrary latent random vector, ξ, ξ' : noise

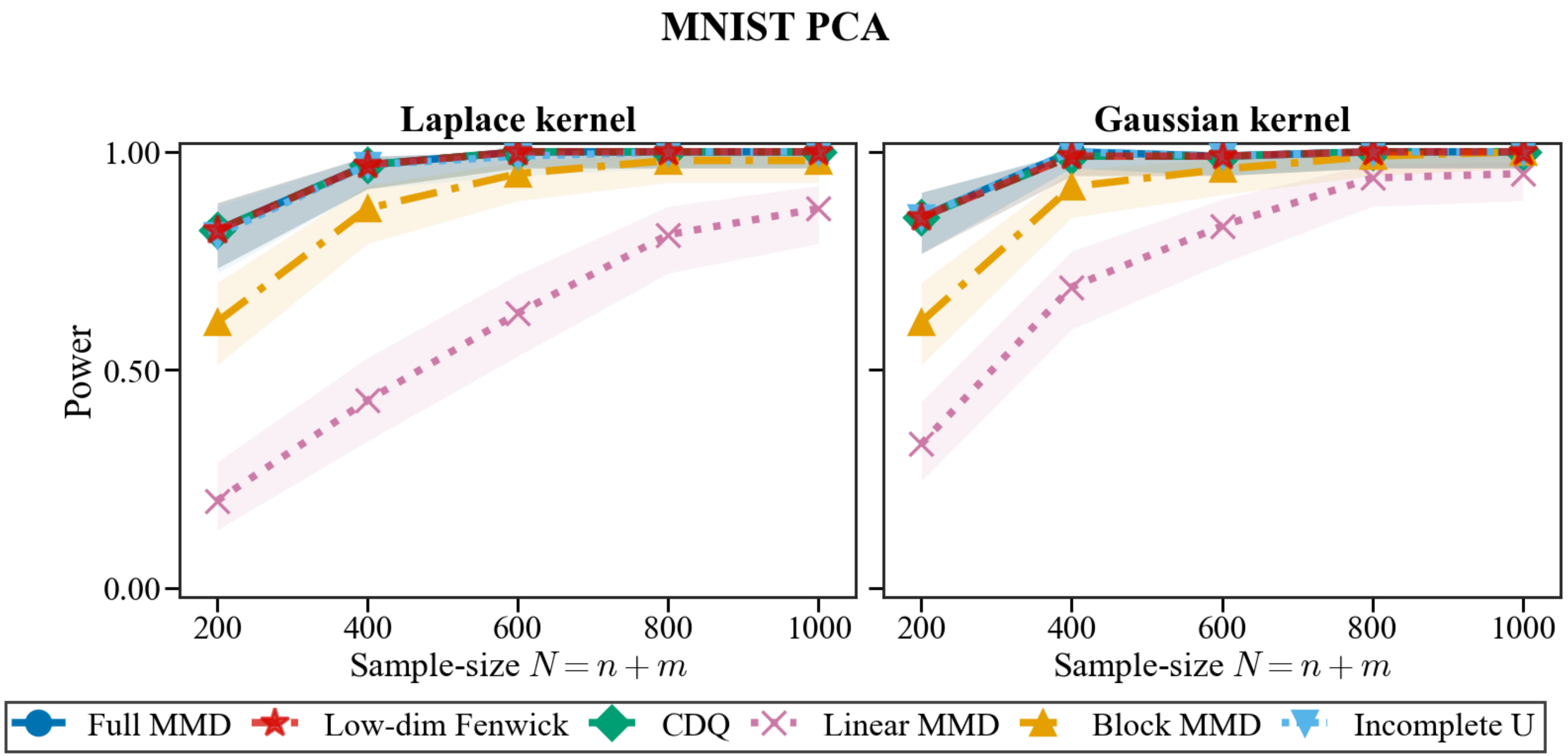
Under regularity assumptions, **PCA-projection** only loses a constant fraction of $L_2(P, Q)$

Experiments



Two sample testing experiments. The left (right) half uses the Laplace (Gaussian) kernel. The powers are averaged over 100 iterations, and the permutation test is run with 100 permutations.

Experiments

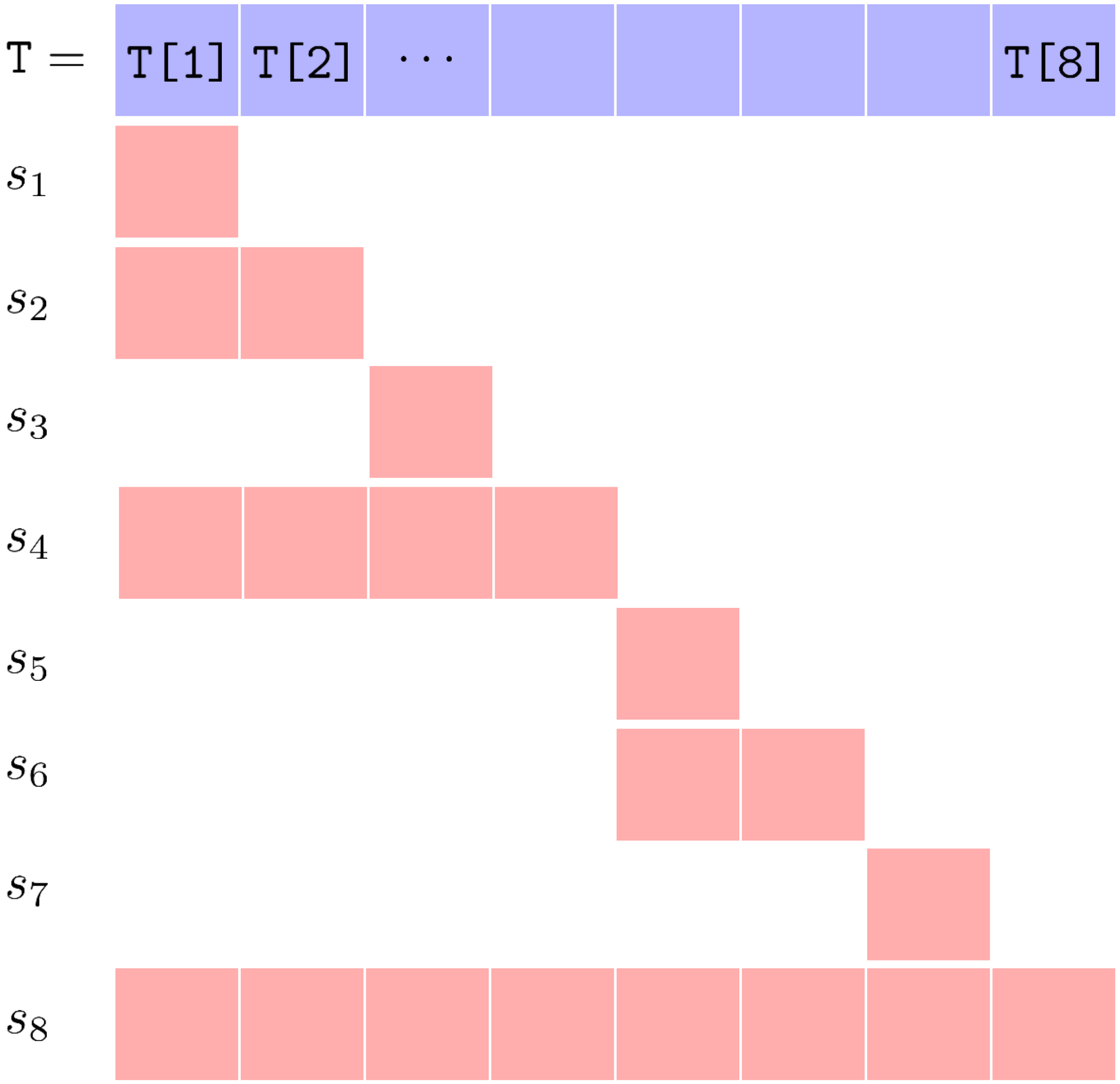


Two sample testing experiments with dimension reduction via PCA. Tests decide between even number vs odd number.

Thank you for your attention!

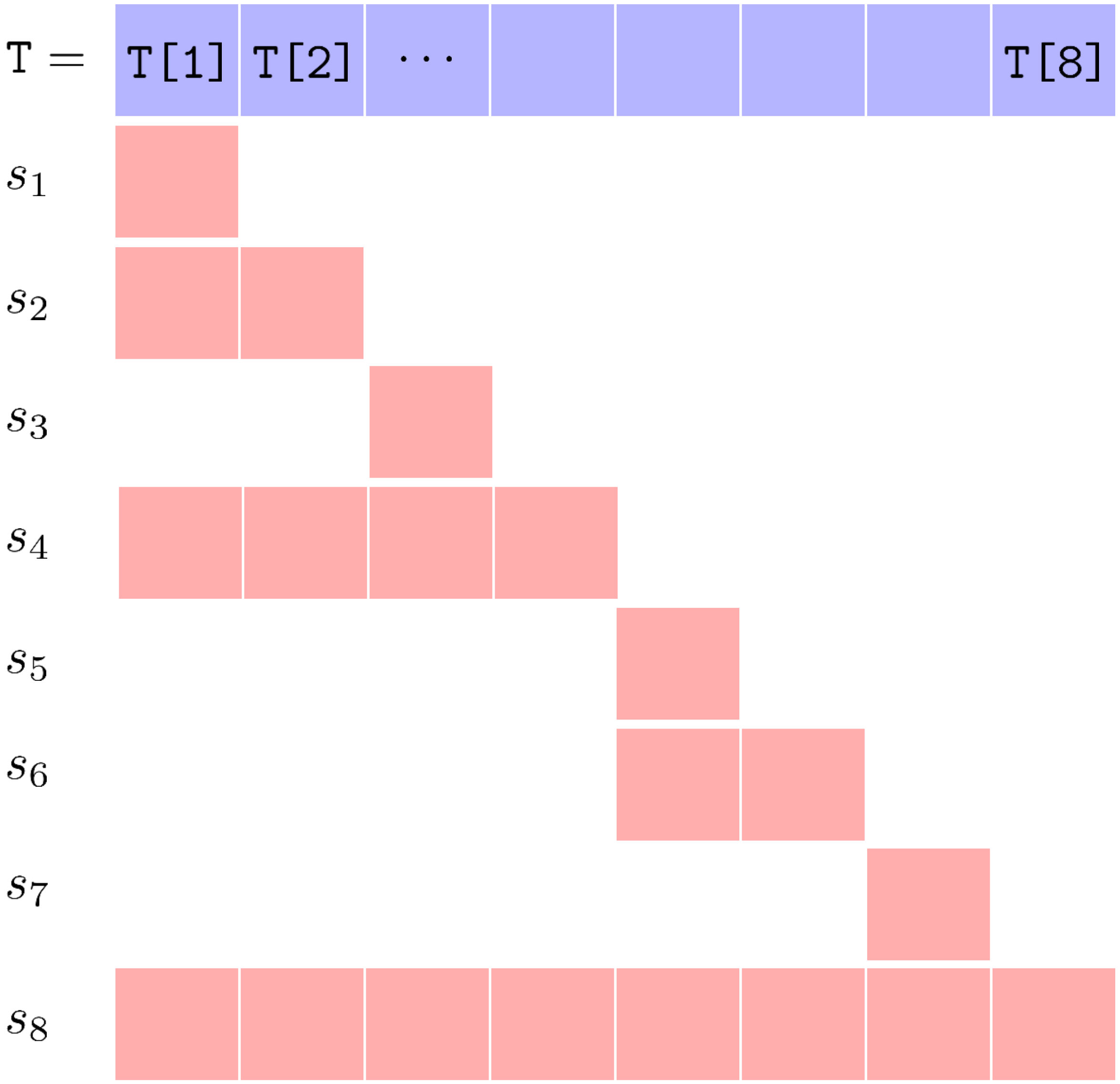
Full paper with more experiments coming soon!

Supplement 1: Permutation test



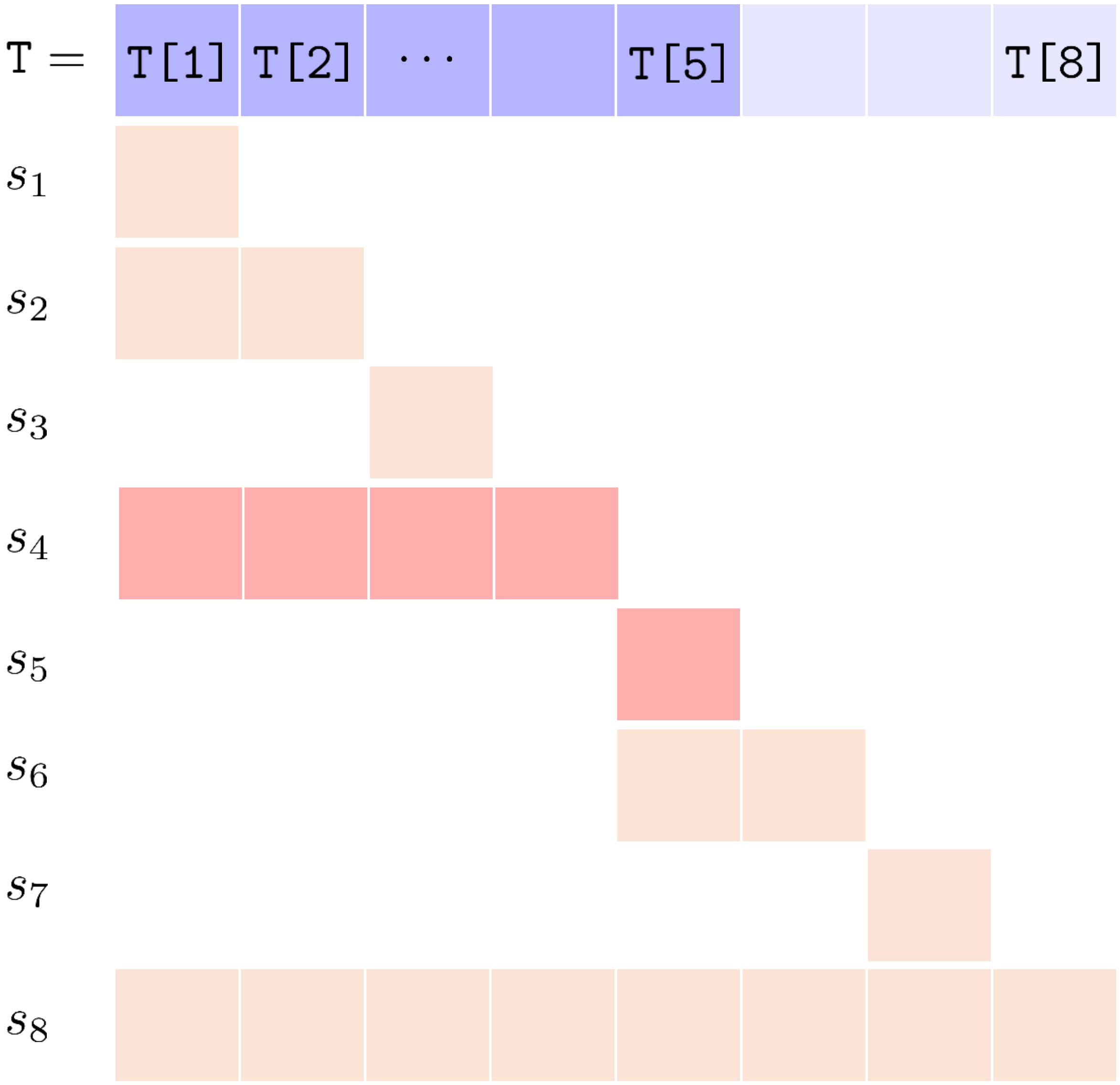
- A data structure saves array as partial sums, supporting two $O(\log n)$ operations.

Supplement 2: Fenwick structure



- A data structure saves array as partial sums, supporting two $O(\log n)$ operations.

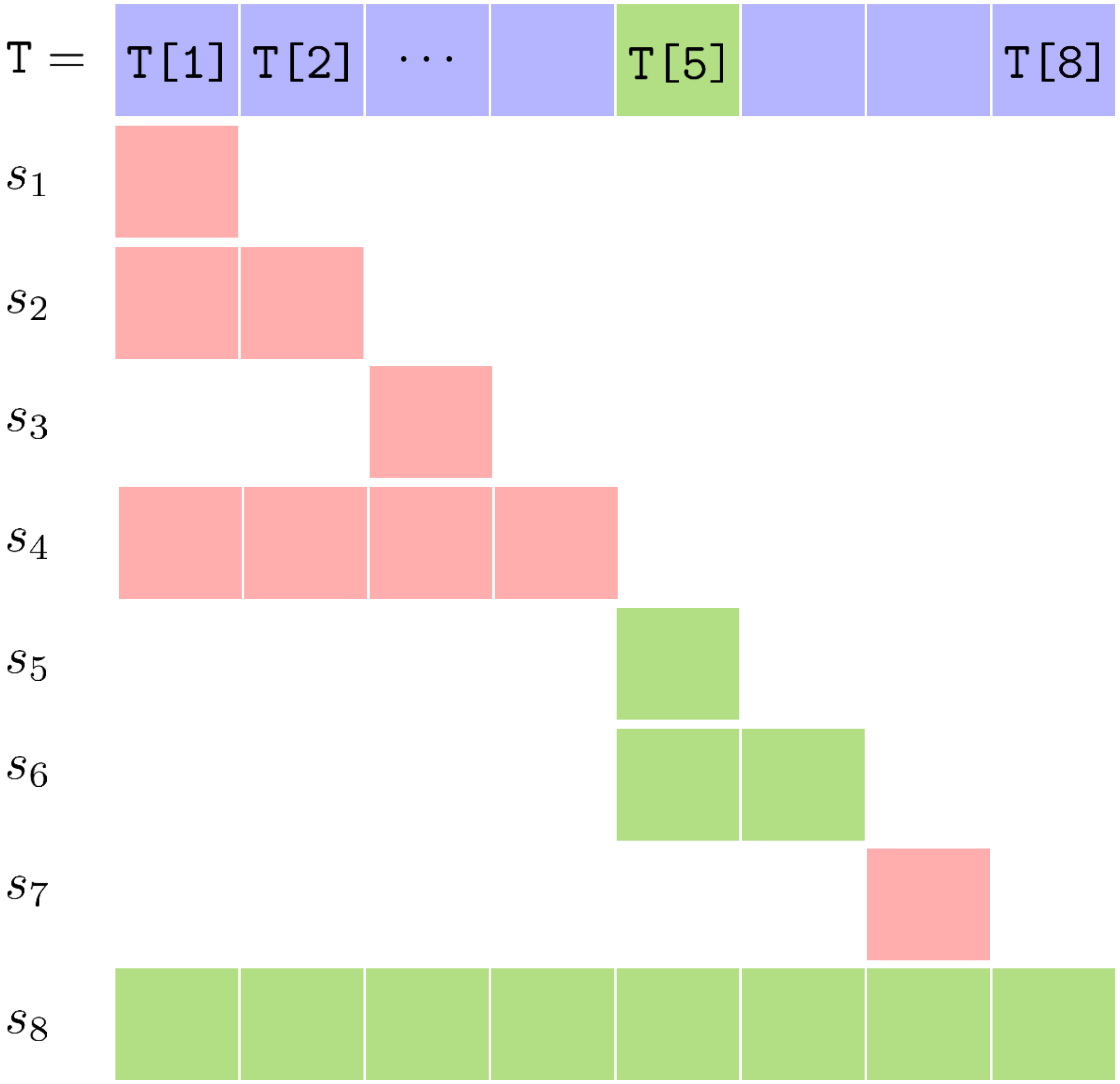
Supplement 2: Fenwick structure



- A data structure saves array as partial sums, supporting two $O(\log n)$ operations.
  1. query prefix sum:

$$\sum_{i=1}^{5} \texttt{T[i]} = s_4 + s_5$$

Supplement 2: Fenwick structure



- A data structure saves array as partial sums, supporting two $O(\log n)$ operations.
  1. query prefix sum:
  2. update element:

$$\texttt{T[5]} \leftarrow \texttt{T[5]+v}$$

$$s_5 \leftarrow s_5 + v$$

$$s_6 \leftarrow s_6 + v$$

$$s_8 \leftarrow s_8 + v$$

Supplement 2: Fenwick structure

Preprocess: Obtain the order of $Z_1, \ldots, Z_n$

Initialize: empty Fenwick array of length m

$$\sum_{1 \le i < t:\, Z_t > Z_i} e^{Z_i}$$

0	0	0	0	0	0	0	0

$(t-1)$- steps done

$t = 6$

e^{Z_4}	e^{Z_2}	0	e^{Z_1}	0	e^{Z_5}	e^{Z_3}	0

Each step,

1. get prefix sum
2. update element

within $O(\log n)$

Supplement 2: Fenwick structure

Preprocess: Obtain the order of $Z_1, \dots, Z_n$

Initialize: empty Fenwick array of length m

$$\boxed{\sum_{1 \le i < t:\, Z_t > Z_i} e^{Z_i}}$$

0	0	0	0	0	0	0	0

$(t-1)$- steps done

$t = 6$

e^{Z_4}	e^{Z_2}	0	e^{Z_1}	0	e^{Z_5}	e^{Z_3}	0

Suppose $\text{order}(Z_6) = 5$

Each step,

1. get prefix sum

$$\sum_{i: Z_6 > Z_i}^{6} e^{Z_i} = s_4 + s_5$$

Supplement 2: Fenwick structure

Preprocess: Obtain the order of $Z_1, \dots, Z_n$

Initialize: empty Fenwick array of length m

$$\sum_{1 \le i < t:\, Z_t > Z_i} e^{Z_i}$$

0	0	0	0	0	0	0	0

$(t-1)$- steps done

$t = 6$

e^{Z_4}	e^{Z_2}	0	e^{Z_1}	e^{Z_6}	e^{Z_5}	e^{Z_3}	0

Suppose $\text{order}(Z_6) = 5$

Each step,

1. get prefix sum
2. update element

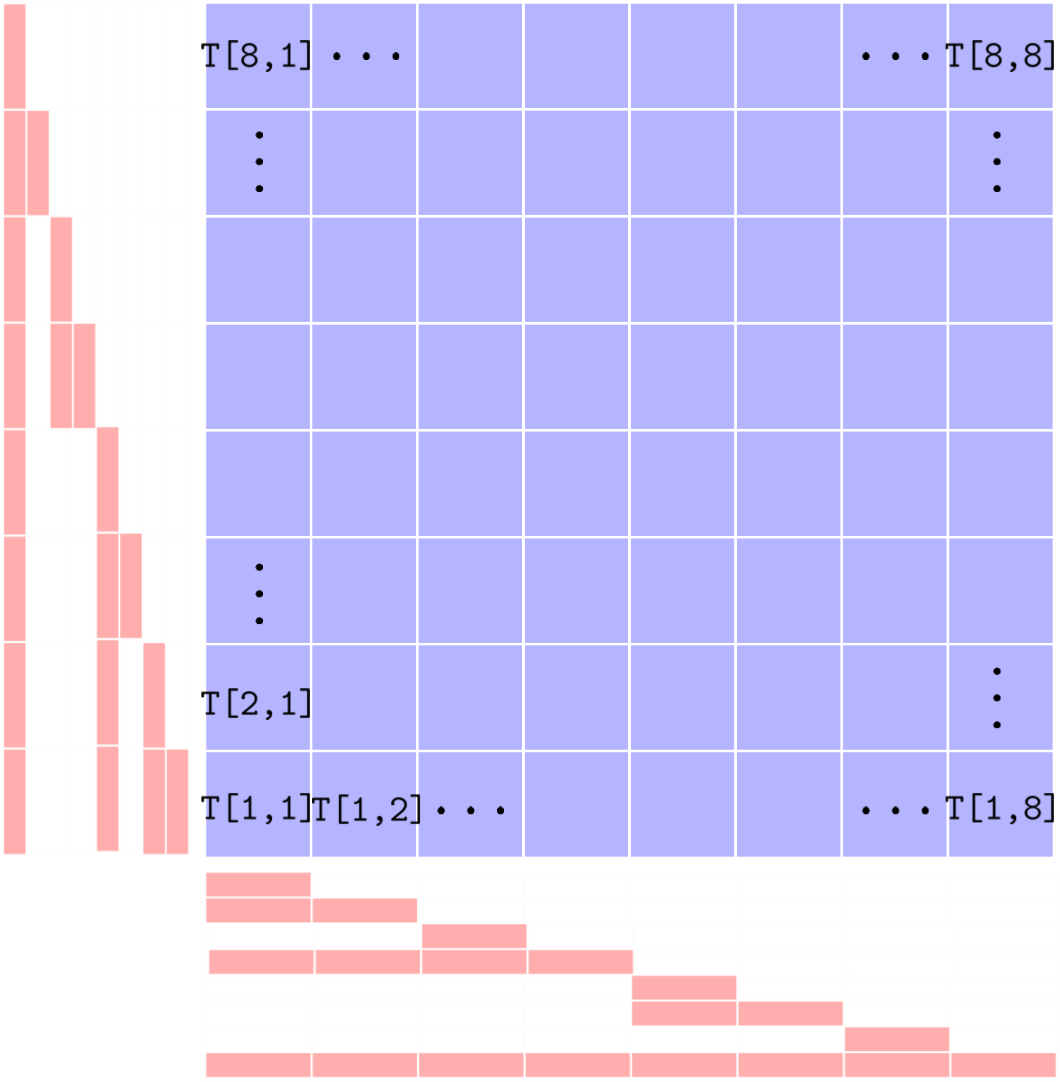
$$\texttt{T[5]} \leftarrow \texttt{T[5]} + e^{Z_6}$$

$$s_5 \leftarrow s_5 + e^{Z_6}$$

$$s_6 \leftarrow s_6 + e^{Z_6}$$

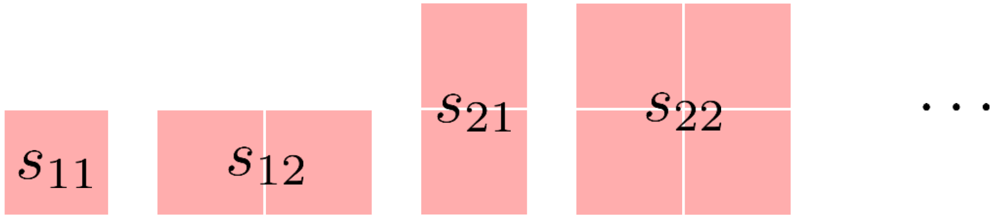
$$s_8 \leftarrow s_8 + e^{Z_6}$$

Supplement 3: multidimensional Fenwick

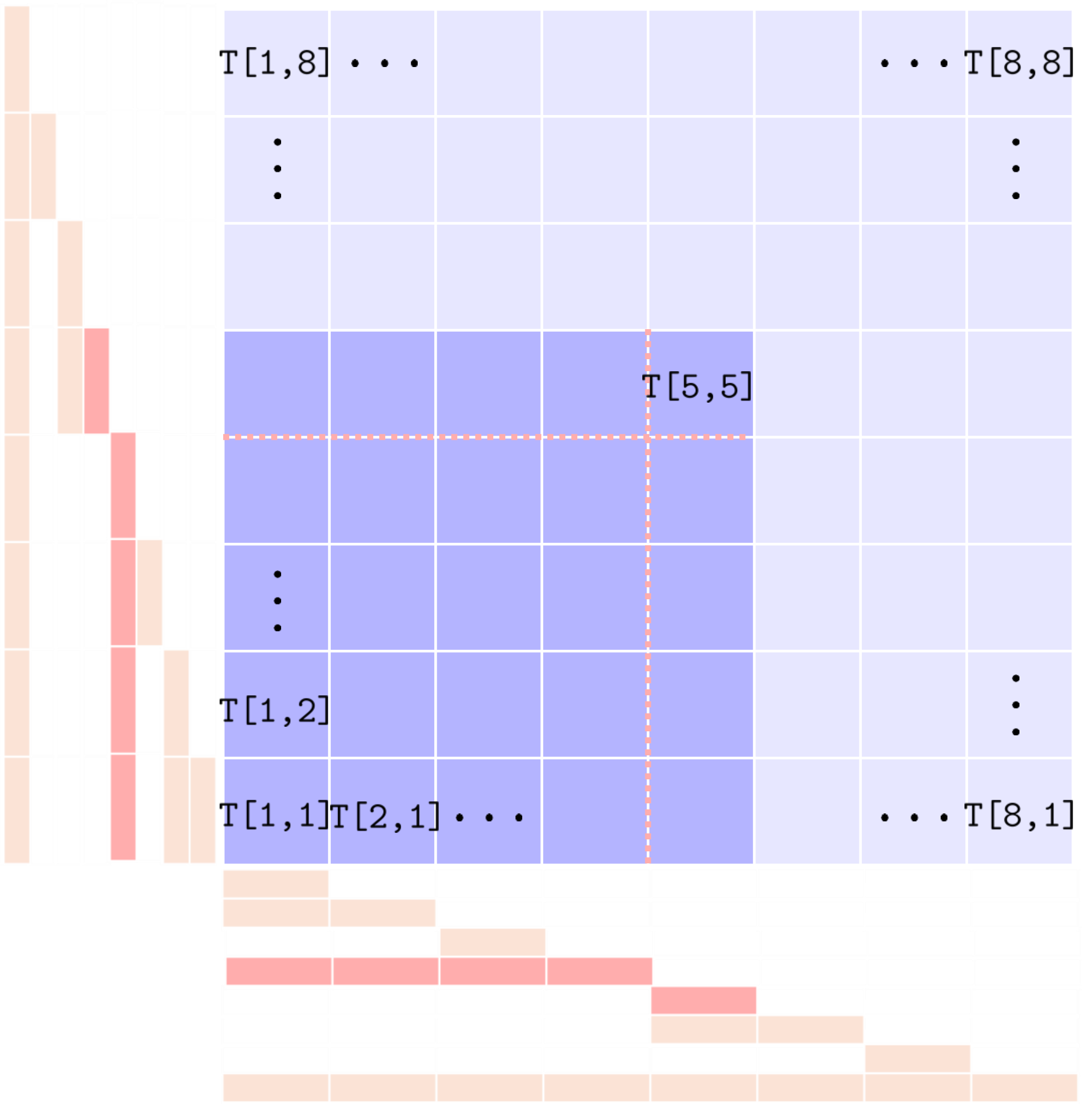


save array as box partial sums.

Update, prefix sum now take $O(\log^d n)$ time

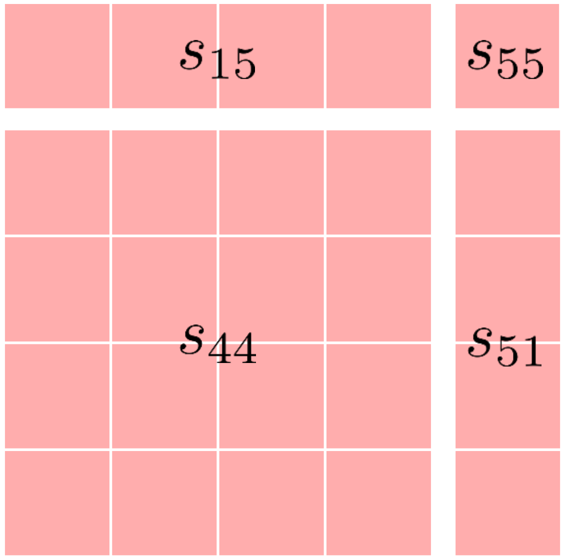


Supplement 3. multidimensional Fenwick

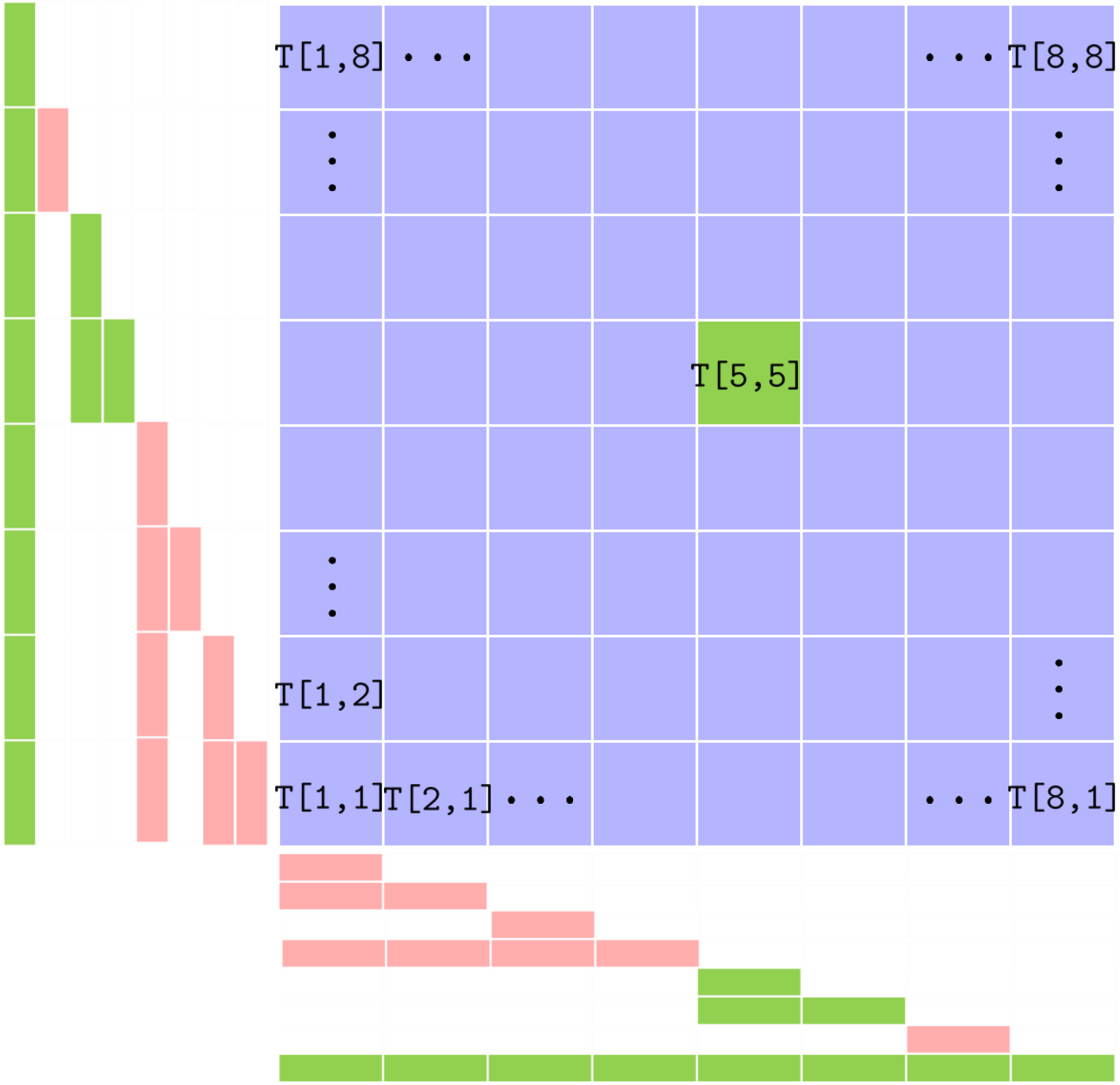


Prefix box sum:

$$\sum_{i,j=1}^{5} = \texttt{T[i,j]} = s_{44} + s_{51} + s_{15} + s_{55}$$



Supplement 3. multidimensional Fenwick



Update element:

$$\texttt{T[5,5]} \leftarrow \texttt{T[5,5]+v}$$

$$s_{58} \leftarrow s_{58} + v \quad s_{68} \leftarrow s_{68} + v \quad s_{88} \leftarrow s_{88} + v$$

$$s_{56} \leftarrow s_{56} + v \quad s_{66} \leftarrow s_{66} + v \quad s_{86} \leftarrow s_{86} + v$$

$$s_{55} \leftarrow s_{55} + v \quad s_{65} \leftarrow s_{65} + v \quad s_{85} \leftarrow s_{85} + v$$

Supplement 4. Sobolev ball details

$$p - q \in \mathcal{S}_d^s(\mathcal{R}),\ \|p\|_\infty, \|q\|_\infty \le M \qquad\qquad M < \infty$$

$$\mathcal{S}_d^s(\mathcal{R}) = \left\{ f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|\xi\|_2^{2s} |\hat{f}(\xi)|^2 d\xi \le (2\pi)^d \mathcal{R}^2 \right\}$$

$\hat{f}(\xi) = \int_{\mathbb{R}^d} f(x) e^{-ix^\top \xi} dx$: Fourier transform

$\mathcal{R}$: Radius of Sobolev ball $\qquad$ s : smoothness parameter

Choice of $\mathcal{R}_d, M_d$ for growing d testing: product of $N(0,1)$