Continuous Distributed Counting for Non-monotonic Streams

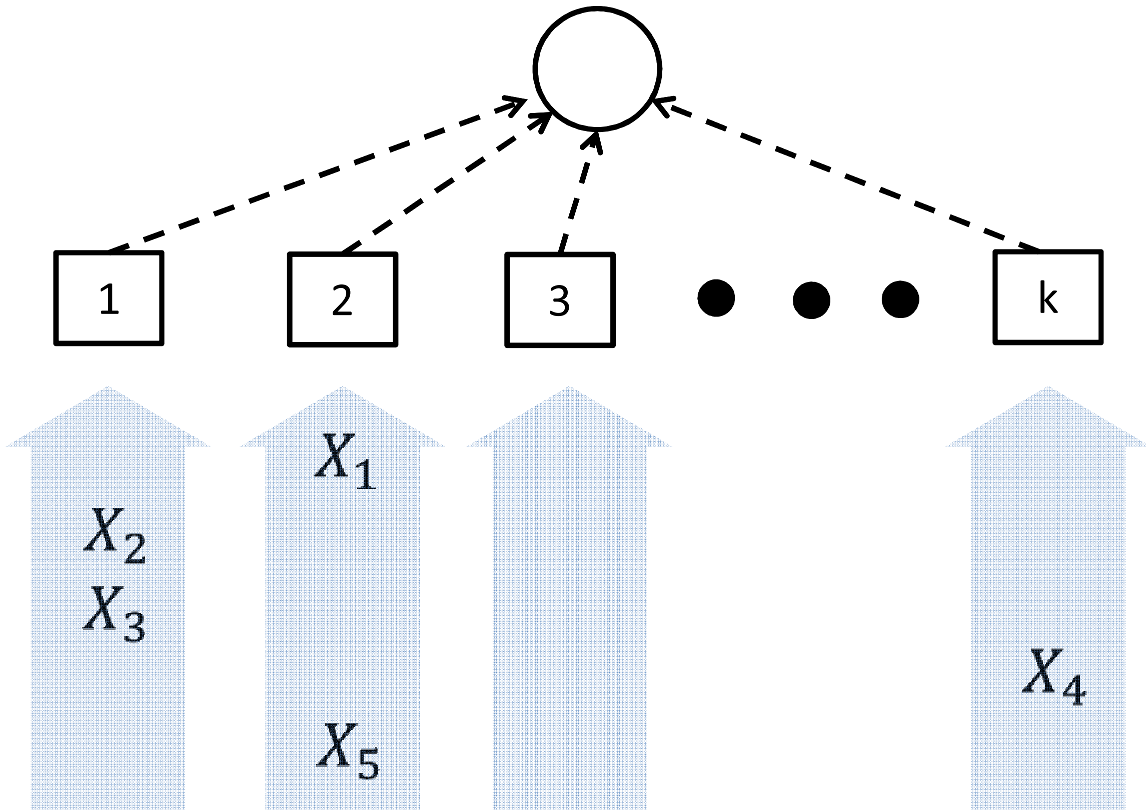
Milan Vojnović
Microsoft Research

Joint work with Zhenming Liu and Božidar Radunović

NII SHONAN MEETING January 14, 2012

SUM Tracking Problem

Track: $f(A): (1-\epsilon)S_t \leq \hat{S}_t \leq (1+\epsilon)S_t$



SUM: $S_t = \sum_{i \leq t} X_i$

SUM Tracking: Applications

- Ex 1: database queries

```
SELECT SUM(AdBids)from Ads
```

- Ex 2: iterative solving

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \gamma f(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$$

input data

Related Work

- Count tracking [Huang, Yi and Zhang, 2011]
  – Worst-case input, monotonic sum
  – Expected communication cost, for $k \leq 1/\epsilon^2$:
    $O\left(\frac{\sqrt{k}}{\epsilon}\log n\right)$ and lower bound $\Omega\left(\frac{\sqrt{k}}{\epsilon}\right)$

- Lower bound for worst case input
  [Arackaparambil, Brody and Chakrabarti, 2009]
  – Expected communication cost: $\Omega(\frac{n}{k})$

Questions

- Worst case complexity $\Omega(n)$

  Ex. +1, -1, +1, …

- Complexity under random input?
  - Random permutation
  - Random i.i.d.
  - Fractional Brownian motion

Outline

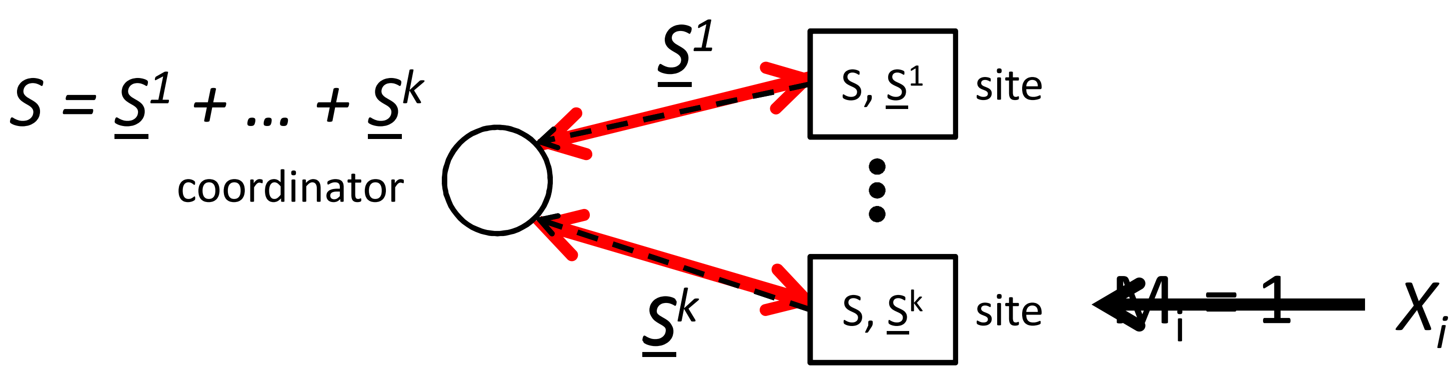
- Upper bounds
- Lower bounds
- Applications

Our Tracker Algorithm

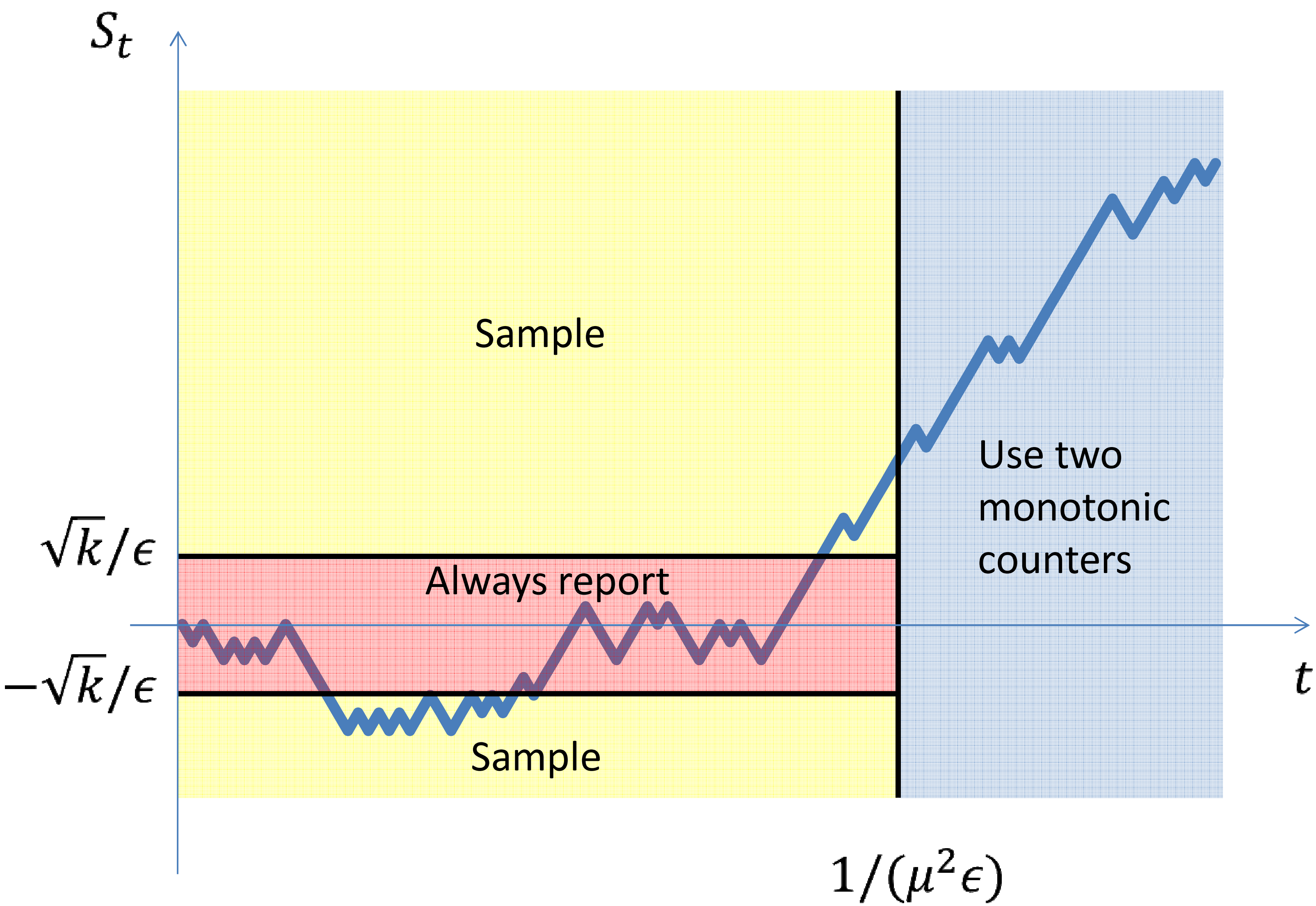
- Sampling based algorithm: upon arrival of update t, send a message to the coordinator w. p.

$$p_t = \min\left\{\frac{\alpha \log^{\beta} n}{(\epsilon S_t)^2}, 1\right\}$$

- If any site sends a message: sync all



Algorithm's Modes



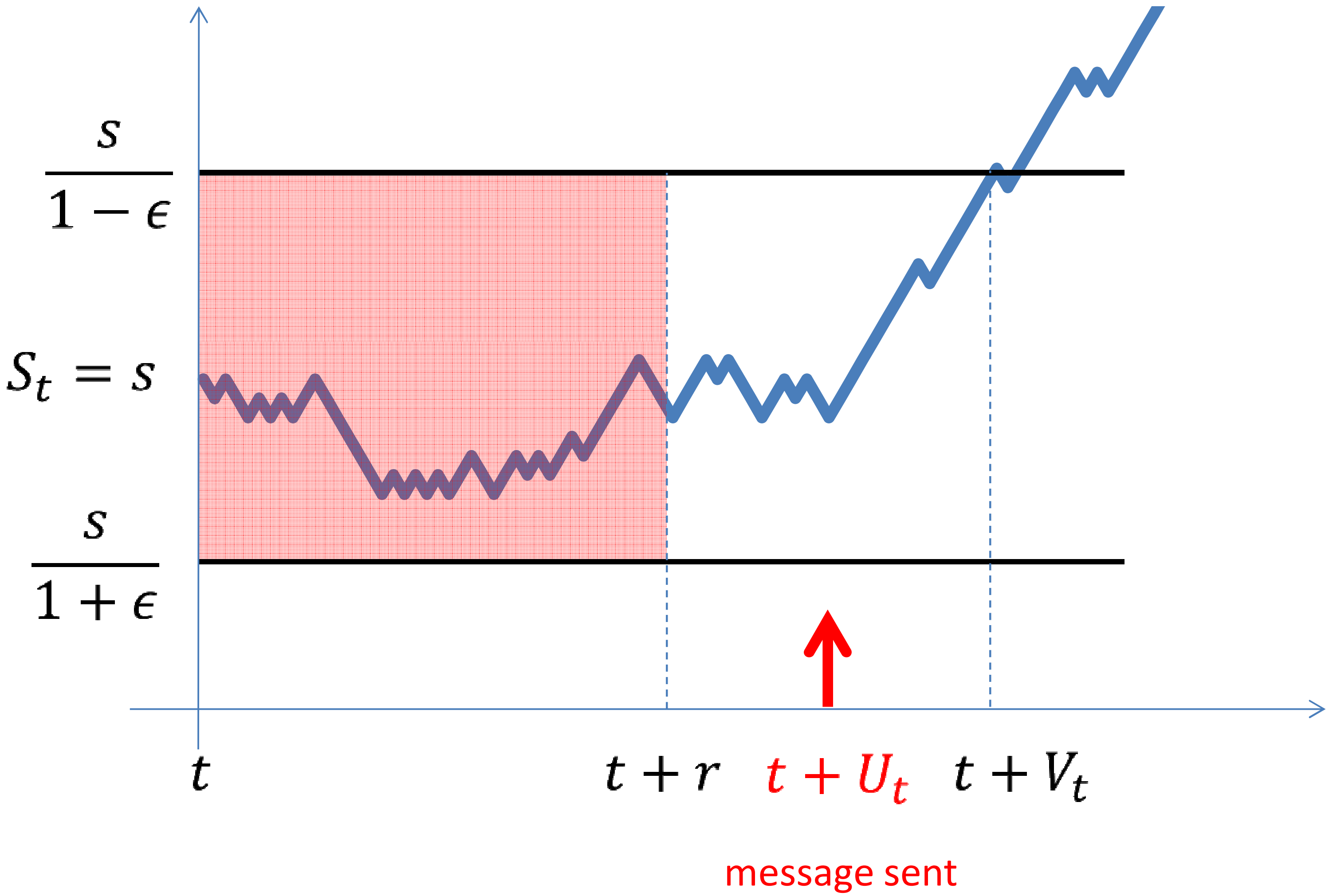
Communication Cost Upper Bound

Single Site

- Input: i.i.d. Bernoulli $\mathbf{P}[X_i = -1] = 1 - \mathbf{P}[X_i = 1] = \frac{1}{2}$
- Sampling probability with $\alpha > 9/2$ and $\beta = 2$

Expected communication cost: $\mathrm{O}(\min\{\frac{1}{\epsilon}\sqrt{n}\log n, n\})$

Proof Key Idea



Communication Cost Upper Bound

Multiple Sites

- k sites
- Updates i.i.d. Bernoulli $\mathbf{P}[X_i = -1] = 1 - \mathbf{P}[X_i = 1] = \frac{1}{2}$
- α large enough and $\beta = 2$

Expected communication cost: $O(\min\{\frac{\sqrt{k}}{\epsilon}\sqrt{n}\log n, n\})$

Communication Cost Upper Bound

Unknown Drift Case

- Input: i.i.d. Bernoulli

$$\mathbf{P}[X_i = -1] = 1 - \mathbf{P}[X_i = 1] = \frac{1+\mu}{2}$$

- $\mu \in [-1,1]$: unknown drift parameter

Expected communication cost:

$$\tilde{O}\left(\frac{\sqrt{k}}{\epsilon}\min\{1/|\mu|, \sqrt{n}\}\right)$$

Communication Cost Upper Bound

Random Permutation Input

- Input: a random permutation of values $a_1, a_2, \dots, a_n$
- α sufficiently large and $\beta = 2$

Expected communication cost:

$$O(\frac{\sqrt{k}}{\epsilon}\sqrt{n}\log n)$$

Communication Cost Upper Bound

Fractional Brownian Motion

- Input: a fractional Brownian motion with Hurst parameter
$$\frac{1}{2} \leq H < 1/\delta$$
- $\mathrm{Sample{-}prob}(S_t, t) = \min\left\{\frac{\alpha_\delta \log^{1+\frac{\delta}{2}} n}{(\epsilon|S_t|)^\delta}, 1\right\}$

Expected communication cost: $O(\min\{\frac{k^{\frac{3-\delta}{2}}}{\epsilon} n^{1-H}, n\})$

Outline

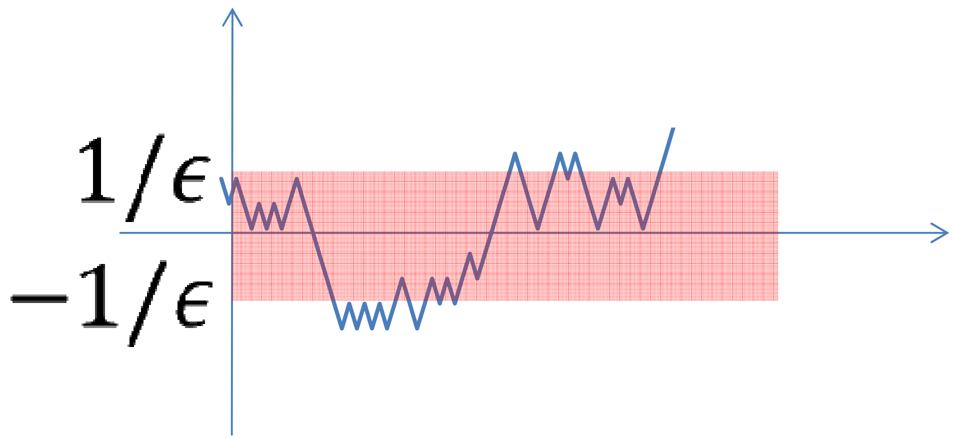
- Upper bounds
- Lower bounds
- Applications

Lower Bounds

Single Site, Zero Drift

- Input: i.i.d. Bernoulli $\mathbf{P}[X_i = -1] = 1 - \mathbf{P}[X_i = 1] = \frac{1}{2}$

Expected communication cost: $\Omega(\min\{\frac{1}{\epsilon}\sqrt{n}, n\})$



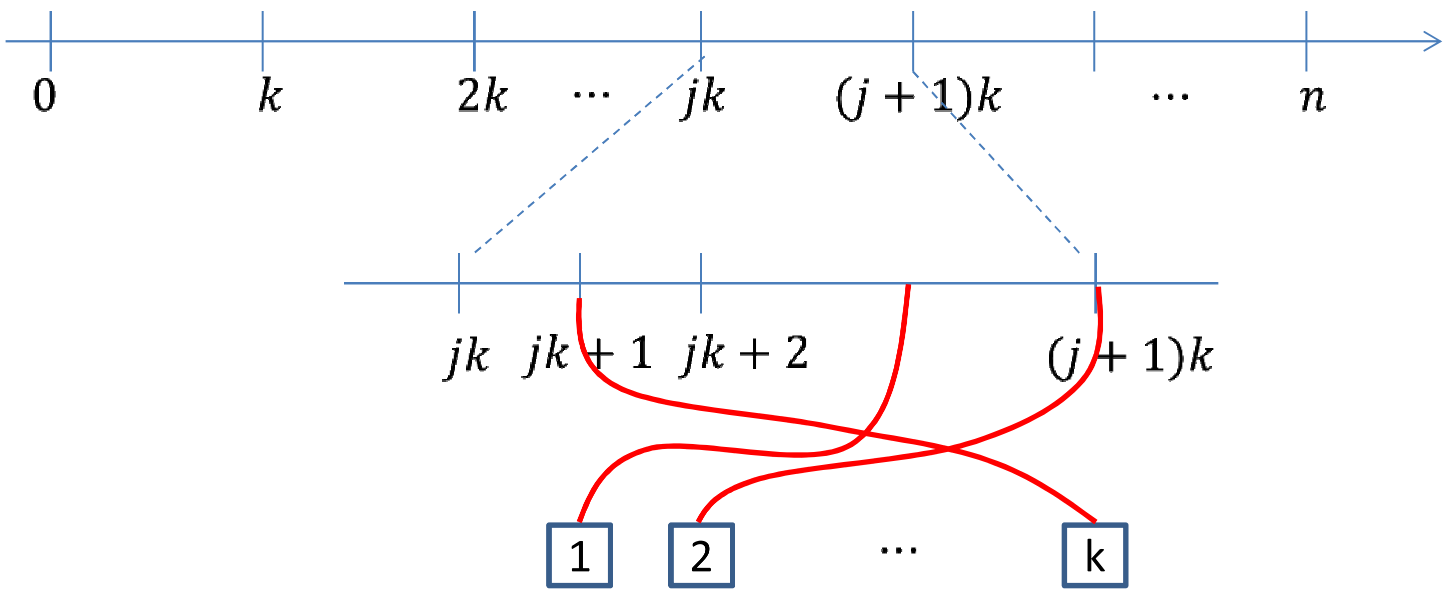
Lower Bounds

Multiple Sites

- Input: i.i.d. Bernoulli $\mathbf{P}[X_i = -1] = 1 - \mathbf{P}[X_i = 1] = \frac{1}{2}$ or a random permutation

Expected communication cost: $\Omega(\min\{\frac{\sqrt{k}}{\epsilon}\sqrt{n}, n\})$

Proof Key Ideas



- $I_j = I\left(S_{kj} \in \left[-\min\left\{\frac{\sqrt{k}}{\epsilon}, \sqrt{jk}\right\}, \min\left\{\frac{\sqrt{k}}{\epsilon}, \sqrt{jk}\right\}\right]\right)$
- Under $I_j = 1$, maximum deviation $\epsilon\left|S_{jk}\right| \leq \sqrt{k}$

Proof Key Ideas (cont'd)

k-input problem:

$$\begin{array}{cccc} \boxed{1} & \boxed{2} & \cdots & \boxed{\mathrm{k}} \\ \uparrow & \uparrow & & \uparrow \\ X_1 & X_2 & & X_k \end{array} \quad \sim \text{i. i. d. } \mathbf{Pr}[X_i = -1] = \mathbf{Pr}[X_i = 1] = \frac{1}{2}$$

- Query: $\mathrm{H}_0 \colon \sum_i X_i > \sqrt{k}$ or $\mathrm{H}_1 \colon \sum_i X_i < -\sqrt{k}$?
- Answer: incorrect only if $|\sum_i X_i| > \sqrt{k}$ and the answer is $\sum_i X_i > \sqrt{k}$ under H_1 or $\sum_i X_i < -\sqrt{k}$ under H_0
- Lemma: $m_k = \Omega(k)$ messages is necessary to answer the query correctly with a constant positive probability

Outline

- Upper bounds
- Lower bounds
- Applications

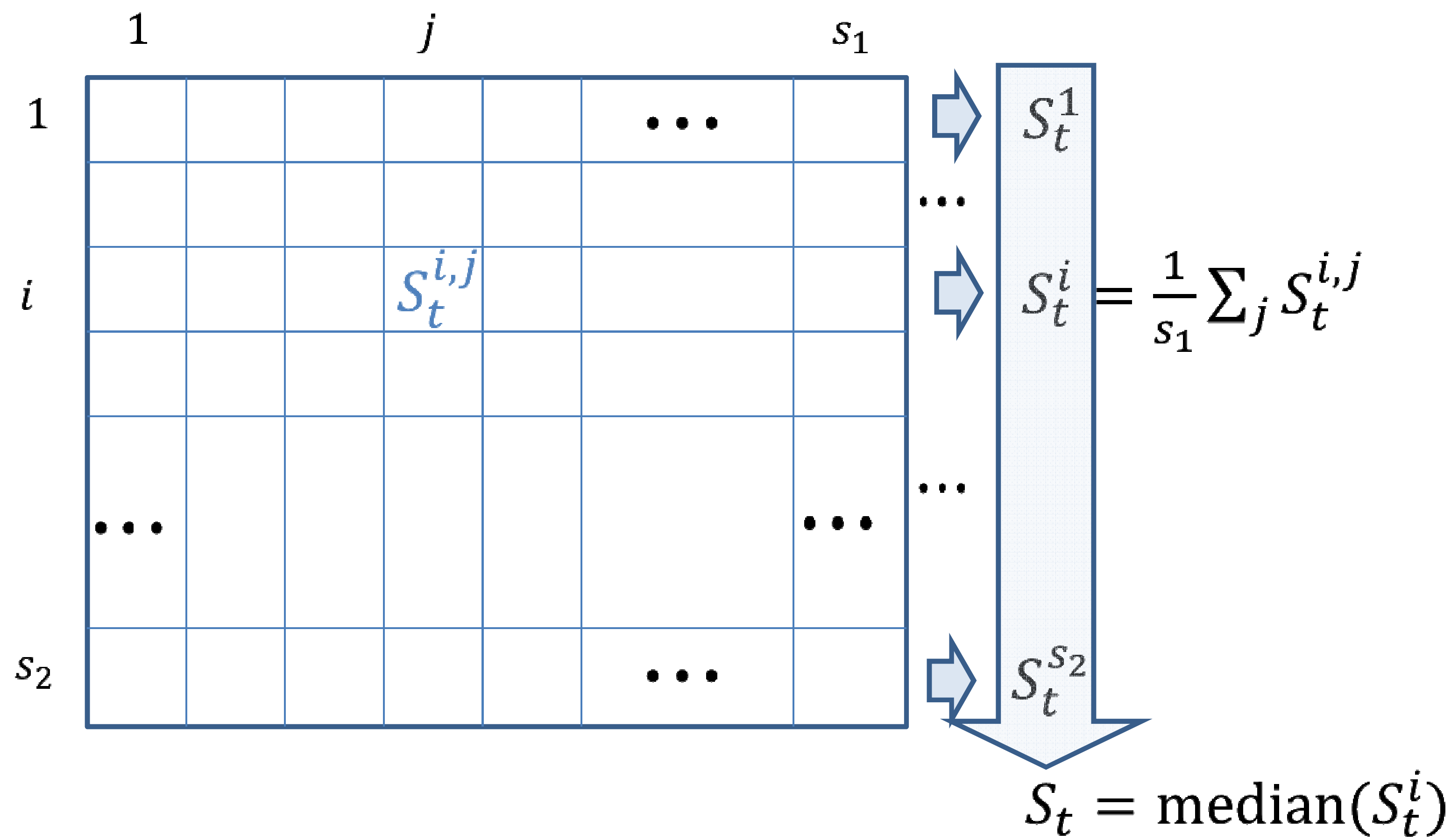
App 1: F_2 Tracking (cont'd)

- Input: random permutation of $a_1, a_2, \dots, a_n$
- $a_t = (\alpha_t, z_t), \alpha_t \in [m], z_t \in \{-1,1\}$

- $m_i(t) = \sum_{s \le t : \alpha_s = i} z_s$
- $F_2(t) = \sum_{i \in [m]} m_i^2(t)$

- Problem: track $F_2(t)$ within a prescribed relative tolerance $\epsilon > 0$ with high probability

AMS Sketch



- $h: [m] \rightarrow \{-1,1\}$, 4-wise independent hash function
- $S_t^{i,j} = \sum_{s \le t} z_s h(\alpha_s) = \sum_{a \in [m]} h(a) m_a(t)$

App 1: F_2 tracking (cont'd)

- AMS: S_t within $(1 \pm \epsilon)F_2(t)$ w. p. $\geq 1 - \delta$ using $s_1 = \frac{16}{\epsilon^2}$ and $s_2 = 2\log\left(\frac{1}{\delta}\right)$
- Sum tracking: $S_{t+1}^{i,j} = S_t^{i,j} + z_t h(\alpha_t)$

Expected total communication:

$\Omega\left(\min\{\frac{\sqrt{k}}{\epsilon}\sqrt{n}, n\}\right)$ $\tilde{O}\left(\min\{\frac{\sqrt{k}}{\epsilon^3}\sqrt{n}, n\}\right)$

App 2: Bayesian Linear Regression

- Feature vector $\boldsymbol{x}_t \in R^d$, output $y_t \in R$
- $y_t = \boldsymbol{w}^T \boldsymbol{A}_t + N(0, \beta^{-1})$, $\boldsymbol{A}_t = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_t)^T$
- Prior $\boldsymbol{w} \sim N(\boldsymbol{m}_0, \boldsymbol{S}_0)$, posterior $\boldsymbol{w} \sim N(\boldsymbol{m}_t, \boldsymbol{S}_t)$

$$\boldsymbol{m}_t = \boldsymbol{S}_t(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta \boldsymbol{A}_t^T \boldsymbol{y}_t)$$
$$\boldsymbol{S}_t^{-1} = \boldsymbol{S}_0^{-1} + \beta \boldsymbol{A}_t^T \boldsymbol{A}_t$$

- Sum tracking: $\boldsymbol{S}_{t+1}^{-1} = \boldsymbol{S}_t^{-1} + \beta \boldsymbol{x}_{t+1}^T \boldsymbol{x}_{t+1}$
- Under random permutation input, the expected communication cost = $\mathrm{O}(d^2 \min\{\frac{\sqrt{k}}{\epsilon}\sqrt{n}\log n, n\})$

Summary

- We considered the sum tracking problem with non-monotonic distributed streams under random permutation, random i. i. d. and fractional Brownian motion
- Derived a practical algorithm that has order optimal communication complexity