Continuous Distributed Counting for Non-monotonic Streams

Milan Vojnović
Microsoft Research

Joint work with Zhenming Liu and Božidar Radunović

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

    \[ x_{t+1} = x_t + \gamma f(x_t, \xi_t) \]
Related Work

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

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

$$S = S^1 + \ldots + S^k$$

$S^1 \quad S^1$ 

site 

$\cdots$ 

$S^k \quad S^k$ 

site 

$X_i = 1$
Algorithm’s Modes

Sample

Use two monotonic counters

$S_t$

$\frac{\sqrt{k}}{\epsilon}$

$-\frac{\sqrt{k}}{\epsilon}$

$1/(\mu^2 \epsilon)$

Sample

Always report
Communication Cost Upper Bound

Single Site

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

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

\[
\frac{s}{1 - \epsilon} \quad \frac{s}{1 + \epsilon}
\]

\[S_t = s\]

message sent
Communication Cost Upper Bound

Multiple Sites

- $k$ sites
- Updates i.i.d. Bernoulli $P[X_i = -1] = 1 - P[X_i = 1] = \frac{1}{2}$
- $\alpha$ 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
  \[ P[X_i = -1] = 1 - P[X_i = 1] = \frac{1 + \mu}{2} \]
- \( \mu \in [-1,1] : \) 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 \( a_1, a_2, \ldots, a_n \)
- \( \alpha \) sufficiently large and \( \beta = 2 \)

Expected communication cost:

\[
O\left(\frac{\sqrt{k}}{\epsilon} \sqrt{n \log n}\right)
\]
Communication Cost Upper Bound

Fractional Brownian Motion

- Input: a fractional Brownian motion with Hurst parameter
  \[ \frac{1}{2} \leq H < \frac{1}{\delta} \]

- Sample probability function:
  \[ \text{Sample–prob}(S_t, t) = \min \left\{ \frac{a_\delta \log^{1+\frac{\delta}{2}} n}{(\epsilon |S_t|)^\delta}, 1 \right\} \]

Expected communication cost:
\[ O(\min\left\{ \frac{k^{3-\delta}}{\epsilon} n^{1-H}, n \right\}) \]
Outline

• Upper bounds

• Lower bounds

• Applications
Lower Bounds
Single Site, Zero Drift

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

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

Multiple Sites

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

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

\[ I_j = I(S_{kj} \in \left[ -\min\left\{ \frac{\sqrt{k}}{\epsilon}, \sqrt{jk} \right\}, \min\left\{ \frac{\sqrt{k}}{\epsilon}, \sqrt{jk} \right\} \right]) \]

- Under \( I_j = 1 \), maximum deviation \( \epsilon |S_{jk}| \leq \sqrt{k} \)
Proof Key Ideas (cont’d)

k-input problem:

\[
\begin{align*}
X_1 & \quad \cdots \quad X_k \\
1 & \quad 2 & \quad \cdots & \quad k
\end{align*}
\]

- \text{i. i. d. } \Pr[X_i = -1] = \Pr[X_i = 1] = \frac{1}{2}

- Query: \(H_0: \Sigma_i X_i > \sqrt{k}\) or \(H_1: \Sigma_i X_i < -\sqrt{k}\)?

- Answer: incorrect only if \(|\Sigma_i X_i| > \sqrt{k}\) and the answer is \(\Sigma_i X_i > \sqrt{k}\) under \(H_1\) or \(\Sigma_i X_i < -\sqrt{k}\) 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 $a_1, a_2, \ldots, a_n$
- $a_t = (\alpha_t, z_t), \alpha_t \in [m], z_t \in \{-1,1\}$

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

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

\[ S_t^i = \frac{1}{s_1} \sum_j S_t^{i,j} \]

\[ S_t = \text{median}(S_t^i) \]

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

$$\Omega \left( \min \{ \frac{\sqrt{k}}{\epsilon} \sqrt{n}, n \} \right) \quad \tilde{\Omega} \left( \min \{ \frac{\sqrt{k}}{\epsilon^3 \sqrt{n}}, n \} \right)$$
App 2: Bayesian Linear Regression

- Feature vector $x_t \in \mathbb{R}^d$, output $y_t \in \mathbb{R}$

- $y_t = w^T A_t + N(0, \beta^{-1}), \quad A_t = (x_1, \ldots, x_t)^T$

- Prior $w \sim N(m_0, S_0)$, posterior $w \sim N(m_t, S_t)$

\[
\begin{align*}
m_t & = S_t(S_0^{-1}m_0 + \beta A_t^T y_t) \\
S_t^{-1} & = S_0^{-1} + \beta A_t^T A_t
\end{align*}
\]

- Sum tracking: $S_{t+1}^{-1} = S_t^{-1} + \beta x_{t+1}^T x_{t+1}$

- Under random permutation input, the expected communication cost $= O(d^2 \min\{\sqrt{k}, \sqrt{n \log n}, n\})$
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