Lower Bounds for Gap-Hamming-Distance and Consequences for Data Stream Algorithms

Joshua Brody and Amit Chakrabarti

DARTMOUTH COLLEGE

$24th$ CCC, 2009, Paris

Counting Distinct Elements in a Data Stream

-----	3	14	1	3	9	9	4	2	1	5	2	3	6	-----

Input: Stream of integers $\sigma =< a_1, \ldots, a_m >$

Output: F_0 := number of distinct elements in σ

Counting Distinct Elements in a Data Stream

-----	3	14	1	3	9	9	4	2	1	5	2	3	6	-----

Input: Stream of integers $\sigma = < a_1, \ldots, a_m >$

Output: F_0 := number of distinct elements in σ

Goal: Minimize space used to compute F_0

Previous Streaming Results

Frequency Moments: $F_k = \sum_{i=1}^{n} \mathrm{freq}(i)^k$ [Alon-Matias-Szegedy '96]

Previous Streaming Results

Frequency Moments: $F_k = \sum_{i=1}^{n} \mathrm{freq}(i)^k$ [Alon-Matias-Szegedy '96]

- $\Omega(n)$ space unless randomization and approximation used
- Upper, lower bounds for randomized algorithms that approximate F_k
- Spawned lots of research, won 2005 Gödel Prize

Previous Streaming Results

Frequency Moments: $F_k = \sum_{i=1}^{n} \mathrm{freq}(i)^k$ [Alon-Matias-Szegedy '96]

- $\Omega(n)$ space unless randomization and approximation used
- Upper, lower bounds for randomized algorithms that approximate F_k
- Spawned lots of research, won 2005 Gödel Prize

One-pass, randomized, ε-approximate: $\left|\frac{\text{output}}{\text{answer}} - 1\right| \leq \varepsilon$

Previous Streaming Results

Frequency Moments: $F_k = \sum_{i=1}^{n} \mathrm{freq}(i)^k$ [Alon-Matias-Szegedy '96]

- $\Omega(n)$ space unless randomization and approximation used
- Upper, lower bounds for randomized algorithms that approximate F_k
- Spawned lots of research, won 2005 Gödel Prize

One-pass, randomized, ε-approximate: $\left|\frac{\text{output}}{\text{answer}} - 1\right| \leq \varepsilon$

Status as of Jan 2009:

- Space upper bound: $\widetilde{O}(\varepsilon^{-2})$
- Space lower bound: $\widetilde{\Omega}(\varepsilon^{-2})$
- Also hold for other problems, e.g. empirical entropy

Do multiple passes help?

Previous Streaming Results

Frequency Moments: $F_k = \sum_{i=1}^{n} \mathrm{freq}(i)^k$ [Alon-Matias-Szegedy '96]

- $\Omega(n)$ space unless randomization and approximation used
- Upper, lower bounds for randomized algorithms that approximate F_k
- Spawned lots of research, won 2005 Gödel Prize

One-pass, randomized, ε-approximate: $\left|\frac{\text{output}}{\text{answer}} - 1\right| \leq \varepsilon$

Status as of Jan 2009:

- Space upper bound: $\widetilde{O}(\varepsilon^{-2})$
- Space lower bound: $\widetilde{\Omega}(\varepsilon^{-2})$
- Also hold for other problems, e.g. empirical entropy

Do multiple passes help? If not, why not?

The Gap-Hamming-Distance Problem

Input: Alice gets $x \in \{0,1\}^n$, Bob gets $y \in \{0,1\}^n$.

Output:

- $\text{GHD}(x,y) = 1$ if $\Delta(x,y) > \frac{n}{2} + \sqrt{n}$
- $\text{GHD}(x,y) = 0$ if $\Delta(x,y) < \frac{n}{2} - \sqrt{n}$

Problem: Design randomized, constant error protocol to solve this

Cost: Worst case number of bits communicated

$x =$	**0**	**1**	**0**	**0**	**1**	**0**	**1**	**1**	**0**	**0**	**0**	**1**
$y =$	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**0**	**0**	**1**

$$n = 12; \quad \Delta(x,y) = 3 \in [6 - \sqrt{12}, 6 + \sqrt{12}]$$

The Reductions

E.g., Distinct Elements (Other problems: similar)

$x =$	**0**	**1**	**0**	**0**	**1**	**0**	**1**	**1**	**0**	**0**	**0**	**1**
σ :	(1,0)	(2,1)	(3,0)	(4,0)	(5,1)	(6,0)	(9,0)	(8,0)	(9,0)	(10,0)	(11,0)	(12,1)
$y =$	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**0**	**0**	**1**
τ :	(1,0)	(2,0)	(3,0)	(4,0)	(5,0)	(6,0)	(9,0)	(8,0)	(9,1)	(10,0)	(11,0)	(12,1)

Alice: $x \longmapsto \sigma = \langle (1, x_1), (2, x_2), \ldots, (n, x_n) \rangle$

Bob: $y \longmapsto \tau = \langle (1, y_1), (2, y_2), \ldots, (n, y_n) \rangle$

Notice: $F_0(\sigma \circ \tau) = n + \Delta(x, y) = \begin{cases} < \frac{3n}{2} - \sqrt{n}, & \text{or} \\ > \frac{3n}{2} + \sqrt{n}. & \end{cases}$ Set $\varepsilon = \frac{1}{\sqrt{n}}$.

Communication to Streaming

p-pass streaming algorithm $\Longrightarrow$ $(2p-1)$-round communication protocol

messages $=$ memory contents of streaming algorithm

And Thus

Previous results [Indyk-Woodruff'03], [Woodruff'04], [C.-Cormode-McGregor'07]:

- For one-round protocols, $\mathrm{R}^{\rightarrow}(\textsc{ghd}) = \Omega(n)$
- Implies the $\widetilde{\Omega}(\varepsilon^{-2})$ streaming lower bounds

Communication to Streaming

p-pass streaming algorithm $\Longrightarrow$ $(2p-1)$-round communication protocol

messages $=$ memory contents of streaming algorithm

And Thus

Previous results [Indyk-Woodruff'03], [Woodruff'04], [C.-Cormode-McGregor'07]:

- For one-round protocols, $\mathrm{R}^{\rightarrow}(\textsc{ghd}) = \Omega(n)$
- Implies the $\widetilde{\Omega}(\varepsilon^{-2})$ streaming lower bounds

Key open questions:

- What is the unrestricted randomized complexity $\mathrm{R}(\textsc{ghd})$?
- Better algorithm for Distinct Elements (or F_k, or H) using two passes?

Our Results

Previous Results (Communication):

- One-round (one-way) lower bound: $\mathrm{R}^{\rightarrow}(\textsc{ghd}) = \Omega(n)$ [Woodruff'04]
- Simplification, clever reduction from $\textsc{index}$ [Jayram-Kumar-Sivakumar]

- Multi-round case: $\mathrm{R}(\textsc{ghd}) = \Omega(\sqrt{n})$ [Folklore]

Our Results

Previous Results (Communication):

- One-round (one-way) lower bound: $\mathrm{R}^{\rightarrow}(\textsc{ghd}) = \Omega(n)$ [Woodruff'04]
- Simplification, clever reduction from $\textsc{index}$ [Jayram-Kumar-Sivakumar]

  Hard distribution "contrived," non-uniform
- Multi-round case: $\mathrm{R}(\textsc{ghd}) = \Omega(\sqrt{n})$ [Folklore]

Our Results

Previous Results (Communication):

- One-round (one-way) lower bound: $\mathrm{R}^{\rightarrow}(\textsc{ghd}) = \Omega(n)$ [Woodruff'04]
- Simplification, clever reduction from $\textsc{index}$ [Jayram-Kumar-Sivakumar]

  Hard distribution "contrived," non-uniform
- Multi-round case: $\mathrm{R}(\textsc{ghd}) = \Omega(\sqrt{n})$ [Folklore]

  Reduction from $\textsc{disjointness}$ using "repetition code"

  Hard distribution again far from uniform

Our Results

Previous Results (Communication):

- One-round (one-way) lower bound: $\mathrm{R}^{\rightarrow}(\textsc{ghd}) = \Omega(n)$ [Woodruff'04]
- Simplification, clever reduction from $\textsc{index}$ [Jayram-Kumar-Sivakumar]

  Hard distribution "contrived," non-uniform

- Multi-round case: $\mathrm{R}(\textsc{ghd}) = \Omega(\sqrt{n})$ [Folklore]

  Reduction from $\textsc{disjointness}$ using "repetition code"

  Hard distribution again far from uniform

What we show:

- Theorem 1: $\Omega(n)$ lower bound for any $O(1)$-round protocol

  Holds under uniform distribution

Our Results

Previous Results (Communication):

- One-round (one-way) lower bound: $\mathrm{R}^{\to}(\textsc{ghd}) = \Omega(n)$ [Woodruff'04]
- Simplification, clever reduction from $\textsc{index}$ [Jayram-Kumar-Sivakumar]

  Hard distribution "contrived," non-uniform
- Multi-round case: $\mathrm{R}(\textsc{ghd}) = \Omega(\sqrt{n})$ [Folklore]

  Reduction from $\textsc{disjointness}$ using "repetition code"

  Hard distribution again far from uniform

What we show:

- Theorem 1: $\Omega(n)$ lower bound for any $O(1)$-round protocol

  Holds under uniform distribution
- Theorem 2: one-round, deterministic: $\mathrm{D}^{\to}(\textsc{ghd}) = n - \Theta(\sqrt{n}\log n)$
- Theorem 3: $\mathrm{R}^{\to}(\textsc{ghd}) = \Omega(n)$ (simpler proof, uniform distrib)

  (independently proved by [Woodruff'09])

Technique: Round Elimination

Base Case Lemma: There is no "nice" 0-round GHD protocol.

Round Elimination Lemma: If there is a "nice" k-round GHD protocol, then there is a "nice" $(k-1)$-round GHD protocol.

Technique: Round Elimination

Base Case Lemma: There is no 0-round $\textsc{ghd}$ protocol with error $\varepsilon < \frac{1}{2}$.

Round Elimination Lemma: If there is a "nice" k-round $\textsc{ghd}$ protocol, then there is a "nice" $(k-1)$-round $\textsc{ghd}'$ protocol.

Technique: Round Elimination

Base Case Lemma: There is no 0-round $\textsc{ghd}$ protocol with error $\varepsilon < \frac{1}{2}$.

Round Elimination Lemma: If there is a "nice" k-round $\textsc{ghd}$ protocol, then there is a "nice" $(k-1)$-round $\textsc{ghd}'$ protocol.

- The $(k-1)$-round protocol will be solving a "simpler" problem
- Parameters degrade with each round elimination step

Parametrized Gap-Hamming-Distance Problem

The problem:

$$\mathrm{GHD}_{c,n}(x,y) = \begin{cases} 1\,, & \text{if } \Delta(x,y) \geq n/2 + c\sqrt{n}\,, \\ 0\,, & \text{if } \Delta(x,y) \leq n/2 - c\sqrt{n}\,, \\ \star\,, & \text{otherwise.} \end{cases}$$

Parametrized Gap-Hamming-Distance Problem

The problem:

$$\mathrm{GHD}_{c,n}(x,y) = \begin{cases} 1\,, & \text{if } \Delta(x,y) \geq n/2 + c\sqrt{n}\,, \\ 0\,, & \text{if } \Delta(x,y) \leq n/2 - c\sqrt{n}\,, \\ \star\,, & \text{otherwise.} \end{cases}$$

Hard input distribution:

$\mu_{c,n}$: uniform over (x,y) such that $|\Delta(x,y) - n/2| \geq c\sqrt{n}$

Parametrized Gap-Hamming-Distance Problem

The problem:

$$\mathrm{GHD}_{c,n}(x,y) = \begin{cases} 1\,, & \text{if } \Delta(x,y) \geq n/2 + c\sqrt{n}\,, \\ 0\,, & \text{if } \Delta(x,y) \leq n/2 - c\sqrt{n}\,, \\ \star\,, & \text{otherwise.} \end{cases}$$

Hard input distribution:

$\mu_{c,n}$: uniform over (x,y) such that $|\Delta(x,y) - n/2| \geq c\sqrt{n}$

Protocol assumptions (eventually, will lead to contradiction):

- Deterministic k-round protocol for $\mathrm{GHD}_{c,n}$
- Each message is $s \ll n$ bits
- Error probability $\leq \varepsilon$, under distribution $\mu_{c,n}$

Round Elimination

Main Construction: Given k-round protocol $\mathcal{P}$ for $\text{GHD}_{c,n}$, construct $(k-1)$-round protocol $\mathcal{Q}$ for $\text{GHD}_{c',n'}$

Round Elimination

Main Construction: Given k-round protocol $\mathcal{P}$ for $\textsc{GHD}_{c,n}$, construct $(k-1)$-round protocol $\mathcal{Q}$ for $\textsc{GHD}_{c',n'}$

First Attempt:

- Fix Alice's first message m in $\mathcal{P}$, suitably

Round Elimination

Main Construction: Given k-round protocol $\mathcal{P}$ for $\text{GHD}_{c,n}$, construct $(k-1)$-round protocol $\mathcal{Q}$ for $\text{GHD}_{c',n'}$

First Attempt:

- Fix Alice's first message m in $\mathcal{P}$, suitably
- Protocol $\mathcal{Q}_1$:
  - Input: $x', y' \in \{0,1\}^A$ where $A \subseteq [n]$, $|A| = n'$
  - Extend $x' \to x$ s.t. Alice sends m on input x
  - Extend $y' \to y$ uniformly at random
  - Output $\mathcal{P}(x,y)$; Note: first message unnecessary

Round Elimination

Main Construction: Given k-round protocol $\mathcal{P}$ for $\text{GHD}_{c,n}$, construct $(k-1)$-round protocol $\mathcal{Q}$ for $\text{GHD}_{c',n'}$

First Attempt:

- Fix Alice's first message m in $\mathcal{P}$, suitably
- Protocol $\mathcal{Q}_1$:
  - Input: $x', y' \in \{0,1\}^A$ where $A \subseteq [n],\ |A| = n'$
  - Extend $x' \to x$ s.t. Alice sends m on input x
  - Extend $y' \to y$ uniformly at random
  - Output $\mathcal{P}(x,y)$; Note: first message unnecessary
- Errors: $\mathcal{Q}_1$ correct, unless
  - BAD_1: $\text{GHD}_{c',n'}(x',y') \neq \text{GHD}_{c,n}(x,y)$.
  - BAD_2: $\text{GHD}_{c,n}(x,y) \neq \mathcal{P}(x,y)$.

Round Elimination

Main Construction: Given k-round protocol $\mathcal{P}$ for $\text{GHD}_{c,n}$, construct $(k-1)$-round protocol $\mathcal{Q}$ for $\text{GHD}_{c',n'}$

First Attempt:

- Fix Alice's first message m in $\mathcal{P}$, suitably
- Protocol $\mathcal{Q}_1$:
  - Input: $x', y' \in \{0,1\}^A$ where $A \subseteq [n],\ |A| = n'$
  - Extend $x' \to x$ s.t. Alice sends m on input x (why possible?)
  - Extend $y' \to y$ uniformly at random
  - Output $\mathcal{P}(x,y)$; Note: first message unnecessary
- Errors: $\mathcal{Q}_1$ correct, unless
  - BAD_1: $\text{GHD}_{c',n'}(x',y') \neq \text{GHD}_{c,n}(x,y)$.
  - BAD_2: $\text{GHD}_{c,n}(x,y) \neq \mathcal{P}(x,y)$.

VC-Dimension

Fixing Alice's first message:

- Call x good if $\Pr_y[\mathcal{P}(x, y) \neq \text{GHD}_{c,n}(x, y)] \leq 2\varepsilon$
  Then $\#\{\text{good } x\} \geq 2^{n-1}$ (Markov)
- Let $M = M_{\mathsf{m}} = \{\text{good } x : \text{Alice sends m on input } x\}$.
- Fix m to maximize $|M|$; then $|M| \geq 2^{n-1-s}$.

VC-Dimension

Fixing Alice's first message:

- Call x good if $\Pr_y[\mathcal{P}(x,y) \neq \textsc{GHD}_{c,n}(x,y)] \leq 2\varepsilon$
  Then $\#\{\text{good } x\} \geq 2^{n-1}$ (Markov)
- Let $M = M_{\mathsf{m}} = \{\text{good } x : \text{Alice sends } \mathsf{m} \text{ on input } x\}$.
- Fix m to maximize $|M|$; then $|M| \geq 2^{n-1-s}$.

$$\begin{array}{ccccc} & & A & & \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{array}$$

Shattering:

- Say $S \subseteq \{0,1\}^n$ shatters $A \subseteq [n]$ if $\#\{x|_A : x \in S\} = 2^{|A|}$
- $\mathrm{VCD}(S)$:= size of largest A shattered by S

VC-Dimension

$$\begin{array}{ccccc} & \text{A} & & \text{A} & \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{array}$$

Fixing Alice's first message:

- Call x good if $\Pr_y[\mathcal{P}(x,y) \neq \text{GHD}_{c,n}(x,y)] \leq 2\varepsilon$

  Then $\#\{\text{good } x\} \geq 2^{n-1}$ (Markov)

- Let $M = M_{\mathsf{m}} = \{\text{good } x :$ Alice sends m on input $x\}$.
- Fix m to maximize $|M|$; then $|M| \geq 2^{n-1-s}$.

Shattering:

- Say $S \subseteq \{0,1\}^n$ shatters $A \subseteq [n]$ if $\#\{x|_A : x \in S\} = 2^{|A|}$
- $\mathrm{VCD}(S)$:= size of largest A shattered by S

Sauer's Lemma: If $\mathrm{VCD}(S) < \alpha n$ then $|S| < 2^{nH(\alpha)}$.

VC-Dimension

$$\begin{array}{ccccccc} & & & A & & & \\ 1 & 0 & 1 & 1 & 0 & \leftarrow \\ 0 & 0 & 1 & 1 & 1 & \\ 1 & 1 & 1 & 1 & 1 & \leftarrow \\ 1 & 1 & 1 & 0 & 1 & \leftarrow \\ 1 & 0 & 1 & 0 & 0 & \leftarrow \\ 0 & 1 & 0 & 1 & 0 & \end{array}$$

Fixing Alice's first message:

- Call x good if $\Pr_y[\mathcal{P}(x,y) \neq \text{GHD}_{c,n}(x,y)] \leq 2\varepsilon$

  Then $\#\{\text{good } x\} \geq 2^{n-1}$ (Markov)
- Let $M = M_{\mathsf{m}} = \{\text{good } x : \text{Alice sends } \mathsf{m} \text{ on input } x\}$.
- Fix m to maximize $|M|$; then $|M| \geq 2^{n-1-s}$.

Shattering:

- Say $S \subseteq \{0,1\}^n$ shatters $A \subseteq [n]$ if $\#\{x|_A : x \in S\} = 2^{|A|}$
- $\text{VCD}(S)$:= size of largest A shattered by S

Sauer's Lemma: If $\text{VCD}(S) < \alpha n$ then $|S| < 2^{nH(\alpha)}$.

Corollary: $\text{VCD}(M) \geq n' := n/3$ (Because $s \ll n$)

VC-Dimension

A

$$\begin{array}{ccccc} 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{array}$$

Fixing Alice's first message:

- Call x good if $\Pr_y[\mathcal{P}(x,y) \neq \text{GHD}_{c,n}(x,y)] \leq 2\varepsilon$

  Then $\#\{\text{good } x\} \geq 2^{n-1}$ (Markov)
- Let $M = M_{\mathsf{m}} = \{\text{good } x : \text{Alice sends } \mathsf{m} \text{ on input } x\}$.
- Fix m to maximize $|M|$; then $|M| \geq 2^{n-1-s}$.

Shattering:

- Say $S \subseteq \{0,1\}^n$ shatters $A \subseteq [n]$ if $\#\{x|_A : x \in S\} = 2^{|A|}$
- $\text{VCD}(S)$:= size of largest A shattered by S

Sauer's Lemma: If $\text{VCD}(S) < \alpha n$ then $|S| < 2^{nH(\alpha)}$.

Corollary: $\text{VCD}(M) \geq n' := n/3$ (Because $s \ll n$)

Extend $x' \to x$: pick $x \in M$ such that $x' = x|_A$

The First Bad Event

Recall BAD_1: $\textsc{ghd}_{c',n'}(x', y') \neq \textsc{ghd}_{c,n}(x, y)$.

Notation: $x = x' \circ \bar{x}$, $y = y' \circ \bar{y}$, $n = n' + \bar{n}$.

The First Bad Event

Recall BAD_1: $\textsc{GHD}_{c',n'}(x', y') \neq \textsc{GHD}_{c,n}(x, y)$.

Notation: $x = x' \circ \bar{x}$, $y = y' \circ \bar{y}$, $n = n' + \bar{n}$.

Definition: $\bar{x}, \bar{y}$ nearly orthogonal if $|\Delta(\bar{x}, \bar{y}) - \bar{n}/2| < 2\sqrt{\bar{n}}$.

The First Bad Event

Recall BAD_1: $\text{GHD}_{c',n'}(x', y') \neq \text{GHD}_{c,n}(x, y)$.

Notation: $x = x' \circ \bar{x}$, $y = y' \circ \bar{y}$, $n = n' + \bar{n}$.

Definition: $\bar{x}, \bar{y}$ nearly orthogonal if $|\Delta(\bar{x}, \bar{y}) - \bar{n}/2| < 2\sqrt{\bar{n}}$.

Lemma: $\Pr_{\bar{y}}[\bar{x}, \bar{y} \text{ nearly orthogonal}] > 7/8$. (Binom distrib tail)

The First Bad Event

Recall BAD_1: $\textsc{ghd}_{c',n'}(x',y') \neq \textsc{ghd}_{c,n}(x,y)$.

Notation: $x = x' \circ \bar{x}$, $y = y' \circ \bar{y}$, $n = n' + \bar{n}$.

Definition: $\bar{x}, \bar{y}$ nearly orthogonal if $|\Delta(\bar{x},\bar{y}) - \bar{n}/2| < 2\sqrt{\bar{n}}$.

Lemma: $\Pr_{\bar{y}}[\bar{x}, \bar{y}$ nearly orthogonal$] > 7/8$. (Binom distrib tail)

Lemma: If $\bar{x}, \bar{y}$ nearly orthogonal and $c' \geq 2c$, then

- $\textsc{ghd}_{c',n'}(x',y') = 1 \implies \textsc{ghd}_{c,n}(x,y) = 1$
- $\textsc{ghd}_{c',n'}(x',y') = 0 \implies \textsc{ghd}_{c,n}(x,y) = 0$

The First Bad Event

Recall BAD_1: $\text{GHD}_{c',n'}(x',y') \neq \text{GHD}_{c,n}(x,y)$.

Notation: $x = x' \circ \bar{x}$, $y = y' \circ \bar{y}$, $n = n' + \bar{n}$.

Definition: $\bar{x}, \bar{y}$ nearly orthogonal if $|\Delta(\bar{x},\bar{y}) - \bar{n}/2| < 2\sqrt{\bar{n}}$.

Lemma: $\Pr_{\bar{y}}[\bar{x},\bar{y}$ nearly orthogonal$] > 7/8$. (Binom distrib tail)

Lemma: If $\bar{x}, \bar{y}$ nearly orthogonal and $c' \geq 2c$, then

- $\text{GHD}_{c',n'}(x',y') = 1 \implies \text{GHD}_{c,n}(x,y) = 1$
- $\text{GHD}_{c',n'}(x',y') = 0 \implies \text{GHD}_{c,n}(x,y) = 0$

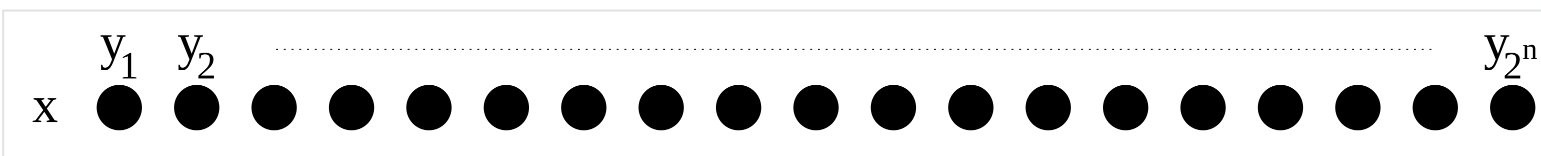
Corollary: $\Pr[BAD_1] < 1/8$.

The Second Bad Event

Recall BAD_2: $\textsc{ghd}_{c,n}(x, y) \neq \mathcal{P}(x, y)$.

Bounding $\Pr[BAD_2]$ is subtle:

- x is good, so $\Pr[\mathcal{P}\text{ errs} \mid x] \leq 2\varepsilon$
  - But this requires $(x, y) \sim \mu_{c,n}$
- Random extension $(x', y') \to (x, y)$ is not $\sim \mu_{c,n}$.

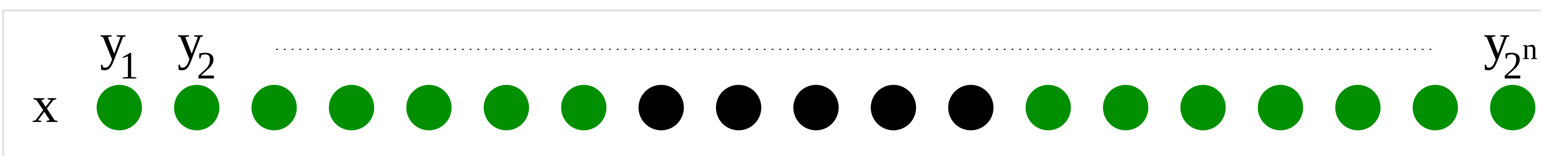


The Second Bad Event

Recall BAD_2: $\textsc{ghd}_{c,n}(x, y) \neq \mathcal{P}(x, y)$.

Bounding $\Pr[BAD_2]$ is subtle:

- x is good, so $\Pr[\mathcal{P} \text{ errs} \mid x] \leq 2\varepsilon$
  - But this requires $(x, y) \sim \mu_{c,n}$
- Random extension $(x', y') \to (x, y)$ is not $\sim \mu_{c,n}$.

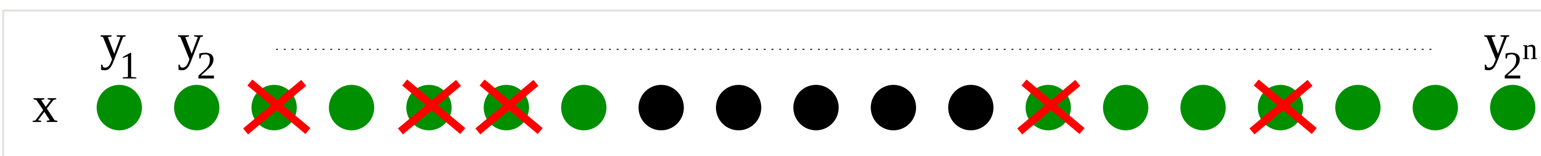


The Second Bad Event

Recall BAD_2: $\textsc{GHD}_{c,n}(x, y) \neq \mathcal{P}(x, y)$.

Bounding $\Pr[BAD_2]$ is subtle:

- x is good, so $\Pr[\mathcal{P}$ errs $\mid x] \leq 2\varepsilon$
  - But this requires $(x, y) \sim \mu_{c,n}$
- Random extension $(x', y') \to (x, y)$ is not $\sim \mu_{c,n}$.

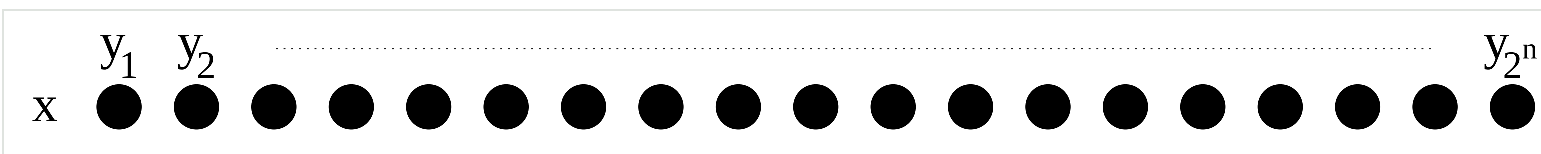


The Second Bad Event

Recall BAD_2: $\text{GHD}_{c,n}(x, y) \neq \mathcal{P}(x, y)$.

Bounding $\Pr[BAD_2]$ is subtle:

- x is good, so $\Pr[\mathcal{P} \text{ errs} \mid x] \leq 2\varepsilon$
  - But this requires $(x, y) \sim \mu_{c,n}$
- Random extension $(x', y') \to (x, y)$ is not $\sim \mu_{c,n}$.
- Actual distrib (fixed x, random y):
  - $(x, y) \sim (\mu_{c',n'} \mid x) \otimes \mathsf{Unif}_{\bar{n}}$
  - y uniform over a subset of $\{0, 1\}^n$, just like in $\mu_{c,n}$

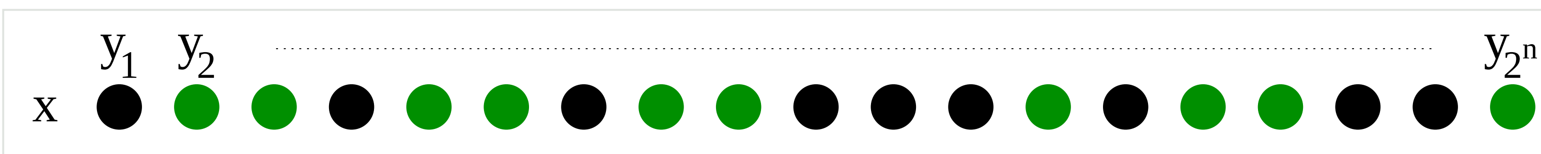


The Second Bad Event

Recall BAD_2: $\textsc{GHD}_{c,n}(x, y) \neq \mathcal{P}(x, y)$.

Bounding $\Pr[BAD_2]$ is subtle:

- x is good, so $\Pr[\mathcal{P} \text{ errs} \mid x] \leq 2\varepsilon$
  - But this requires $(x, y) \sim \mu_{c,n}$
- Random extension $(x', y') \to (x, y)$ is not $\sim \mu_{c,n}$.
- Actual distrib (fixed x, random y):
  - $(x, y) \sim (\mu_{c',n'} \mid x) \otimes \mathsf{Unif}_{\bar{n}}$
  - y uniform over a subset of $\{0, 1\}^n$, just like in $\mu_{c,n}$

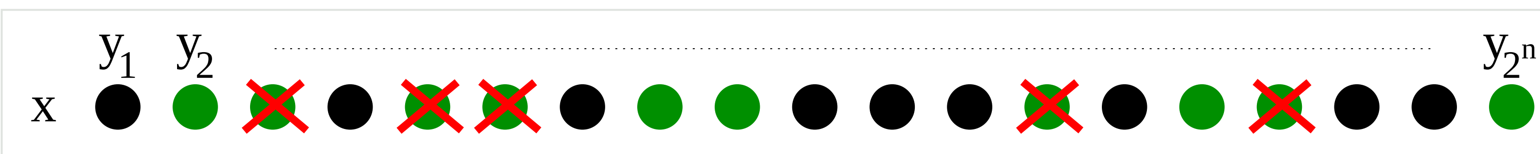


The Second Bad Event

Recall BAD_2: $\textsc{GHD}_{c,n}(x,y) \neq \mathcal{P}(x,y)$.

Bounding $\Pr[BAD_2]$ is subtle:

- x is good, so $\Pr[\mathcal{P}$ errs $\mid x] \leq 2\varepsilon$
  - But this requires $(x,y) \sim \mu_{c,n}$
- Random extension $(x',y') \to (x,y)$ is not $\sim \mu_{c,n}$.
- Actual distrib (fixed x, random y):
  - $(x,y) \sim (\mu_{c',n'} \mid x) \otimes \mathsf{Unif}_{\bar{n}}$
  - y uniform over a subset of $\{0,1\}^n$, just like in $\mu_{c,n}$

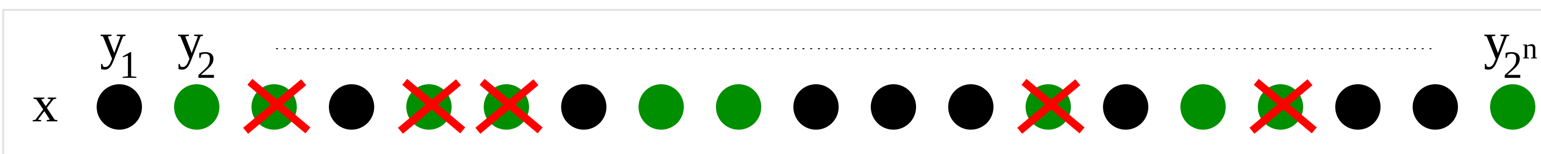


The Second Bad Event

Recall BAD_2: $\textsc{GHD}_{c,n}(x, y) \neq \mathcal{P}(x, y)$.

Bounding $\Pr[BAD_2]$ is subtle:

- x is good, so $\Pr[\mathcal{P}$ errs $| \; x] \leq 2\varepsilon$
  - But this requires $(x, y) \sim \mu_{c,n}$
- Random extension $(x', y') \to (x, y)$ is not $\sim \mu_{c,n}$.
- Actual distrib (fixed x, random y):
  - $(x, y) \sim (\mu_{c',n'} \mid x) \otimes \mathsf{Unif}_{\bar{n}}$
  - y uniform over a subset of $\{0, 1\}^n$, just like in $\mu_{c,n}$



Lemma: $\Pr[BAD_2] = O(\varepsilon)$.

Round Elimination, First Attempt (Recap)

Putting it together:

- $\mathcal{P}$ is k-round ε-error protocol for $\text{GHD}_{c,n}$
- $\mathcal{Q}_1$ is $(k-1)$-round ε'-error protocol for $\text{GHD}_{c',n'}$ with
  - $c' = 2c,\ n' = n/3$
  - $\varepsilon' = 1/8 + O(\varepsilon)$

Round Elimination, First Attempt (Recap)

Putting it together:

- $\mathcal{P}$ is k-round ε-error protocol for $\textsc{GHD}_{c,n}$
- $\mathcal{Q}_1$ is $(k-1)$-round ε'-error protocol for $\textsc{GHD}_{c',n'}$ with
  - $c' = 2c,\ n' = n/3$
  - $\varepsilon' \leq 1/8 + 16\varepsilon$ $\longleftarrow$ Can't repeat this argument!

Round Elimination, Second Attempt

Putting it together:

- $\mathcal{P}$ is k-round ε-error protocol for $\mathrm{GHD}_{c,n}$
- $\mathcal{Q}_1$ is $(k-1)$-round ε'-error protocol for $\mathrm{GHD}_{c',n'}$ with
  - $c' = 2c,\ n' = n/3$
  - $\varepsilon' \le 1/8 + 16\varepsilon$ $\longleftarrow$ Can't repeat this argument!

Second attempt: protocol $\mathcal{Q}$:

- Repeat $\mathcal{Q}_1$ $2^{O(k)}$ times in parallel, take majority
- Blows up communication by $2^{O(k)}$
- Error analysis even more subtle: not just a Chernoff bound

Round Elimination, Second Attempt

Putting it together:

- $\mathcal{P}$ is k-round ε-error protocol for $\text{GHD}_{c,n}$
- $\mathcal{Q}_1$ is $(k-1)$-round ε'-error protocol for $\text{GHD}_{c',n'}$ with
  - $c' = 2c,\ n' = n/3$
  - $\varepsilon' \leq 1/8 + 16\varepsilon$ $\longleftarrow$ Can't repeat this argument!

Second attempt: protocol $\mathcal{Q}$:

- Repeat $\mathcal{Q}_1$ $2^{O(k)}$ times in parallel, take majority
- Blows up communication by $2^{O(k)}$
- Error analysis even more subtle: not just a Chernoff bound

Lemma: $\Pr[Q \text{ errs}] = O(\varepsilon)$.

Eventual Round Elimination Lemma

Lemma: If there is a k-round, ε-error protocol for $\text{GHD}_{c,n}$ in which each player sends $s \ll n$ bits, then there is a $(k-1)$-round, $O(\varepsilon)$-error protocol for $\text{GHD}_{2c,n/3}$ in which each player sends $2^{O(k)}s$ bits.

Recall Base Case Lemma: There is no zero-round protocol with error $< 1/2$.

Eventual Round Elimination Lemma

Lemma: If there is a k-round, ε-error protocol for $\text{GHD}_{c,n}$ in which each player sends $s \ll n$ bits, then there is a $(k-1)$-round, $O(\varepsilon)$-error protocol for $\text{GHD}_{2c,n/3}$ in which each player sends $2^{O(k)}s$ bits.

Recall Base Case Lemma: There is no zero-round protocol with error $< 1/2$.

Consequence: Main Theorem

Theorem: There is no $o(n)$-bit, $\frac{1}{3}$-error, $O(1)$-round randomized protocol for $\text{GHD}_{c,n}$. In other words, $\mathrm{R}^{O(1)}(\text{GHD}) = \Omega(n)$.

Eventual Round Elimination Lemma

Lemma: If there is a k-round, ε-error protocol for $\mathrm{GHD}_{c,n}$ in which each player sends $s \ll n$ bits, then there is a $(k-1)$-round, $O(\varepsilon)$-error protocol for $\mathrm{GHD}_{2c,n/3}$ in which each player sends $2^{O(k)}s$ bits.

Recall Base Case Lemma: There is no zero-round protocol with error $< 1/2$.

Consequence: Main Theorem

Theorem: There is no $o(n)$-bit, $\frac{1}{3}$-error, $O(1)$-round randomized protocol for $\mathrm{GHD}_{c,n}$. In other words, $\mathrm{R}^{O(1)}(\mathrm{GHD}) = \Omega(n)$.

More Specific: $\mathrm{R}^{k}(\mathrm{GHD}) = n/2^{O(k^2)}$.

Why Did This Take So Long?

Multi-pass lower bounds for Distinct Elements and F_k has been an important open question since at least 2003. Why did it remain open for so long?

Why Did This Take So Long?

Multi-pass lower bounds for Distinct Elements and F_k has been an important open question since at least 2003. Why did it remain open for so long?

Underlying communication problem thorny!

Why Did This Take So Long?

Multi-pass lower bounds for Distinct Elements and F_k has been an important open question since at least 2003. Why did it remain open for so long?

Underlying communication problem thorny! Resists the "usual" attacks:

- Rectangle-based methods (discrepancy/corruption)
- Approximate polynomial degree
- Pattern matrix, Factorization norms [Sherstov'08], [Linial-Shraibman'07]
- Information complexity [C.-Shi-Wirth-Yao'01], [BarYossef-J.-K.-S.'02]

Why Did This Take So Long?

Multi-pass lower bounds for Distinct Elements and F_k has been an important open question since at least 2003. Why did it remain open for so long?

Underlying communication problem thorny! Resists the "usual" attacks:

- Rectangle-based methods (discrepancy/corruption)

  Matrix has large near-monochromatic rectangles

- Approximate polynomial degree

- Pattern matrix, Factorization norms [Sherstov'08], [Linial-Shraibman'07]

- Information complexity [C.-Shi-Wirth-Yao'01], [BarYossef-J.-K.-S.'02]

Why Did This Take So Long?

Multi-pass lower bounds for Distinct Elements and F_k has been an important open question since at least 2003. Why did it remain open for so long?

Underlying communication problem thorny! Resists the "usual" attacks:

- Rectangle-based methods (discrepancy/corruption)

  Matrix has large near-monochromatic rectangles

- Approximate polynomial degree

  Underlying predicate has approx degree $\widetilde{O}(\sqrt{n})$

- Pattern matrix, Factorization norms [Sherstov'08], [Linial-Shraibman'07]

- Information complexity [C.-Shi-Wirth-Yao'01], [BarYossef-J.-K.-S.'02]

Why Did This Take So Long?

Multi-pass lower bounds for Distinct Elements and F_k has been an important open question since at least 2003. Why did it remain open for so long?

Underlying communication problem thorny! Resists the "usual" attacks:

- Rectangle-based methods (discrepancy/corruption)

  Matrix has large near-monochromatic rectangles

- Approximate polynomial degree

  Underlying predicate has approx degree $\widetilde{O}(\sqrt{n})$

- Pattern matrix, Factorization norms [Sherstov'08], [Linial-Shraibman'07]

  Quantum communication upper bound $O(\sqrt{n}\log n)$

- Information complexity [C.-Shi-Wirth-Yao'01], [BarYossef-J.-K.-S.'02]

Why Did This Take So Long?

Multi-pass lower bounds for Distinct Elements and F_k has been an important open question since at least 2003. Why did it remain open for so long?

Underlying communication problem thorny! Resists the "usual" attacks:

- Rectangle-based methods (discrepancy/corruption)

  Matrix has large near-monochromatic rectangles

- Approximate polynomial degree

  Underlying predicate has approx degree $\widetilde{O}(\sqrt{n})$

- Pattern matrix, Factorization norms [Sherstov'08], [Linial-Shraibman'07]

  Quantum communication upper bound $O(\sqrt{n}\log n)$

- Information complexity [C.-Shi-Wirth-Yao'01], [BarYossef-J.-K.-S.'02]

  Hmm! Can't see a concrete obstacle

Why Did This Take So Long?

Multi-pass lower bounds for Distinct Elements and F_k has been an important open question since at least 2003. Why did it remain open for so long?

Underlying communication problem thorny! Resists the "usual" attacks:

- Rectangle-based methods (discrepancy/corruption)

  Matrix has large near-monochromatic rectangles

- Approximate polynomial degree

  Underlying predicate has approx degree $\widetilde{O}(\sqrt{n})$

- Pattern matrix, Factorization norms [Sherstov'08], [Linial-Shraibman'07]

  Quantum communication upper bound $O(\sqrt{n}\log n)$

- Information complexity [C.-Shi-Wirth-Yao'01], [BarYossef-J.-K.-S.'02]

  Hmm! Can't see a concrete obstacle

  We're biased (Amit helped invent it, so it's his pet technique)

Open Problems

1. The key problem here: Settle $\mathrm{R}(\textsc{ghd})$.
2. More generally: Understand communication complexity of "gap problems" better.
3. This should help with other streaming problems, e.g., longest increasing subsequence.

Questions? Comments? Post-Doc/Job offers?

Contact jbrody@cs.dartmouth.edu