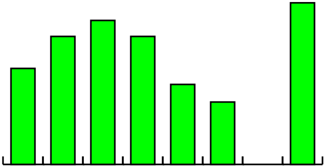
Communication Complexity of Learning Discrete Distributions

Krzysztof Onak
IBM T.J. Watson Research Center

Joint work with **Ilias Diakonikolas**, **Elena Grigorescu**, and **Abhiram Natarajan**.

Distribution Learning and Testing

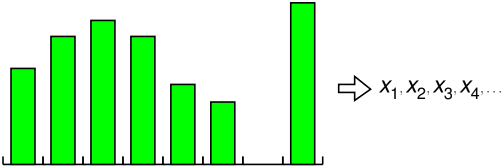
Input: Stream of independent samples
from an unknown distribution $\mathcal{D}$



$\Rightarrow x_1, x_2, x_3, x_4, \ldots$

Distribution Learning and Testing

Input: Stream of independent samples
from an unknown distribution $\mathcal{D}$



Goal:

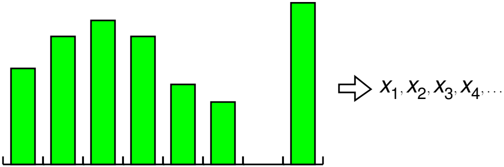
Learn the distribution

or test a property

or estimate a parameter

Distribution Learning and Testing

Input: Stream of independent samples from an unknown distribution $\mathcal{D}$



Goal:

Learn the distribution

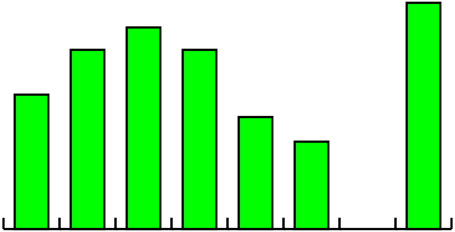
or test a property

or estimate a parameter

- Small total variation distance error acceptable
- Traditional focus: sample complexity

Learning Discrete Distributions

$\mathcal{D}$ = probability distribution on $\{1, \ldots, n\}$
Input: Independent samples from $\mathcal{D}$



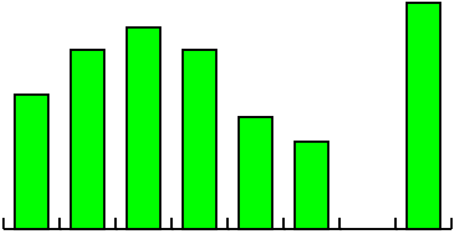
$\Rightarrow x_1, x_2, x_3, x_4, \ldots$

Goal:

Output a distribution $\mathcal{D}'$ such that $\|\mathcal{D} - \mathcal{D}'\|_1 < \epsilon$

Learning Discrete Distributions

$\mathcal{D}$ = probability distribution on $\{1, \ldots, n\}$
Input: Independent samples from $\mathcal{D}$



$\Rightarrow x_1, x_2, x_3, x_4, \ldots$

Goal:

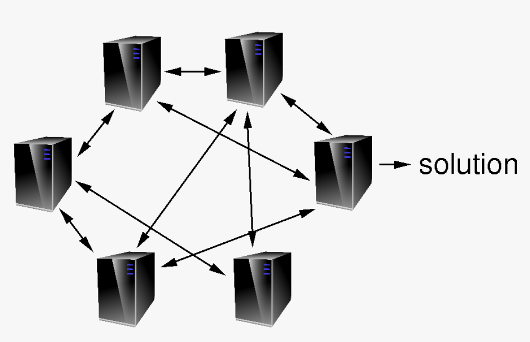
Output a distribution $\mathcal{D}'$ such that $\|\mathcal{D} - \mathcal{D}'\|_1 < \epsilon$

Sample complexity: $\Theta(n/\epsilon^2)$

Communication Complexity

Distributed data: samples held by different players

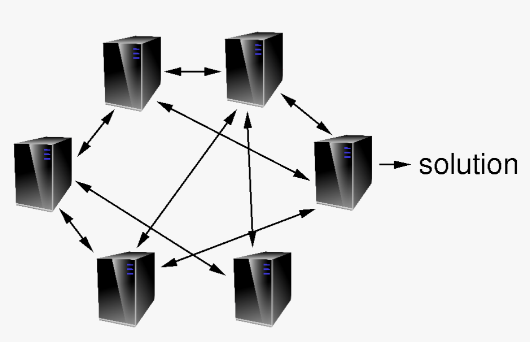
Example: Samples in different data centers



Communication Complexity

Distributed data: samples held by different players

Example: Samples in different data centers



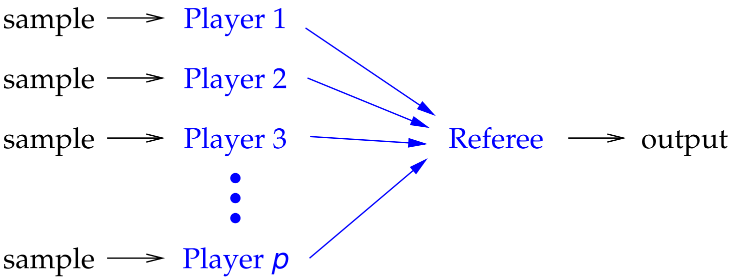
How much do players have to communicate to solve the problem?

Is sublinear communication possible?

"Survey" Complexity

This talk will focus on the simplest setting:

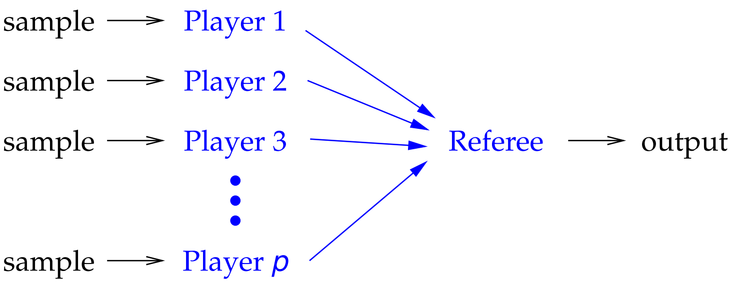
- Each player has one sample
  and sends a single message to a referee
- The referee outputs solution



"Survey" Complexity

This talk will focus on the simplest setting:

- Each player has one sample
  and sends a single message to a referee
- The referee outputs solution



- Each sample is $\Theta(\log n)$ bits
- Can average communication be made $o(\log n)$?

Related Work

A lot of recent interest in communication-efficient learning:

DAW12, ZDW13, ZX15, GMN14, KVW14, LBKW14, SSZ14, DJWZ14, LSLT15, BGMNW15

- Both upper and lower bounds.
- Usually more continuous problems.
- Sample problem: estimating the mean of a Gaussian distribution.

Related Work

A lot of recent interest in communication-efficient learning:

DAW12, ZDW13, ZX15, GMN14, KVW14, LBKW14, SSZ14, DJWZ14, LSLT15, BGMNW15

- Both upper and lower bounds.
- Usually more continuous problems.
- Sample problem: estimating the mean of a Gaussian distribution.

See Mark Braverman's talk tomorrow

Outline

1. $O(n/\epsilon^2)$ Sample Complexity Review
2. Communication Complexity Lower Bound
3. Quick Distribution Testing Example

Outline

1. $O(n/\epsilon^2)$ Sample Complexity Review
2. Communication Complexity Lower Bound
3. Quick Distribution Testing Example

Upper Bound Review

Solution: $\mathcal{D}' =$ empirical distribution of $O(n/\epsilon^2)$ samples

Upper Bound Review

Solution: $\mathcal{D}' =$ empirical distribution of $O(n/\epsilon^2)$ samples

Why this works:

- For every subset of $\{1, \ldots, n\}$ the probabilities under $\mathcal{D}$ and $\mathcal{D}'$ within $\epsilon/2$ with probability $1 - 2^{-2n}$

Upper Bound Review

Solution: $\mathcal{D}' =$ empirical distribution of $O(n/\epsilon^2)$ samples

Why this works:

- For every subset of $\{1, \dots, n\}$ the probabilities under $\mathcal{D}$ and $\mathcal{D}'$ within $\epsilon/2$ with probability $1 - 2^{-2n}$
- Union bound: $\|\mathcal{D} - \mathcal{D}'\|_1 \leq \epsilon$ with probability $1 - o(1)$

Lower Bound Review

Fact: Hoeffding's inequality is optimal

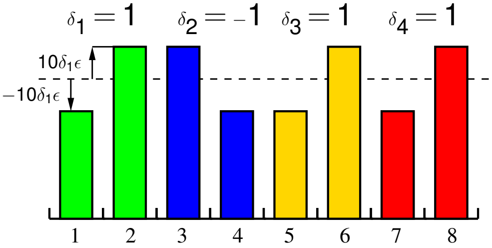
- ϵ-biased coin, determine direction of the bias
- $\Omega(\epsilon^{-2})$ coin tosses needed

Lower Bound Review

Fact: Hoeffding's inequality is optimal

- ϵ-biased coin, determine direction of the bias
- $\Omega(\epsilon^{-2})$ coin tosses needed

Construction:

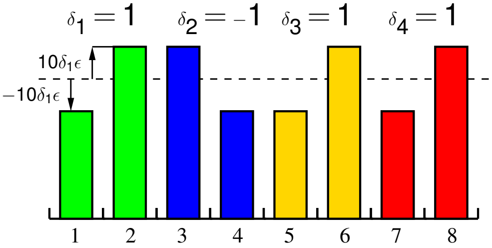


Lower Bound Review

Fact: Hoeffding's inequality is optimal

- ϵ-biased coin, determine direction of the bias
- $\Omega(\epsilon^{-2})$ coin tosses needed

Construction:



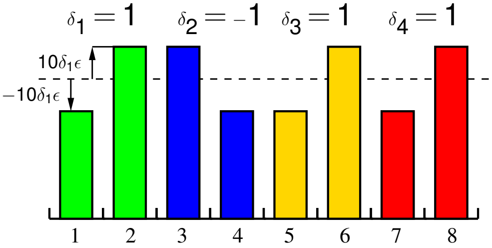
- Each pair randomly biased by 10ϵ

Lower Bound Review

Fact: Hoeffding's inequality is optimal

- ϵ-biased coin, determine direction of the bias
- $\Omega(\epsilon^{-2})$ coin tosses needed

Construction:



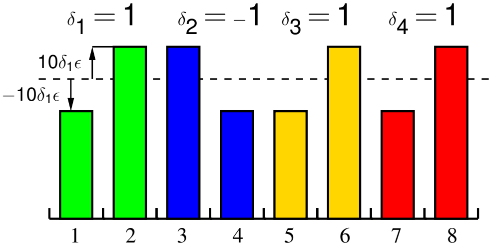
- Each pair randomly biased by 10ϵ
- Need to predict bias of more than $\frac{9}{10}$ pairs
  (via averaging/Markov's bound)

Lower Bound Review

Fact: Hoeffding's inequality is optimal

- ϵ-biased coin, determine direction of the bias
- $\Omega(\epsilon^{-2})$ coin tosses needed

Construction:



- Each pair randomly biased by 10ϵ
- Need to predict bias of more than $\frac{9}{10}$ pairs (via averaging/Markov's bound)
- This requires $\Omega(n/\epsilon^2)$ samples

Outline

1. $O(n/\epsilon^2)$ Sample Complexity Review
2. Communication Complexity Lower Bound
3. Quick Distribution Testing Example

Our Claim

No protocol with $o\left(\frac{n}{\epsilon^2}\log n\right)$
communication on average
that succeeds learning the distribution
with probability $99/100$.

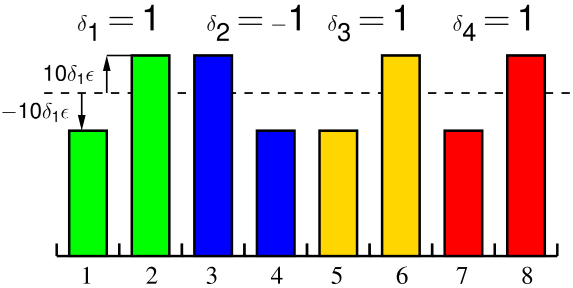
Our Claim

No protocol with $o\left(\frac{n}{\epsilon^2}\log n\right)$
communication on average
that succeeds learning the distribution
with probability $99/100$.

(Can assume at most $O\left(n/\epsilon^2 \log n\right)$ players in the proof)

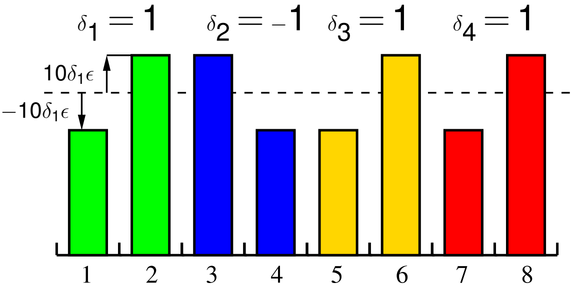
Hard Distribution

Reuse the hard distribution for sampling:



Hard Distribution

Reuse the hard distribution for sampling:



Can assume the protocol is deterministic:

- Slight loss in the probability of success
- Expected communication goes up by constant factor

The Proof Plan

- Assume $o(n\epsilon^{-2}\log n)$ communication protocol

The Proof Plan

- Assume $o(n\epsilon^{-2}\log n)$ communication protocol
- For random i, show that:
  - Messages reveal very little about δ_i
    (even if the referee knows all other δ_i's)
  - The referee can predict δ_i with probability $\frac{1}{2} + o(1)$

The Proof Plan

- Assume $o(n\epsilon^{-2} \log n)$ communication protocol
- For random i, show that:
  - Messages reveal very little about δ_i (even if the referee knows all other δ_i's)
  - The referee can predict δ_i with probability $\frac{1}{2} + o(1)$
- The original protocol correct only on $\frac{1}{2} + o(1)$ fraction of δ_i's most of the time

The Proof Plan

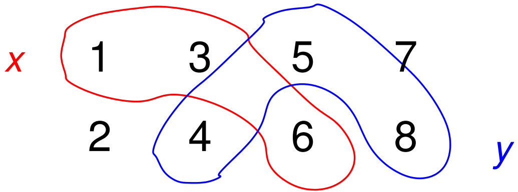
- Assume $o(n\epsilon^{-2} \log n)$ communication protocol
- For random i, show that:
  - Messages reveal very little about δ_i (even if the referee knows all other δ_i's)
  - The referee can predict δ_i with probability $\frac{1}{2} + o(1)$
- The original protocol correct only on $\frac{1}{2} + o(1)$ fraction of δ_i's most of the time

CONTRADICTION!!!

Messages of Single Player

Modify protocol for each pair $2j - 1$ and $2j$:

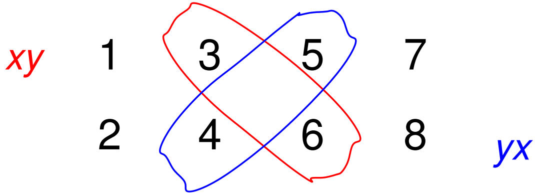
- Before: x sent for $2j - 1$ and y sent for $2j$
- After: send xy for $2j - 1$ and yx for $2j$



Messages of Single Player

Modify protocol for each pair $2j-1$ and $2j$:

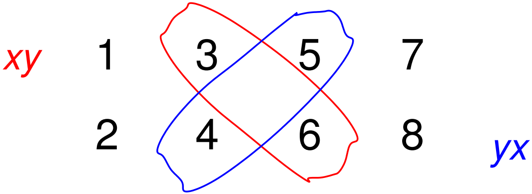
- Before: x sent for $2j-1$ and y sent for $2j$
- After: send xy for $2j-1$ and yx for $2j$



Messages of Single Player

Modify protocol for each pair $2j-1$ and $2j$:

- Before: x sent for $2j-1$ and y sent for $2j$
- After: send xy for $2j-1$ and yx for $2j$



Result:

- Communication complexity only doubles.
- This partitions pairs. Each message reveals bias on a specific subset of pairs.

Messages of Single Player

Three cases for a pair $2i - 1$ and $2i$
and corresponding messages xy and yx:

Messages of Single Player

Three cases for a pair $2i - 1$ and $2i$
and corresponding messages xy and yx:

1. $|xy| > \frac{\log n}{100}$
   - Happens for $o(n/\epsilon^2)$ fraction of players
   - Can assume the message reveals the sample
   - $I(\text{message}; \delta_i) \leq I(\text{sample}; \delta_i) = O(\epsilon^2/n)$

Messages of Single Player

Three cases for a pair $2i - 1$ and $2i$
and corresponding messages xy and yx:

1. $|xy| > \frac{\log n}{100}$
2. $|xy| \leq \frac{\log n}{100}$ & $\leq \sqrt{n}$ pairs with these messages
    - Random i: happens with probability $\frac{n^{0.01} \cdot \sqrt{n}}{n}$
    - Can assume the message reveals the sample
    - $I(\text{message}; \delta_i) \leq I(\text{sample}; \delta_i) = O(\epsilon^2/n)$

Messages of Single Player

Three cases for a pair $2i - 1$ and $2i$
and corresponding messages xy and yx:

1. $|xy| > \frac{\log n}{100}$
2. $|xy| \leq \frac{\log n}{100}$ & $\leq \sqrt{n}$ pairs with these messages
3. $|xy| \leq \frac{\log n}{100}$ & $> \sqrt{n}$ pairs with these messages
    - Can happen always
    - δ_i has little impact on probabilities of xy and yx
    - $I(\text{sample}; \delta_i) = O(\epsilon^2/(n \cdot \#\text{pairs})) = O(\epsilon^2/n^{1.5})$

Total Information about δ_i

$M_j =$ message of the j-th player $\quad M = (M_1, M_2, \ldots, M_p)$

Total Information about δ_i

M_j = message of the j-th player $\quad M = (M_1, M_2, \ldots, M_p)$

For all but $o(1)$ fraction of i's:

$$\sum_j I(\delta_i; M_j) = o\left(\frac{n}{\epsilon^2}\right) \cdot O\left(\frac{\epsilon^2}{n}\right) + O\left(\frac{n^{0.52}}{\epsilon^2}\right) \cdot O\left(\frac{\epsilon^2}{n}\right)$$
$$+ O\left(\frac{n \log n}{\epsilon^2}\right) \cdot O\left(\frac{\epsilon^2}{n^{1.5}}\right) = o(1)$$

Total Information about δ_i

M_j = message of the j-th player $\qquad M = (M_1, M_2, \ldots, M_p)$

For all but $o(1)$ fraction of i's:

$$\sum_j I(\delta_i; M_j) = o\left(\frac{n}{\epsilon^2}\right) \cdot O\left(\frac{\epsilon^2}{n}\right) + O\left(\frac{n^{0.52}}{\epsilon^2}\right) \cdot O\left(\frac{\epsilon^2}{n}\right)$$
$$+ O\left(\frac{n \log n}{\epsilon^2}\right) \cdot O\left(\frac{\epsilon^2}{n^{1.5}}\right) = o(1)$$

Then $I(\delta_i; M) = o(1)$:

- Messages M_j independent once δ_i is fixed
- This implies that $I(\delta_i; M) \le \sum_j I(\delta_i, M_j)$

Total Information about δ_i

M_j = message of the j-th player $\qquad M = (M_1, M_2, \ldots, M_p)$

For all but $o(1)$ fraction of i's:

$$\sum_j I(\delta_i; M_j) = o\left(\frac{n}{\epsilon^2}\right) \cdot O\left(\frac{\epsilon^2}{n}\right) + O\left(\frac{n^{0.52}}{\epsilon^2}\right) \cdot O\left(\frac{\epsilon^2}{n}\right) + O\left(\frac{n \log n}{\epsilon^2}\right) \cdot O\left(\frac{\epsilon^2}{n^{1.5}}\right) = o(1)$$

Then $I(\delta_i; M) = o(1)$:

- Messages M_j independent once δ_i is fixed
- This implies that $I(\delta_i; M) \leq \sum_j I(\delta_i, M_j)$

And $H(\delta_i | M) = H(\delta_i) - I(\delta_i; M) = 1 - o(1)$

Total Information about δ_i

M_j = message of the j-th player $\quad M = (M_1, M_2, \ldots, M_p)$

For all but $o(1)$ fraction of i's:

$$\sum_j I(\delta_i; M_j) = o\left(\frac{n}{\epsilon^2}\right) \cdot O\left(\frac{\epsilon^2}{n}\right) + O\left(\frac{n^{0.52}}{\epsilon^2}\right) \cdot O\left(\frac{\epsilon^2}{n}\right)$$
$$+ O\left(\frac{n \log n}{\epsilon^2}\right) \cdot O\left(\frac{\epsilon^2}{n^{1.5}}\right) = o(1)$$

Then $I(\delta_i; M) = o(1)$:

- Messages M_j independent once δ_i is fixed
- This implies that $I(\delta_i; M) \leq \sum_j I(\delta_i, M_j)$

And $H(\delta_i|M) = H(\delta_i) - I(\delta_i; M) = 1 - o(1)$

Algorithm correct with probability $\frac{1}{2} + o(1)$

Outline

1. $O(n/\epsilon^2)$ Sample Complexity Review

2. Communication Complexity Lower Bound

3. Quick Distribution Testing Example

Uniformity Testing

Problem:

- Distinguish $\mathcal{D} = \mathcal{U}$ vs. $\|\mathcal{D} - \mathcal{U}\|_1 \geq \epsilon$

Uniformity Testing

Problem:

- Distinguish $\mathcal{D} = \mathcal{U}$ vs. $\|\mathcal{D} - \mathcal{U}\|_1 \geq \epsilon$
- Sample complexity: $\Theta\left(\sqrt{n}/\epsilon^2\right)$

Uniformity Testing

Problem:

- Distinguish $\mathcal{D} = \mathcal{U}$ vs. $\|\mathcal{D} - \mathcal{U}\|_1 \geq \epsilon$
- Sample complexity: $\Theta\left(\sqrt{n}/\epsilon^2\right)$

Communication complexity bound:

- Assume lengths of all messages $o(\log n)$
- Methods presented here imply:
  - Referee likely learns $n^{-\Omega(1)}$-fraction of samples
  - Other messages provide little information
  - Not enough to distinguish hard instances

This talk:

- Communication lower bounds
- Players have to essentially transmit their samples

This talk:

- Communication lower bounds
- Players have to essentially transmit their samples

Longer goals

- Reinterpret known distribution testing and learning results in this framework
- Design non-trivial protocols with sublinear amount of communication

This talk:

- Communication lower bounds
- Players have to essentially transmit their samples

Longer goals

- Reinterpret known distribution testing and learning results in this framework
- Design non-trivial protocols with sublinear amount of communication

Questions?