Communication-Efficient Distributed Learning of Discrete Distributions

Abhiram Natarajan (Purdue)

Joint work with Ilias Diakonikolas (USC), Elena Grigorescu (Purdue), Jerry Li (MIT), Krzysztof Onak (IBM), and Ludwig Schmidt (MIT)

Nonparametric Discrete Density Estimation

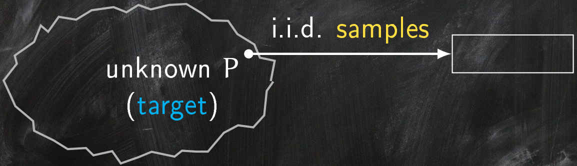
Distribution family $\mathcal{D}$
over domain $\{1, \ldots, d\}$

Nonparametric Discrete Density Estimation

Distribution family $\mathcal{D}$
over domain $\{1, \ldots, d\}$

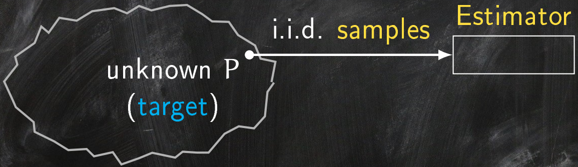
Nonparametric Discrete Density Estimation

Distribution family $\mathcal{D}$
over domain $\{1, \ldots, d\}$



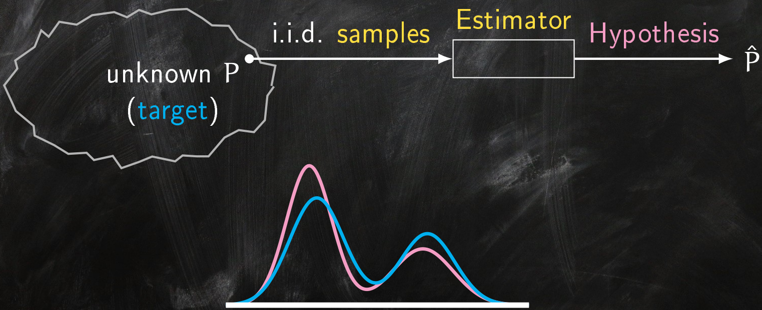
Nonparametric Discrete Density Estimation

Distribution family $\mathcal{D}$
over domain $\{1, \ldots, d\}$



Nonparametric Discrete Density Estimation

Distribution family $\mathcal{D}$
over domain $\{1, \ldots, d\}$



Nonparametric Discrete Density Estimation

- For small error ε, we want

$$\mathbb{E}\left[\mathrm{dist}(P, \hat{P})\right] \leqslant \varepsilon \qquad \text{dist is } \ell_1 \text{ or } \ell_2 \text{ distance}$$

Nonparametric Discrete Density Estimation

- For small error ε, we want

$$\mathbb{E}\left[\mathrm{dist}(P, \hat{P})\right] \leqslant \varepsilon \qquad \text{dist is } \ell_1 \text{ or } \ell_2 \text{ distance}$$

- Fundamental learning problem with many applications

Nonparametric Discrete Density Estimation

- For small error ε, we want

  $\mathbb{E}\left[\mathrm{dist}(\mathrm{P}, \hat{\mathrm{P}})\right] \leqslant \varepsilon$ dist is ℓ_1 or ℓ_2 distance

- Fundamental learning problem with many applications

Sample Size vs Runtime vs Communication

Distributed Density Estimation

- Data is distributed amongst machines

Distributed Density Estimation

- Data is distributed amongst machines
- Need communication-efficient distributed protocols

Distributed Density Estimation

- Data is distributed amongst machines
- Need communication-efficient distributed protocols
- Communication complexity - practical and fundamental

Communication Model

Blackboard

Referee

Samples:
$X_1^{(1)}$
$X_2^{(1)}$
$\vdots$

Machine 1

Samples:
$X_1^{(2)}$
$X_2^{(2)}$
$\vdots$

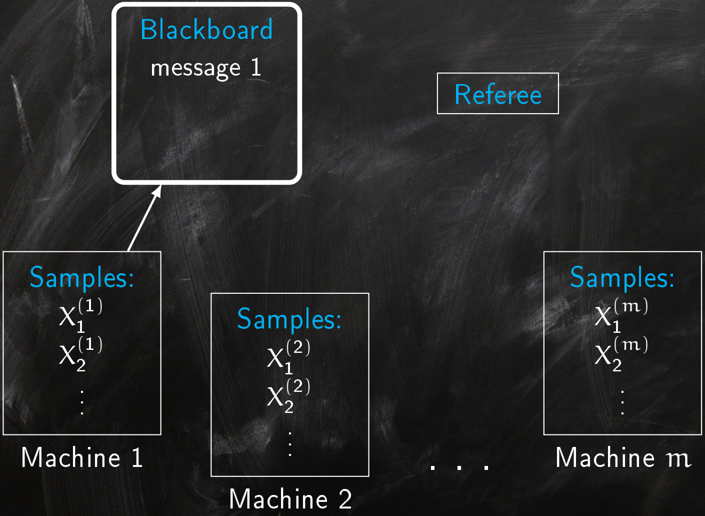
Machine 2

. . .

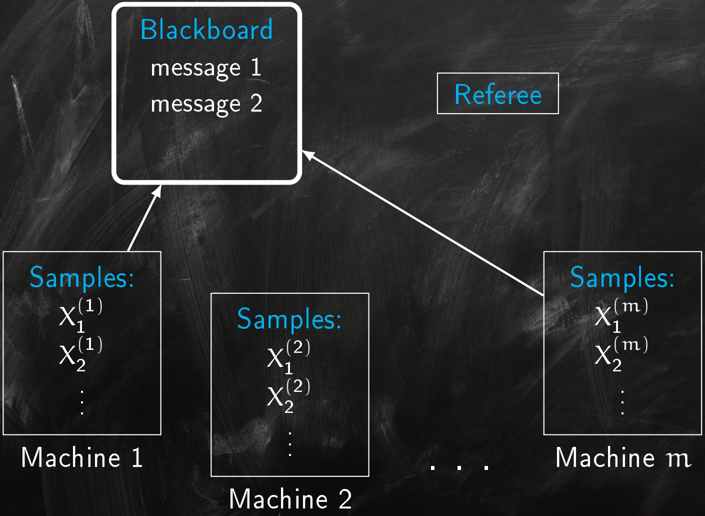
Samples:
$X_1^{(m)}$
$X_2^{(m)}$
$\vdots$

Machine m

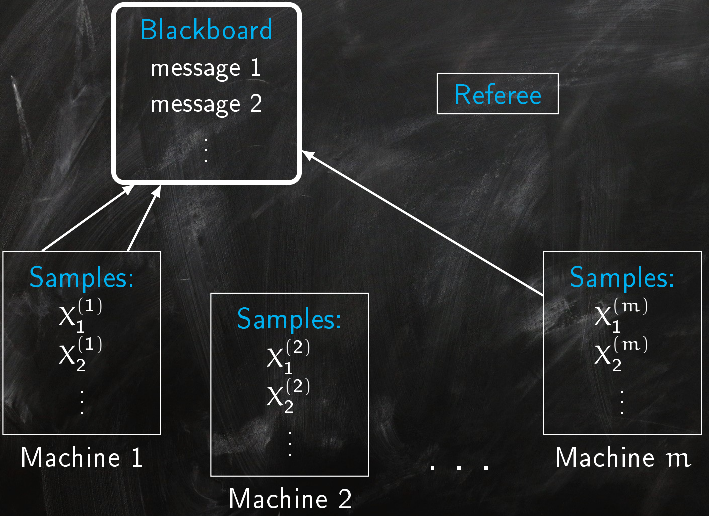
Communication Model



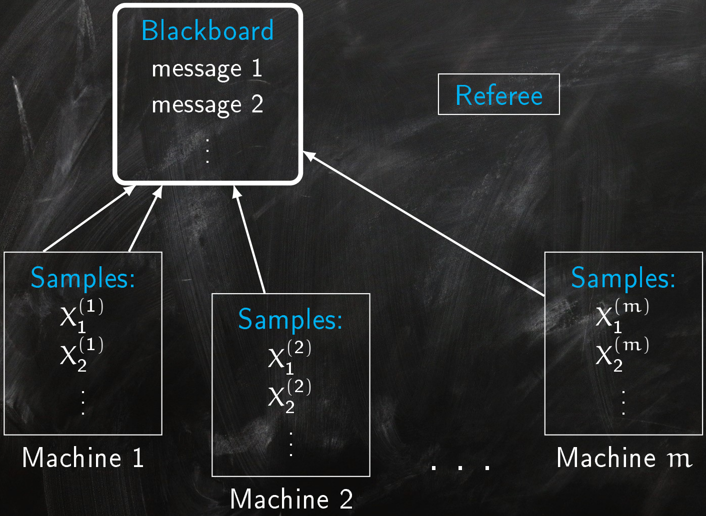
Communication Model



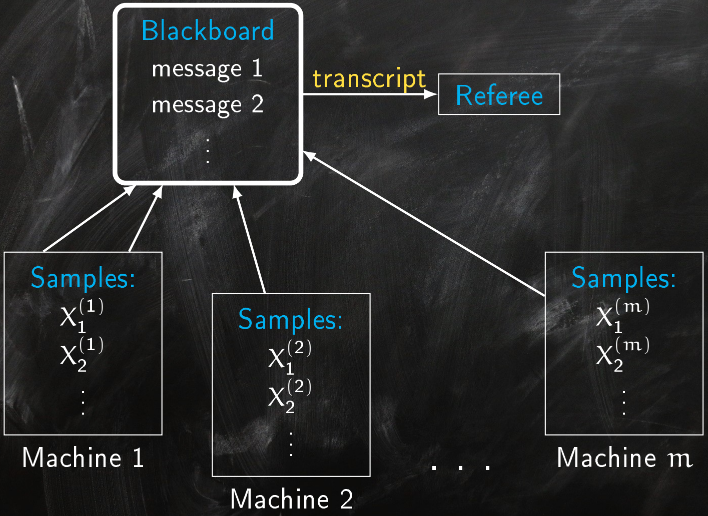
Communication Model



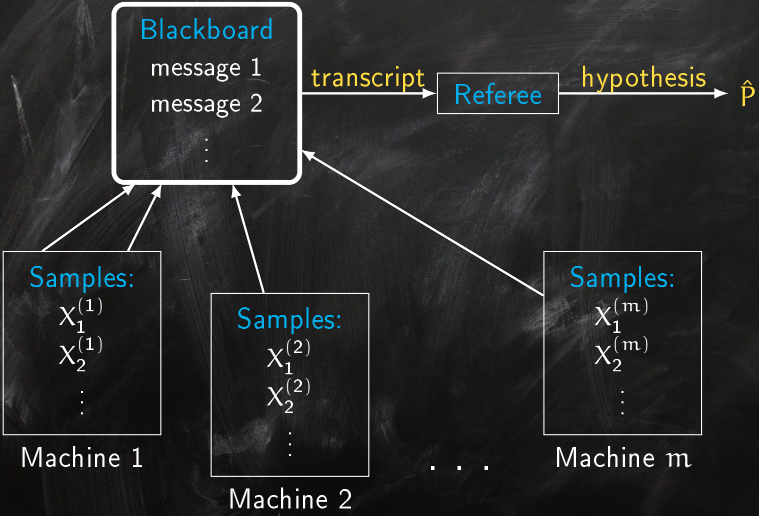
Communication Model



Communication Model

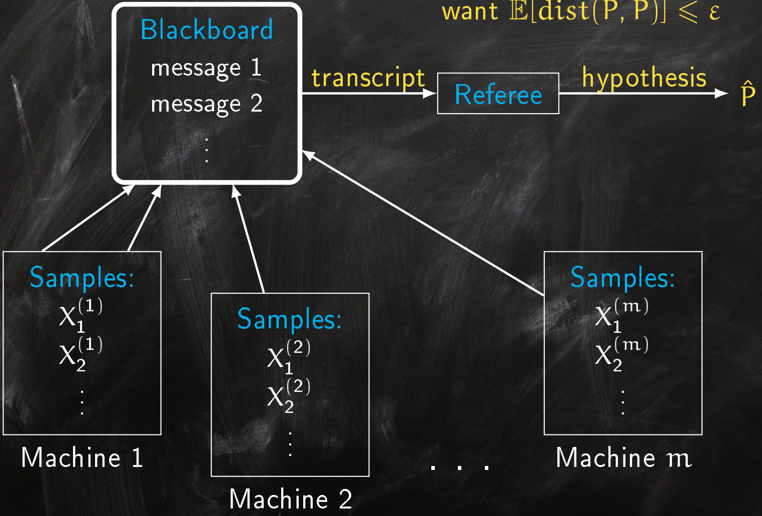


Communication Model



Communication Model

want $\mathbb{E}[\mathrm{dist}(\hat{P}, P)] \leqslant \varepsilon$



Distributed Density Estimation

- n is a sufficient sample size for learning family $\mathcal{D}$

Distributed Density Estimation

- n is a sufficient sample size for learning family $\mathcal{D}$
- There exists non-distributed algorithm to learn $\mathcal{D}$ using n samples

Distributed Density Estimation

- n is a sufficient sample size for learning family $\mathcal{D}$
- There exists non-distributed algorithm to learn $\mathcal{D}$ using n samples
- From unknown $P \in \mathcal{D}$, we have s samples each on $\frac{n}{s}$ machines

Distributed Density Estimation

- n is a sufficient sample size for learning family $\mathcal{D}$
- There exists non-distributed algorithm to learn $\mathcal{D}$ using n samples
- From unknown $P \in \mathcal{D}$, we have s samples each on $\frac{n}{s}$ machines
- How many bits are in transcript of protocol?

Distributed Density Estimation

- n is a sufficient sample size for learning family $\mathcal{D}$
- There exists non-distributed algorithm to learn $\mathcal{D}$ using n samples
- From unknown $P \in \mathcal{D}$, we have s samples each on $\frac{n}{s}$ machines
- How many bits are in transcript of protocol?

Fact (Baseline Protocol)

There exists protocol with $O(n \log d)$ bits of communication.

High-Level Summary of Results

- In the absence of structural assumptions on the distribution, the baseline protocol is optimal

High-Level Summary of Results

- In the absence of structural assumptions on the distribution, the baseline protocol is optimal
- When distribution is structured (k-histograms, monotone, etc.), structure can be exploited for improvement

Unstructured Distributions in ℓ_1

- $\Theta\left(\frac{d}{\varepsilon^2}\right)$ samples necessary and sufficient for learning any distribution over $\{1, \ldots, d\}$ in ℓ_1 distance

Unstructured Distributions in ℓ_1

- $\Theta\left(\frac{d}{\varepsilon^2}\right)$ samples necessary and sufficient for learning any distribution over $\{1, \ldots, d\}$ in ℓ_1 distance
- Baseline protocol uses $O\left(\frac{d}{\varepsilon^2}\log d\right)$ bits of communication

Unstructured Distributions in ℓ_1

- $\Theta\left(\frac{d}{\varepsilon^2}\right)$ samples necessary and sufficient for learning any distribution over $\{1, \ldots, d\}$ in ℓ_1 distance

- Baseline protocol uses $O\left(\frac{d}{\varepsilon^2} \log d\right)$ bits of communication

Theorem (Communication Lower Bound)

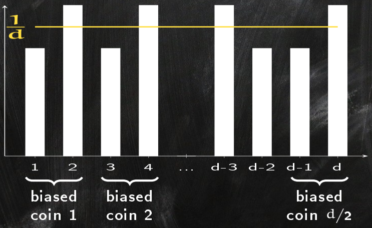
$\Omega\left(\frac{d}{\varepsilon^2} \log d\right)$ *bits is the best possible protocol when there is one sample per machine*

Lower Bound Proof Ideas

- Construct *hard to learn* family of distributions on $\{1, \ldots, d\}$:

$$\mathbb{P}(2i-1) = \frac{1+10\delta_i\varepsilon}{d} \qquad \mathbb{P}(2i) = \frac{1-10\delta_i\varepsilon}{d},$$

δ_i uniform on $\{-1, 1\}$

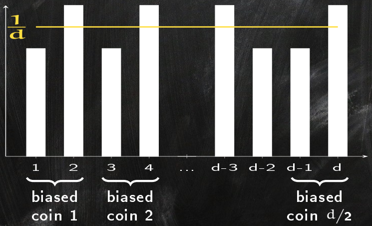


Lower Bound Proof Ideas

- Construct *hard to learn* family of distributions on $\{1, \ldots, d\}$:

$$\mathbb{P}(2i-1) = \frac{1 + 10\delta_i \varepsilon}{d} \qquad \mathbb{P}(2i) = \frac{1 - 10\delta_i \varepsilon}{d},$$

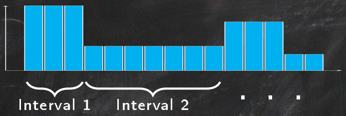
δ_i uniform on $\{-1, 1\}$



- Using *information complexity* machinery, we show that large number of bits is required to get information about all coins

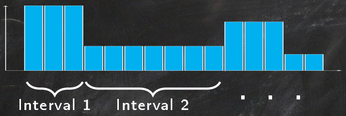
K-Histogram Distributions

- Piecewise-constant over some set of k intervals over $\{1, ..., d\}$

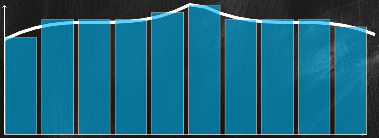


K-Histogram Distributions

- Piecewise-constant over some set of k intervals over $\{1, \ldots, d\}$

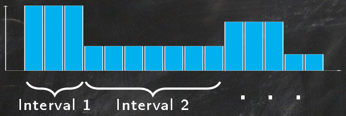


- Motivation - histogram approximations exist for large class of distributions (log-concave, unimodal, etc.)

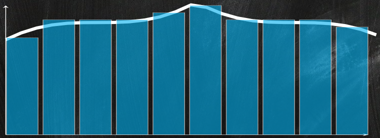


k-Histogram Distributions

- Piecewise-constant over some set of k intervals over $\{1, ..., d\}$



- Motivation - histogram approximations exist for large class of distributions (log-concave, unimodal, etc.)



- Need learning algorithm that is robust to model mis-specification

Learning k-Histograms in ℓ_2

- $\Theta\left(\frac{1}{\varepsilon^2}\right)$ samples necessary and sufficient to learn k-Histograms in ℓ_2

Learning k-Histograms in ℓ_2

- $\Theta\left(\frac{1}{\varepsilon^2}\right)$ samples necessary and sufficient to learn k-Histograms in ℓ_2
- When partition known, reduces to unstructured case

Learning k-Histograms in ℓ_2

- $\Theta\left(\frac{1}{\varepsilon^2}\right)$ samples necessary and sufficient to learn k-Histograms in ℓ_2
- When partition known, reduces to unstructured case
- When partition unknown, baseline protocol $O(\frac{1}{\varepsilon^2}\log d)$ bits

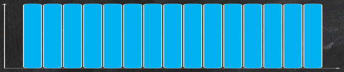
Learning k-Histograms in ℓ_2

- $\Theta\left(\frac{1}{\varepsilon^2}\right)$ samples necessary and sufficient to learn k-Histograms in ℓ_2
- When partition known, reduces to unstructured case
- When partition unknown, baseline protocol $O(\frac{1}{\varepsilon^2}\log d)$ bits

Theorem (Communication Upper bound)
There exists robust protocol with $\tilde{O}\left(mk\log\frac{1}{\varepsilon}\log d\right)$ *bits of communication, where* m *is number of machines*

Upper Bound Proof Ideas

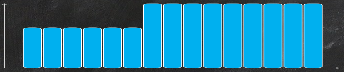
Blackboard



Samples: $\vec{X}_1$	Samples: $\vec{X}_2$	. . .	Samples: $\vec{X}_m$
Machine 1	Machine 2		Machine m

Upper Bound Proof Ideas

Blackboard



Samples: $\vec{X}_1$	Samples: $\vec{X}_2$	. . .	Samples: $\vec{X}_m$
Machine 1	Machine 2		Machine m

Upper Bound Proof Ideas

Blackboard

Samples: $\vec{X}_1$	Samples: $\vec{X}_2$	$\cdots$	Samples: $\vec{X}_m$
Machine 1	Machine 2		Machine m

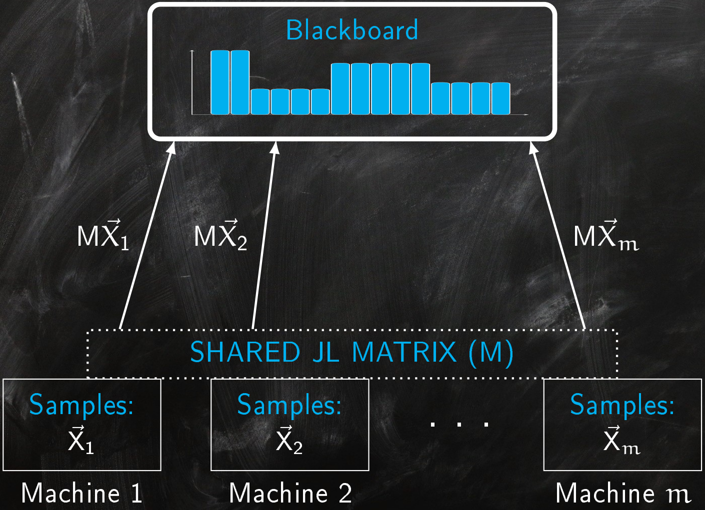
Upper Bound Proof Ideas

Blackboard

SHARED JL MATRIX (M)

Samples: $\vec{X}_1$	Samples: $\vec{X}_2$	$\cdots$	Samples: $\vec{X}_m$
Machine 1	Machine 2		Machine m

Upper Bound Proof Ideas



Other Results

- Additionally, communication bounds in multiple regimes for:
  - Unstructured distributions in ℓ_2
  - k-Histograms in ℓ_1
  - Monotone distributions in ℓ_1

Other Results

- Additionally, communication bounds in multiple regimes for:
  - Unstructured distributions in ℓ_2
  - k-Histograms in ℓ_1
  - Monotone distributions in ℓ_1
- All structured learners are robust to model mis-specification

Conclusion & Open Problems

- We provide first communication bounds for a large class of discrete distributions

Conclusion & Open Problems

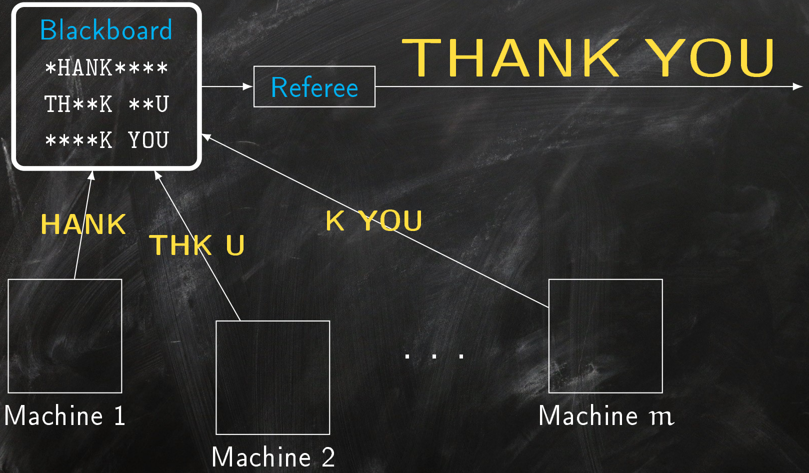
- We provide first communication bounds for a large class of discrete distributions
- Some open problems:
  - Tighten upper and lower bounds in some regimes

Conclusion & Open Problems

- We provide first communication bounds for a large class of discrete distributions
- Some open problems:
  - Tighten upper and lower bounds in some regimes
  - Other classes of distributions - densities, etc.

Conclusion & Open Problems

- We provide first communication bounds for a large class of discrete distributions
- Some open problems:
  - Tighten upper and lower bounds in some regimes
  - Other classes of distributions - densities, etc.
  - Multivariate distribution estimation


Blackboard
*HANK****
TH**K **U
****K YOU
Referee
THANK YOU
HANK
THK U
K YOU
Machine 1
Machine 2
. . .
Machine m