Communication-Efficient Distributed Algorithms for Density Estimation

Abhiram Natarajan

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

Outline

Motivation and Problem Definition

Learning Unstructured Distributions in ℓ_1

Learning k-Histograms in ℓ_2

Other Results

Conclusion

Can Distributed Cooks make Good Broth?

- too much data to store on one machine

Source: Google Images

Can Distributed Cooks make Good Broth?

- too much data to store on one machine

Source: Google Images

- distributed computation is necessary

Can Distributed Cooks make Good Broth?

- too much data to store on one machine

Source: Google Images

- distributed computation is necessary
- need communication-efficient distributed algorithms

Our Work

- study density estimation - a fundamental statistical task

Our Work

- study density estimation - a fundamental statistical task
- communication-efficient algorithms vs instrinsic limits

Our Work

- study density estimation - a fundamental statistical task
- communication-efficient algorithms vs instrinsic limits
- obtain optimal and near-optimal algorithms in a variety of settings

Our Work

- study density estimation - a fundamental statistical task
- communication-efficient algorithms vs instrinsic limits
- obtain optimal and near-optimal algorithms in a variety of settings
- time-complexity vs sample-complexity vs communication-complexity

Density Estimation

- draw samples from unknown distributon (*target* distribution)

Density Estimation

- draw samples from unknown distributon (*target* distribution)

- run algorithm on samples to output *hypothesis* distribution

Density Estimation

- draw samples from unknown distributon (*target* distribution)
- run algorithm on samples to output *hypothesis* distribution
- hope hypothesis is *close* to target distribution

Density Estimation

- draw samples from unknown distributon (*target* distribution)

  $\mathcal{D}$ family of distributions over $[n]$, $P \in \mathcal{D}$ target distribution
  draw m i.i.d. samples $X_1, \ldots, X_m$ from P

- run algorithm on samples to output *hypothesis* distribution

- hope hypothesis is *close* to target distribution

Density Estimation

- draw samples from unknown distributon (*target* distribution)

  $\mathcal{D}$ family of distributions over $[n]$, $P \in \mathcal{D}$ target distribution
  draw m i.i.d. samples $X_1, \ldots, X_m$ from P

- run algorithm on samples to output *hypothesis* distribution

  $\theta : [n]^m \to \mathcal{D}$ estimator
  output hypothesis distribution $\hat{P} = \theta(X_1, \ldots, X_m)$

- hope hypothesis is *close* to target distribution

Density Estimation

- draw samples from unknown distributon (*target* distribution)

  $\mathcal{D}$ family of distributions over $[n]$, $P \in \mathcal{D}$ target distribution
  draw m i.i.d. samples $X_1, \ldots, X_m$ from P

- run algorithm on samples to output *hypothesis* distribution

  $\theta : [n]^m \to \mathcal{D}$ estimator
  output hypothesis distribution $\hat{P} = \theta(X_1, \ldots, X_m)$

- hope hypothesis is *close* to target distribution

  error metric $d : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$
  $d(\hat{P}, P)$ must be low

Error in Density Estimation

- ℓ_1 and ℓ_2 error: $d(\hat{P}, P) := \|\hat{P} - P\|_{\rho \in \{1,2\}}$

Error in Density Estimation

- ℓ_1 and ℓ_2 error: $d(\hat{P}, P) := \|\hat{P} - P\|_{\rho \in \{1,2\}}$
- $\hat{P}$ is a random variable, so is $\|\hat{P} - P\|_{\rho}$

Error in Density Estimation

- ℓ_1 and ℓ_2 error: $d(\hat{P}, P) := \|\hat{P} - P\|_{\rho \in \{1,2\}}$
- $\hat{P}$ is a random variable, so is $\|\hat{P} - P\|_\rho$
- fix error $\varepsilon \in (0, 1)$ we need

$$\mathsf{E}\left[\|\hat{P} - P\|_\rho\right] \leqslant \varepsilon$$

Error in Density Estimation

- ℓ_1 and ℓ_2 error: $d(\hat{P}, P) := \|\hat{P} - P\|_{\rho \in \{1,2\}}$
- $\hat{P}$ is a random variable, so is $\|\hat{P} - P\|_\rho$
- fix error $\varepsilon \in (0, 1)$ we need

$$\mathsf{E}\left[\|\hat{P} - P\|_\rho\right] \leqslant \varepsilon$$

called Den-Est$(\mathcal{D}, \varepsilon, \ell_\rho)$ problem

Sample Complexity of Density Estimation

Definition

$m_1 = m_1(n, \varepsilon)$ is sufficient sample size for Den-Est$(\mathcal{D}, \varepsilon, \ell_\rho)$

- there exists algorithm $\mathcal{A}_\mathcal{D}$ takes m_1 samples and

$$\mathsf{E}\left[\|\hat{P} - P\|_\rho\right] \leqslant \varepsilon \qquad \forall\, P \in \mathcal{D}$$

Sample Complexity of Density Estimation

Definition

$m_1 = m_1(n, \varepsilon)$ is sufficient sample size for Den-Est$(\mathcal{D}, \varepsilon, \ell_\rho)$

- there exists algorithm $\mathcal{A}_\mathcal{D}$ takes m_1 samples and

$$\mathsf{E}\left[\|\hat{P} - P\|_\rho\right] \leqslant \varepsilon \qquad \forall\, P \in \mathcal{D}$$

$m_2 = m_2(n, \varepsilon)$ is necessary sample size for Den-Est$(\mathcal{D}, \varepsilon, \ell_\rho)$

- any conceivable algorithm must take m_2 samples

Communication Complexity

- communication complexity introduced by [Yao, 1979]:
  - players contain information $X_1, \ldots, X_n$ known only to them
  - communicate to referee via a protocol to compute $f(X_1, \ldots, X_n)$
  - we care about number of bits communicated

Communication Complexity

- communication complexity introduced by [Yao, 1979]:
  - players contain information $X_1, \ldots, X_n$ known only to them
  - communicate to referee via a protocol to compute $f(X_1, \ldots, X_n)$
  - we care about number of bits communicated
- communication complexity - pratical and more!

Communication Complexity

- communication complexity introduced by [Yao, 1979]:
  - players contain information $X_1, \ldots, X_n$ known only to them
  - communicate to referee via a protocol to compute $f(X_1, \ldots, X_n)$
  - we care about number of bits communicated
- communication complexity - pratical and more!
- applications in seemingly unrelated complexity theory areas - turing machines, decision trees, geometric problems, etc.

Distributed Density Estimation

- α sufficient sample size for Den-Est($\mathcal{D}$, ε, ℓ_p)

Distributed Density Estimation

- α sufficient sample size for Den-Est($\mathcal{D}, \varepsilon, \ell_\rho$)
- distribute α samples from some $P \in \mathcal{D}$ amongst m machines each machine gets $s = \frac{\alpha}{m}$ samples

Distributed Density Estimation

- α sufficient sample size for Den-Est($\mathcal{D}$, ε, ℓ_ρ)
- distribute α samples from some $P \in \mathcal{D}$ amongst m machines
  each machine gets $s = \frac{\alpha}{m}$ samples
- machines communicate to a referee, transcript is Π

Distributed Density Estimation

- α sufficient sample size for Den-Est$(\mathcal{D}, \varepsilon, \ell_p)$
- distribute α samples from some $P \in \mathcal{D}$ amongst m machines each machine gets $s = \frac{\alpha}{m}$ samples
- machines communicate to a referee, transcript is Π
- referee runs algorithm on Π to output hypothesis distribution $\hat{P}$

$$\sup_{P \in \mathcal{D}} \mathsf{E}\left[\|\hat{P} - P\|_1\right] \leqslant \varepsilon$$

Distributed Density Estimation

- α sufficient sample size for Den-Est$(\mathcal{D}, \varepsilon, \ell_\rho)$
- distribute α samples from some $P \in \mathcal{D}$ amongst m machines each machine gets $s = \frac{\alpha}{m}$ samples
- machines communicate to a referee, transcript is Π
- referee runs algorithm on Π to output hypothesis distribution $\hat{P}$

$$\sup_{P \in \mathcal{D}} \mathbb{E}\left[\|\hat{P} - P\|_1\right] \leqslant \varepsilon$$

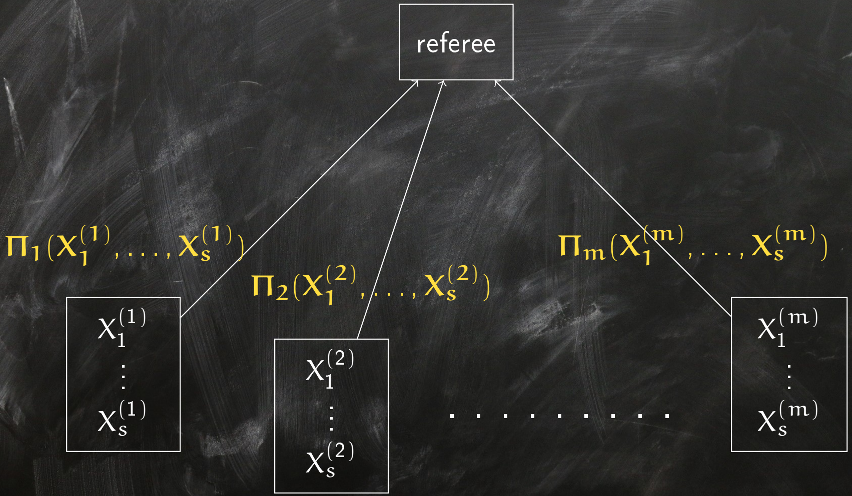
$\boxed{\text{called Dist-DE}(\mathcal{D}, m, \varepsilon, \ell_\rho) \text{ problem}}$

Communication Model – Simultaneous

referee

$$\boxed{\begin{matrix} X_1^{(1)} \\ \vdots \\ X_s^{(1)} \end{matrix}} \quad \boxed{\begin{matrix} X_1^{(2)} \\ \vdots \\ X_s^{(2)} \end{matrix}} \quad \cdots\cdots\cdots \quad \boxed{\begin{matrix} X_1^{(m)} \\ \vdots \\ X_s^{(m)} \end{matrix}}$$

Communication Model – Simultaneous



Communication Model – Interactive

blackboard (ξ)

referee

$X_1^{(1)}$
$\vdots$
$X_s^{(1)}$

$X_1^{(2)}$
$\vdots$
$X_s^{(2)}$

. . . . . .

$X_1^{(m)}$
$\vdots$
$X_s^{(m)}$

Communication Model – Interactive

blackboard (ξ)

referee

$\Pi_{1,1}(\xi, X_1^{(1)}, \ldots, X_s^{(1)})$

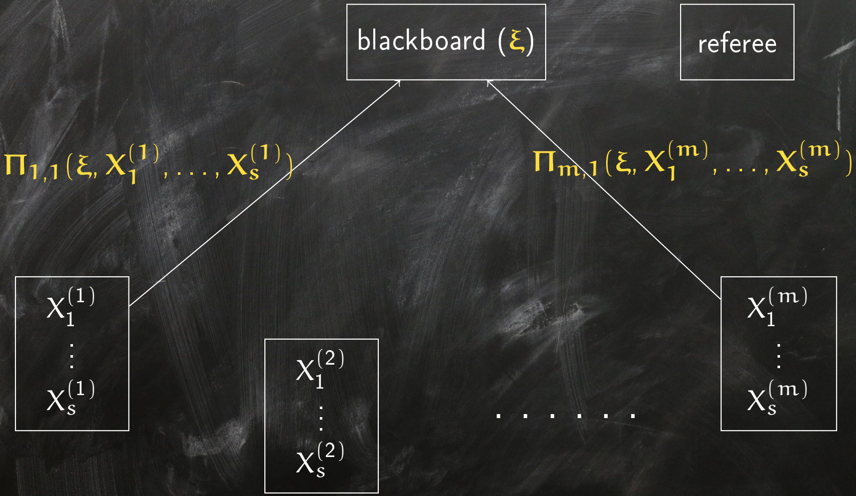
$X_1^{(1)}$
$\vdots$
$X_s^{(1)}$

$X_1^{(2)}$
$\vdots$
$X_s^{(2)}$

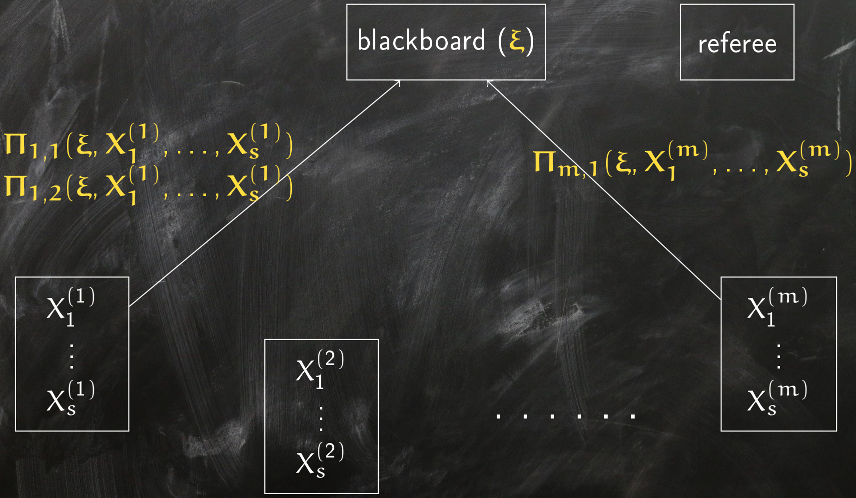
$\ldots\ldots$

$X_1^{(m)}$
$\vdots$
$X_s^{(m)}$

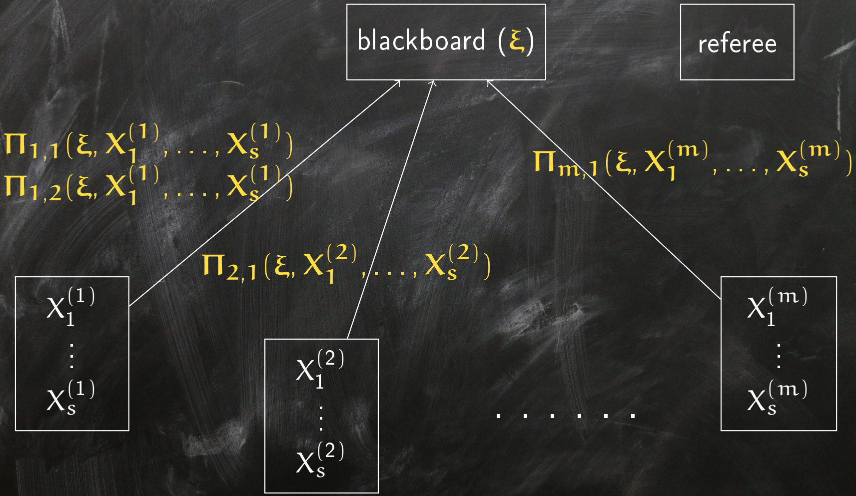
Communication Model – Interactive



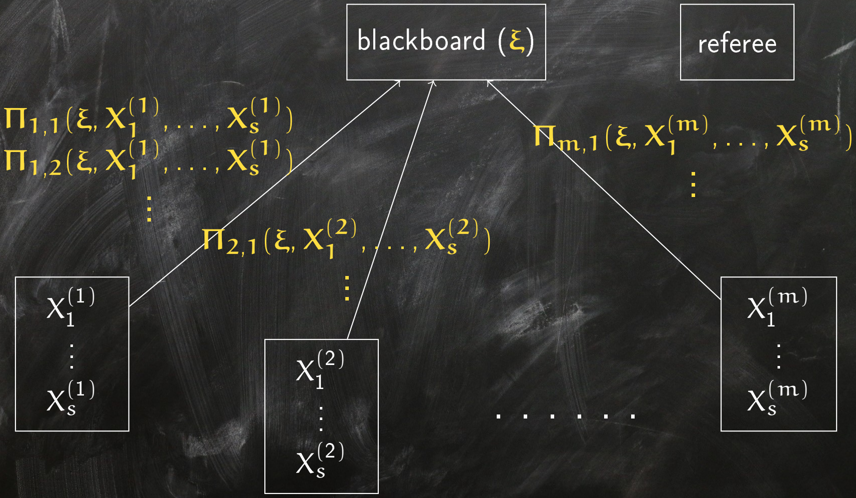
Communication Model – Interactive



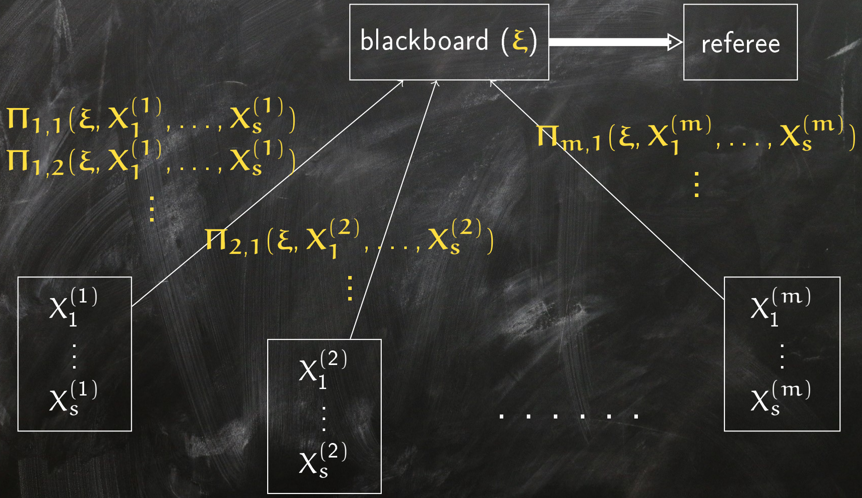
Communication Model – Interactive



Communication Model – Interactive



Communication Model – Interactive



Main Conceptual Messages

- when unstructured, naive protocol is best we can do

Main Conceptual Messages

- when unstructured, naive protocol is best we can do
- when structure is present (k-histograms, monotone), can be exploited for non-trivial improvement

Communication Complexity of Density Estimation

Definition

Protocol $\Pi_{\mathcal{D}}$ solves Dist-DE$(\mathcal{D}, m, \varepsilon, \ell_\rho)$ with $\beta_1 := \mathcal{CC}(\Pi_{\mathcal{D}})$ bits

- in $\Pi_{\mathcal{D}}$ machines communicate at most β_1 bits and referee outputs hypothesis

$$\mathrm{E}\left[\|\hat{P} - P\|_\rho\right] \leqslant \varepsilon \qquad \forall P \in \mathcal{D}$$

Communication Complexity of Density Estimation

Definition

Protocol $\Pi_{\mathcal{D}}$ solves Dist-DE$(\mathcal{D}, m, \varepsilon, \ell_\rho)$ with $\beta_1 := \mathcal{CC}(\Pi_{\mathcal{D}})$ bits

- in $\Pi_{\mathcal{D}}$ machines communicate at most β_1 bits and referee outputs hypothesis

$$\mathsf{E}\left[\|\hat{P} - P\|_\rho\right] \leqslant \varepsilon \qquad \forall P \in \mathcal{D}$$

$\beta_2 := \mathcal{CC}(\text{Dist-DE}(\mathcal{D}, m, \varepsilon, \ell_\rho))$

- any conceivable protocol must take β_2 bits to solve Dist-DE$(\mathcal{D}, m, \varepsilon, \ell_\rho)$

Outline

Motivation and Problem Definition

Learning Unstructured Distributions in ℓ_1

Learning k-Histograms in ℓ_2

Other Results

Conclusion

Folklore Result in Density Estimation

Theorem (Learning unstructured dists. in ℓ_1)

$\mathcal{D}_n$ - *unstructured distributions over* $[n]$. *For* Den-Est$(\mathcal{D}_n, \varepsilon, \ell_1)$

- $m_1 = O\left(\frac{n}{\varepsilon^2}\right)$ *is sufficient sample size*

Folklore Result in Density Estimation

Theorem (Learning unstructured dists. in ℓ_1)

$\mathcal{D}_n$ - *unstructured distributions over* $[n]$. *For* Den-Est$(\mathcal{D}_n, \varepsilon, \ell_1)$

- $m_1 = O\left(\frac{n}{\varepsilon^2}\right)$ *is sufficient sample size*
- $m_2 = \Omega\left(\frac{n}{\varepsilon^2}\right)$ *is necessary sample size*

Folklore Result in Density Estimation

Theorem (Learning unstructured dists. in ℓ_1)

$\mathcal{D}_n$ - *unstructured distributions over* $[n]$. *For* Den-Est$(\mathcal{D}_n, \varepsilon, \ell_1)$

- $m_1 = O\left(\frac{n}{\varepsilon^2}\right)$ *is sufficient sample size*
- $m_2 = \Omega\left(\frac{n}{\varepsilon^2}\right)$ *is necessary sample size*

Moreover, algorithm $\mathcal{A}_{\mathcal{D}_n}$ *outputs empirical distribution of samples*

$$\hat{P}(i) = \frac{\textit{number of } i \textit{ amongst samples}}{m_1} \qquad \forall i \in [n]$$

Communication Upper and Lower Bounds

$\mathcal{D}_n$ - unstructured distributions over $[n]$
$\alpha = \frac{cn}{\varepsilon^2}$ is sufficient sample size for Den-Est$(\mathcal{D}_n, \varepsilon, \ell_1)$

Communication Upper and Lower Bounds

$\mathcal{D}_n$ - unstructured distributions over $[n]$

$\alpha = \frac{cn}{\varepsilon^2}$ is sufficient sample size for Den-Est$(\mathcal{D}_n, \varepsilon, \ell_1)$

Theorem (Communication upper bound)

There exists trivial protocol $\Pi_{\mathcal{D}_n}$ *solves* Dist-DE$(\mathcal{D}_n, m, \varepsilon, \ell_1)$ *with*

$$\mathcal{CC}(\Pi_{\mathcal{D}_n}) = O\left(\frac{n}{\varepsilon^2} \log n\right),$$

for all $1 \leqslant m \leqslant \alpha$.

Communication Upper and Lower Bounds

$\mathcal{D}_n$ - unstructured distributions over $[n]$

$\alpha = \frac{cn}{\varepsilon^2}$ is sufficient sample size for Den-Est$(\mathcal{D}_n, \varepsilon, \ell_1)$

Theorem (Communication upper bound)

There exists trivial protocol $\Pi_{\mathcal{D}_n}$ solves Dist-DE$(\mathcal{D}_n, m, \varepsilon, \ell_1)$ *with*

$$\mathcal{CC}(\Pi_{\mathcal{D}_n}) = O\left(\frac{n}{\varepsilon^2}\log n\right),$$

for all $1 \leqslant m \leqslant \alpha$.

$\Pi_{\mathcal{D}_n}$ *just makes every machine send it's sample using* $\log n$ *bits.*

Communication Upper and Lower Bounds

$\mathcal{D}_n$ - unstructured distributions over $[n]$
$\alpha = \frac{cn}{\varepsilon^2}$ is sufficient sample size for Den-Est$(\mathcal{D}_n, \varepsilon, \ell_1)$

Theorem (Communication upper bound)

There exists trivial protocol $\Pi_{\mathcal{D}_n}$ *solves* Dist-DE$(\mathcal{D}_n, m, \varepsilon, \ell_1)$ *with*

$$\mathcal{CC}(\Pi_{\mathcal{D}_n}) = O\left(\frac{n}{\varepsilon^2}\log n\right),$$

for all $1 \leqslant m \leqslant \alpha$.
$\Pi_{\mathcal{D}_n}$ *just makes every machine send it's sample using* $\log n$ *bits.*

Theorem (Communication lower bound)

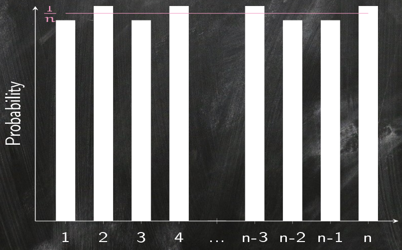
$$\mathcal{CC}(\text{Dist-DE}(\mathcal{D}, \alpha, \varepsilon, \ell_\rho)) = \Omega\left(\frac{n}{\varepsilon^2}\log n\right).$$

Lower Bound Proof Ideas

- construct family of *nearly uniform* distributions on $[n]$: for elements $2i-1$ and $2i$, probabilities are $\frac{1+100\delta_i\varepsilon}{n}$ and $\frac{1-100\delta_i\varepsilon}{n}$, δ_i uniform on $\{-1,1\}$

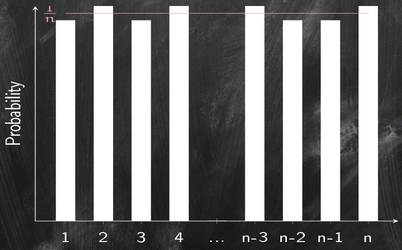
Lower Bound Proof Ideas

- construct family of *nearly uniform* distributions on $[n]$: for elements $2i-1$ and $2i$, probabilities are $\frac{1+100\delta_i\varepsilon}{n}$ and $\frac{1-100\delta_i\varepsilon}{n}$, δ_i uniform on $\{-1,1\}$



Lower Bound Proof Ideas

- construct family of *nearly uniform* distributions on $[n]$: for elements $2i-1$ and $2i$, probabilities are $\frac{1+100\delta_i\varepsilon}{n}$ and $\frac{1-100\delta_i\varepsilon}{n}$, δ_i uniform on $\{-1,1\}$



- learning distribution is equivalent to learning $\{\delta_i\}$

Lower Bound Proof Ideas

- contradiction: there is protocol sends $o\left(\frac{n}{\varepsilon^2}\log n\right)$ bits
  - can't send too many long messages
  - can't send too many short messages with few repetitions
  - there must be lots of repetitions

Lower Bound Proof Ideas

- contradiction: there is protocol sends $o\left(\frac{n}{\varepsilon^2}\log n\right)$ bits
  - can't send too many long messages
  - can't send too many short messages with few repetitions
  - there must be lots of repetitions
- information content in message is only $O(\varepsilon^2/t)$ when t repetitions, while coin toss provides $\Theta(\varepsilon^2)$ information

Lower Bound Proof Ideas

- contradiction: there is protocol sends $o\left(\frac{n}{\varepsilon^2}\log n\right)$ bits
  - can't send too many long messages
  - can't send too many short messages with few repetitions
  - there must be lots of repetitions
- information content in message is only $O(\varepsilon^2/t)$ when t repetitions, while coin toss provides $\Theta(\varepsilon^2)$ information
- less information means more error (Fano's inequality)

Other Regimes of Unstructured Dists. in ℓ_1

samp. per mach.	lower bound	upper bound
1	$\Omega\left(\frac{n}{\varepsilon^2}\log n\right)$	$O\left(\frac{n}{\varepsilon^2}\log n\right)$
$s=\Theta\left(\frac{n}{\varepsilon}\right)$	$\Omega\left(n\log\frac{1}{\varepsilon}\right)$	$O\left(\frac{n}{\varepsilon}\log\frac{1}{\varepsilon}\right)$
$s=\Theta\left(\frac{n}{\varepsilon^2}\right)$	$\Omega\left(n\log\frac{1}{\varepsilon}\right)$	$O\left(n\log\frac{1}{\varepsilon}\right)$

Outline

k-Histogram Distributions

- k-histogram over $[n]$ is a probability distribution that is piecewise constant over some set of k intervals over $[n]$

k-Histogram Distributions

- k-histogram over $[n]$ is a probability distribution that is piecewise constant over some set of k intervals over $[n]$

- $\Theta\left(\frac{k}{\varepsilon^2}\right)$ samples necessary and sufficient

k-Histogram Distributions

- k-histogram over $[n]$ is a probability distribution that is piecewise constant over some set of k intervals over $[n]$

- $\Theta\left(\frac{k}{\varepsilon^2}\right)$ samples necessary and sufficient
- when partition known, reduces to unstructured $\Theta(\frac{k}{\varepsilon^2}\log k)$ bits

k–Histogram Distributions

- k-histogram over $[n]$ is a probability distribution that is piecewise constant over some set of k intervals over $[n]$

- $\Theta\left(\frac{k}{\varepsilon^2}\right)$ samples necessary and sufficient
- when partition known, reduces to unstructured $\Theta(\frac{k}{\varepsilon^2}\log k)$ bits
- when partition unknown, trivial protocol uses too much communication $\Theta(\frac{k}{\varepsilon^2}\log n)$ bits

Learning k-histograms in ℓ_2

- at each step, algorithm maintains a partition of $[n]$

Learning k-Histograms in ℓ_2

- at each step, algorithm maintains a partition of $[n]$
- in every iteration splits partition at lowest error point

Learning k-Histograms in ℓ_2

- at each step, algorithm maintains a partition of $[n]$
- in every iteration splits partition at lowest error point
- returns flattening over final partition

Learning k–Histograms in ℓ_2

- at each step, algorithm maintains a partition of $[n]$
- in every iteration splits partition at lowest error point
- returns flattening over final partition

Key idea to approximate error:

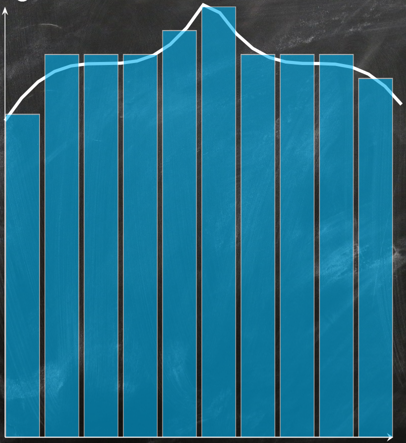
- server i has access to $x_i \in \mathbb{R}^n$ (vector of counts)
- using [Johnson and Lindenstrauss, 1984] lemma, get accurate estimate of $\|x\|_2^2$, where $x = \sum_i x_i$

Some Regimes for k–Histograms in ℓ_2

ε	samp. per mach.	lower bound	upper bound
$\Theta\left(\frac{1}{\sqrt{k}}\right)$	$\leqslant \tilde{O}(k \log n)$	$\Omega(k \log \frac{n}{k} + \sqrt{k} \log k)$	$O(k \log n)$
	$> \tilde{O}(k \log n)$	$\Omega(k \log \frac{n}{k} + \sqrt{k} \log k)$	$\tilde{O}(\frac{k^2}{s} \log n)$

Histogram Approximation

Our algorithm is agnostic!



Outline

Motivation and Problem Definition

Learning Unstructured Distributions in ℓ_1

Learning k-Histograms in ℓ_2

Other Results

Conclusion

Other results

- near optimal bounds for distributed learning in all regimes for:
  - unstructured distributions in ℓ_2 (similar to ℓ_1)
  - k-histograms in ℓ_1 (quite different from ℓ_2, need to approximate $\mathcal{A}_k$ distance)
  - monotone distributions in ℓ_1 (uses Birgé oblivious decomposition [Birgé, 1987b, Birgé, 1987a])

Other results

- near optimal bounds for distributed learning in all regimes for:
  - unstructured distributions in ℓ_2 (similar to ℓ_1)
  - k-histograms in ℓ_1 (quite different from ℓ_2, need to approximate $\mathcal{A}_k$ distance)
  - monotone distributions in ℓ_1 (uses Birgé oblivious decomposition [Birgé, 1987b, Birgé, 1987a])
- our algorithms are agnostic

Other results

- near optimal bounds for distributed learning in all regimes for:
  - unstructured distributions in ℓ_2 (similar to ℓ_1)
  - k-histograms in ℓ_1 (quite different from ℓ_2, need to approximate $\mathcal{A}_k$ distance)
  - monotone distributions in ℓ_1 (uses Birgé oblivious decomposition [Birgé, 1987b, Birgé, 1987a])
- our algorithms are agnostic
- can be extended to a huge class of distributions - unimodal, O(1)-modal, log-concave, monotone hazard rate (MHR) distributions, certain piecewise-polynomial continuous distributions, etc.

Outline

Motivation and Problem Definition

Learning Unstructured Distributions in ℓ_1

Learning k-Histograms in ℓ_2

Other Results

Conclusion

Open Problems

Many obvious next questions:

- tighten bounds in regimes where not tight

Open Problems

Many obvious next questions:

- tighten bounds in regimes where not tight
- prove bounds for other classes of distributions - densities, etc.

Open Problems

Many obvious next questions:

- tighten bounds in regimes where not tight
- prove bounds for other classes of distributions - densities, etc.
- go from univariate to multivariate

Open Problems

Many obvious next questions:

- tighten bounds in regimes where not tight
- prove bounds for other classes of distributions - densities, etc.
- go from univariate to multivariate
- study distribution testing in this model

Open Problems

Many obvious next questions:

- tighten bounds in regimes where not tight
- prove bounds for other classes of distributions - densities, etc.
- go from univariate to multivariate
- study distribution testing in this model
- generalize bounds in terms of entropy of distribution

Open Problems

Many obvious next questions:

- tighten bounds in regimes where not tight
- prove bounds for other classes of distributions - densities, etc.
- go from univariate to multivariate
- study distribution testing in this model
- generalize bounds in terms of entropy of distribution
- more than sufficient sample, also unequal number of samples

Conclusion

- we provide first near-optimal bounds for a huge class of discrete distributions

Conclusion

- we provide first near-optimal bounds for a huge class of discrete distributions
- communication complexity of learning tasks - can it shed fundamental insights on the nature of learning?

References

Birgé, L. (1987a).
Estimating a density under order restrictions: Nonasymptotic minimax risk.
The Annals of Statistics, pages 995–1012.

Birgé, L. (1987b).
On the risk of histograms for estimating decreasing densities.
The Annals of Statistics, pages 1013–1022.

Diakonikolas, I. (2016).
Learning structured distributions.
In Bühlmann, P., Drineas, P., Kane, M., and van Der Laan, M., editors, *Handbook of Big Data*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, chapter 15, pages 267–284. Taylor & Francis.

Diakonikolas, I., Grigorescu, E., Li, J., Natarajan, A., Onak, K., and Schmidt, L. (2017).
Communication-efficient distributed learning of discrete distributions.
To appear.

Johnson, W. B. and Lindenstrauss, J. (1984).
Extensions of lipschitz mappings into a hilbert space.
Contemporary mathematics, 26(189-206):1–1.

Yao, A. (1979).
Some complexity questions related to distributive computing(preliminary report).
In *Proceedings of the Eleventh Annual ACM Symposium on Theory of Computing*, STOC '79, pages 209–213. ACM.