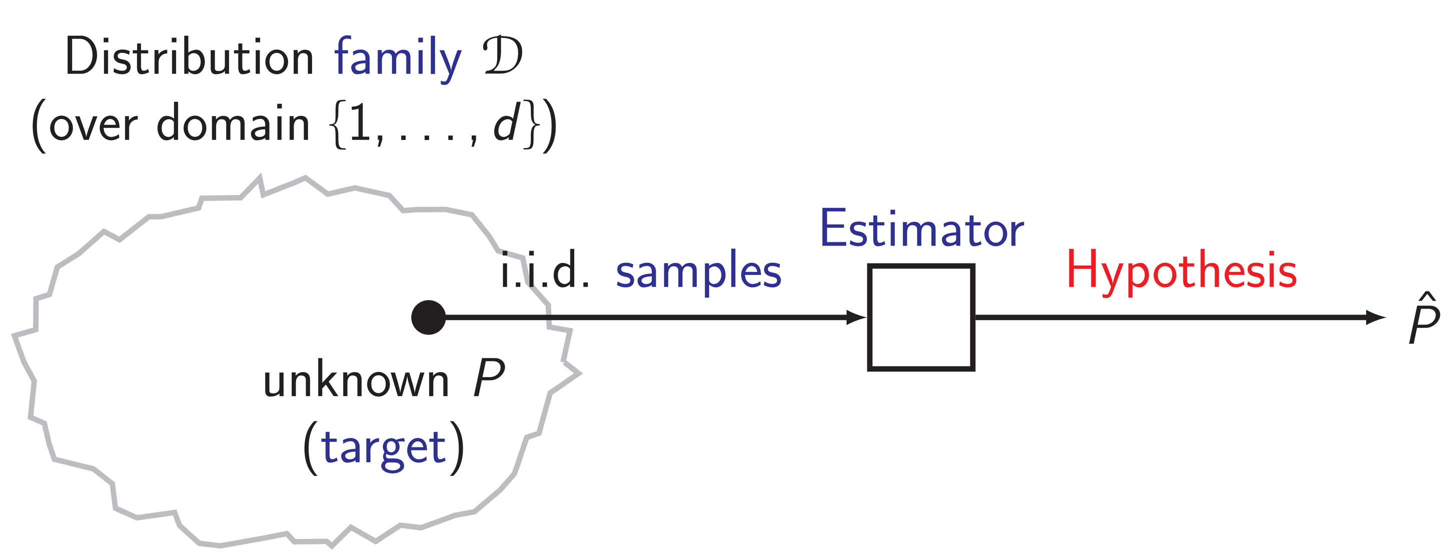
Communication-Efficient Distributed Learning of Discrete Distributions

Ilias Diakonikolas[1], Elena Grigorescu[2], Jerry Li[3], Abhiram Natarajan[2], Krzysztof Onak[4], Ludwig Schmidt[3]

[1]University of Southern California, [2]Purdue University, [3]MIT CSAIL, [4]IBM NY

Density Estimation

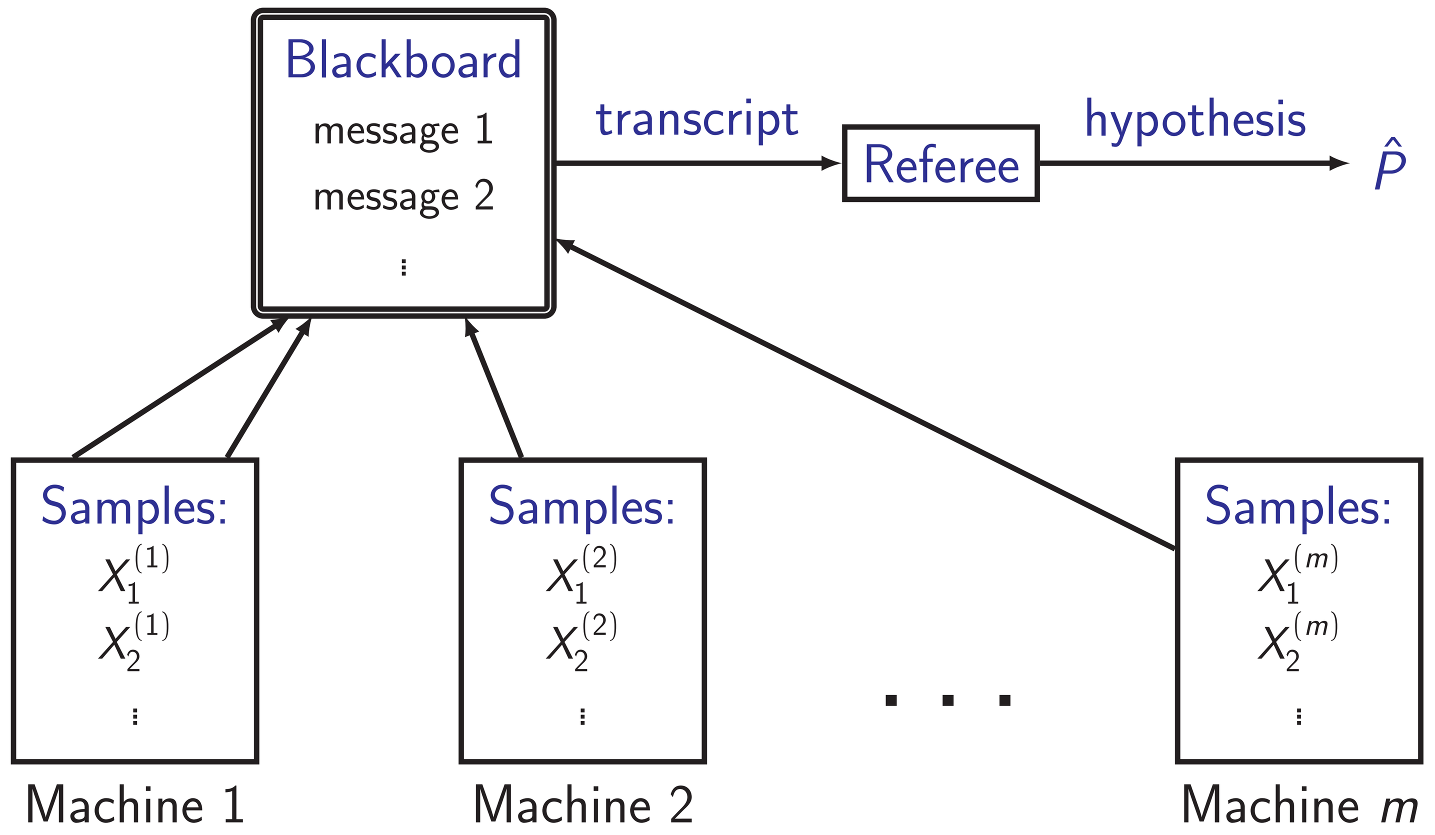
Goal: Accurately estimate a distribution from few samples.



For given $0 < \varepsilon < 1$, we want to achieve the following guarantee:

$$\mathbb{E}\left[\|\hat{P} - P\|_p\right] \leqslant \varepsilon, \qquad p \in \{1, 2\} .$$

Communication Model



Distributed Density Estimation

- Let $\mathcal{D}_d$ be the set of all discrete distributions over $\{1, \ldots, d\}$.
- Let n be a sufficient sample size for learning family $\mathcal{D} \subseteq \mathcal{D}_d$ upto error ε under the ℓ_p-distance.
- From unknown $P \in \mathcal{D}$, we draw s samples on $\frac{n}{s}$ machines each.
- Machines communicate according to protocol, yielding transcript Π.
- Referee runs an estimator θ on the transcript Π to output a hypothesis distribution $\hat{P} = \theta(\Pi)$.

Baseline: There exists a protocol to learn $\mathcal{D}$ up to error ε in ℓ_p-distance that uses $O(n \log d)$ bits of communication.

Our Contribution

We study the following settings (see paper for full results):

- Learning arbitrary distributions in ℓ_1 and ℓ_2 error
- Learning k-histograms in ℓ_1 and ℓ_2 error
- Learning monotone distributions in ℓ_1 error

Main message:

- **Without structural assumptions**, baseline is best possible.
- **When the distribution has stucture**, this can be leveraged to improve algorithms.

Learning Arbitrary Discrete Distributions in ℓ_1-error

Folklore fact: $\Theta(\frac{d}{\varepsilon^2})$ samples are necessary and sufficient to learn any discrete distribution up to ℓ_1-error ε.

$\rightarrow$ **Baseline** communication protocol that uses $O(\frac{d}{\varepsilon^2} \log d)$ bits of communication.

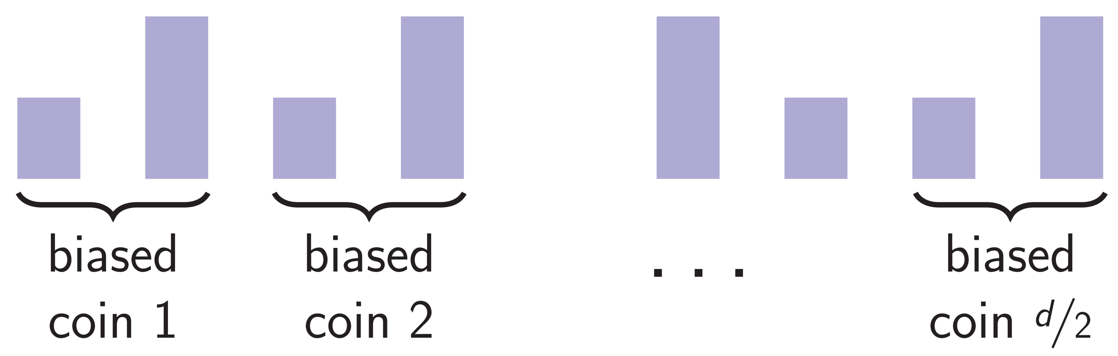
Theorem: Any protocol that learns any distribution from $\mathcal{D}_d$ upto ℓ_1 error ε must use $\Omega(\frac{d}{\varepsilon^2} \log d)$ bits of communication when there is one sample per machine.

Regime	Lower Bound	Upper Bound
$s = 1$	$\Omega(\frac{d}{\varepsilon^2} \log d)$	$O(\frac{d}{\varepsilon^2} \log d)$
$s = \Theta(d)$	$\Omega(d \log \frac{1}{\varepsilon})$	$O(\frac{d}{\varepsilon^2})$
$s = \Theta(\frac{d}{\varepsilon^2})$	$\Omega(d \log \frac{1}{\varepsilon})$	$O(d \log \frac{1}{\varepsilon})$

Lower Bound Proof Technique

- We construct a **hard to learn** distribution family as follows:

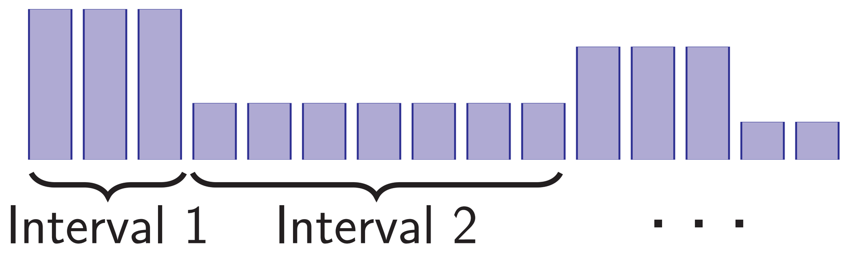
$$\mathbb{P}[2i-1] = \frac{1 + 100\delta_i\varepsilon}{d} \qquad \mathbb{P}[2i] = \frac{1 - 100\delta_i\varepsilon}{d} .$$



- The output of any good learning protocol can be used to learn the bias δ_i of most of the pairs.
- We show: If the messages sent are repeated for lots of different samples, the contribution of this player is not very helpful.
- A coin toss usually provides $\Theta(\varepsilon^2)$ information. We show that the information content is $O(\varepsilon^2/t)$, when there are t repetitions.

k-Histograms

- Piecewise-constant over a partition of $\{1, \ldots, d\}$ into k intervals.



- Many classes of distributions can be approximated by k-histograms
  $\rightarrow$ If we can robustly learn k-histograms, we can learn these distributions as well.

Upper Bounds for Robustly Learning k-Histograms in ℓ_2-error

Formal Problem Statement: Given $\varepsilon > 0$ and n i.i.d. samples from a distribution $P : \{1, \ldots, d\} \rightarrow \mathbb{R}$ evenly distributed over m machines, output a k-histogram $\widehat{h}$ so that

$$\mathbb{E}\left[\|\widehat{h} - P\|_2\right] \leqslant C \cdot \mathrm{OPT}_k + \varepsilon ,$$

where $\mathrm{OPT}_k := \min_{k\text{-histograms } h} \|h - P\|_2$.

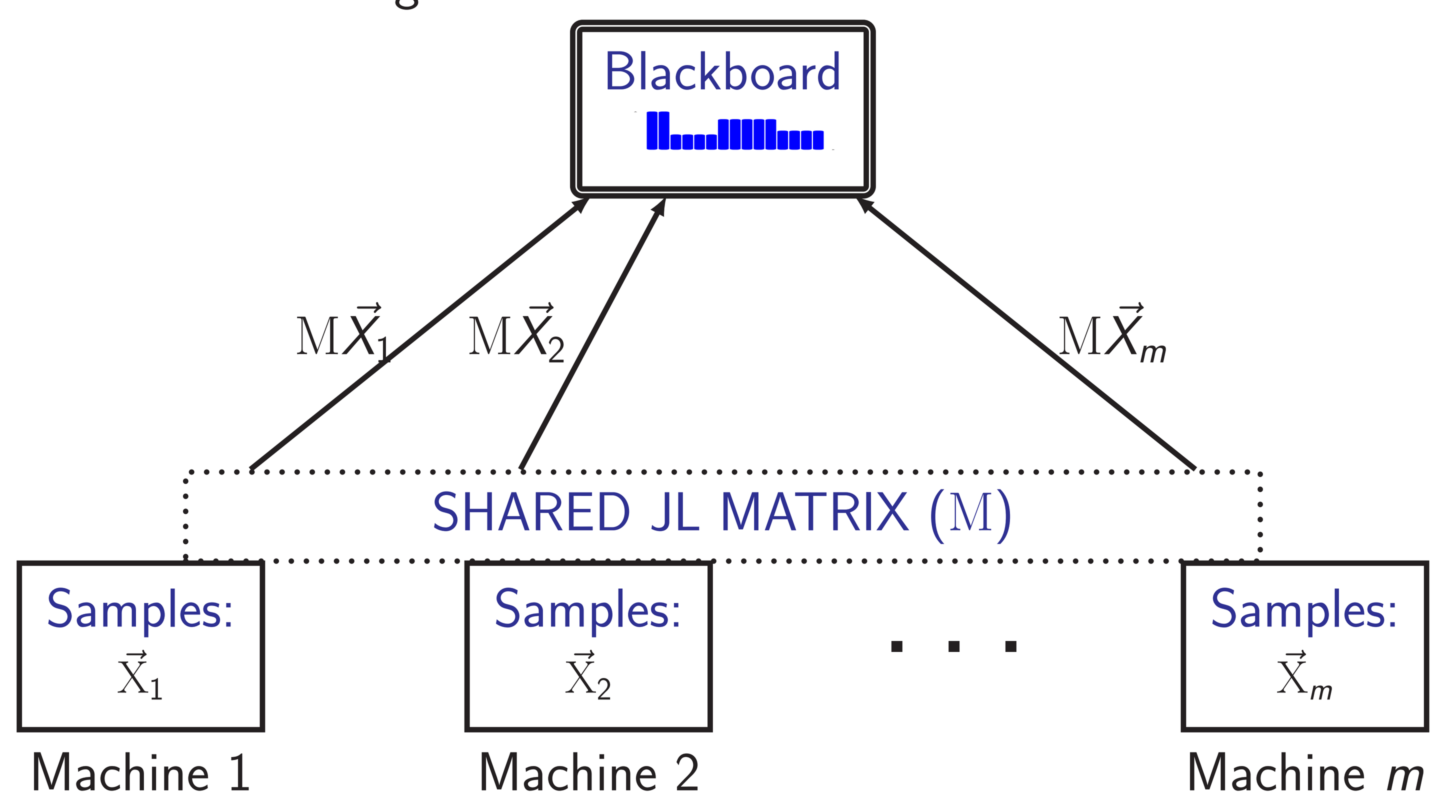
Theorem: For any $\varepsilon > 0$ there is an algorithm, which given $n = \Omega(1/\varepsilon^2)$ samples distributed over m machines, learns a k-histogram to ε error in ℓ_2 using $\widetilde{O}(mk \log \frac{1}{\varepsilon} \log d)$ bits of communication.

Our Techniques

Let $\widehat{P}$ be the empirical distribution. For any interval I, let

$$e(I) = \sum_{i \in I} \left(\widehat{P}(i) - \mathrm{avg}(\widehat{P}, I)\right)^2 .$$

Key insight: This can be computed with few bits of communication via linear sketching.



We design an algorithm that only interacts through the data via queries to $e(I)$ and uses few queries.

$\rightarrow$ **Communication-efficient** algorithms for learning k-histograms.