Communication Lower Bounds for Statistical Estimation Problems via a Distributed Data Processing Inequality

Mark Braverman Ankit Garg Tengyu Ma

Huy Nguyen David Woodruff

DIMACS Workshop on Big Data through the Lens of Sublinear Algorithms

Aug 28, 2015

DIMACS

Center for Discrete Mathematics & Theoretical Computer Science
Founded as a National Science Foundation Science and Technology Center

Distributed mean estimation

Statistical estimation:

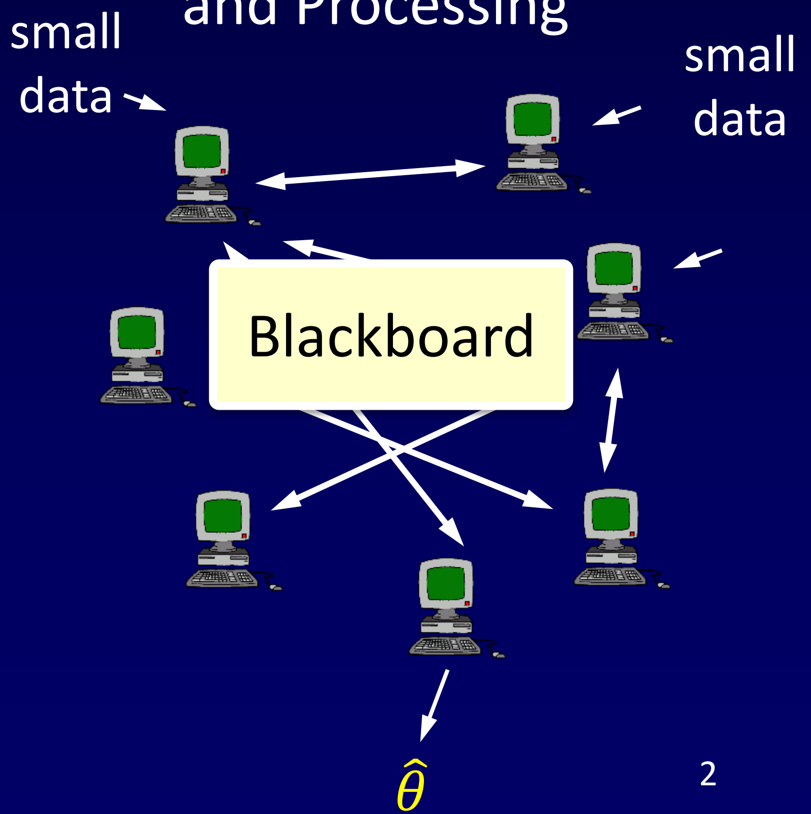
- Unknown parameter θ.
- Inputs to machines: i.i.d. data points $\sim D_\theta$.
- Output estimator $\hat{\theta}$.

Objectives:

- Low communication $C = |\Pi|$.
- Small loss

$$R := \mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right].$$

Big Data!
Distributed Storage and Processing



Goal: estimate $(\theta_1, \ldots, \theta_d)$

…buted sparse Gaussian …mean estimation

- Ambient dimension d.
- Sparsity parameter k: $\|\theta\|_0 \leq k$.
- Number of machines m.
- Each machine holds n samples.
- Standard deviation σ.
- Thus each sample is a vector $X_j^{(t)} \sim \left(\mathcal{N}(\theta_1, \sigma^2), \ldots, \mathcal{N}(\theta_d, \sigma^2)\right) \in \mathbb{R}^d$

Goal: estimate $(\theta_1, \ldots, \theta_d)$

Higher value makes estimation:

- Ambient dimension d. *harder*
- Sparsity parameter k: $\|\theta\|_0 \leq k$. *harder*
- Number of machines m. *easier**
- Each machine holds n samples. *easier*
- Standard deviation σ. *harder*
- Thus each sample is a vector $X_j^{(t)} \sim \left(\mathcal{N}(\theta_1, \sigma^2), \ldots, \mathcal{N}(\theta_d, \sigma^2)\right) \in \mathbb{R}^d$

Distributed sparse Gaussian mean estimation

Statistical limit

- Main result: if $|\Pi| = C$, then

$$R \geq \Omega\left(\max\left(\frac{\sigma^2 dk}{nC}, \frac{\sigma^2 k}{nm}\right)\right)$$

- Tight up to a $\log d$ factor [GMN14]. Up to a const. factor in the dense case.
- For optimal performance, $C \gtrsim md$ (not mk) is needed!

- d – dim
- k – sparsity
- m – machine
- n – samp. each
- σ – deviation
- R – sq. loss

Prior work (partial list)

- [Zhang-Duchi-Jordan-Wainwright'13]: the case when $d = 1$ and general communication; and the dense case for simultaneous-message protocols.
- [Shamir'14]: Implies the result for $k = 1$ in a restricted communication model.
- [Duchi-Jordan-Wainwright-Zhang'14, Garg-Ma-Nguyen'14]: the dense case (up to logarithmic factors).
- A lot of recent work on communication-efficient distributed learning.

Reduction from Gaussian mean detection

- $R \geq \Omega\left(\max\left(\frac{\sigma^2 dk}{nC}, \frac{\sigma^2 k}{nm}\right)\right)$
- Gaussian mean detection
  - A one-dimensional problem.
  - Goal: distinguish between $\mu_0 = \mathcal{N}(0, \sigma^2)$ and $\mu_1 = \mathcal{N}(\delta, \sigma^2)$.
  - Each player gets n samples.

- Assume $R \ll \max\left(\frac{\sigma^2 dk}{nC}, \frac{\sigma^2 k}{nm}\right)$
- Distinguish between $\mu_0 = \mathcal{N}(0, \sigma^2)$ and $\mu_1 = \mathcal{N}(\delta, \sigma^2)$.
- Theorem: If can attain $R \leq \frac{1}{16} k\delta^2$ in the estimation problem using C communication, then we can solve the detection problem at $\sim C/d$ *min-information cost.*
- Using $\delta^2 \ll \sigma^2 d/(C\, n)$, get detection using $I \ll \frac{\sigma^2}{n\, \delta^2}$ *min-information cost.*

The detection problem

- Distinguish between $\mu_0 = \mathcal{N}(0,1)$ and $\mu_1 = \mathcal{N}(\delta, 1)$.
- Each player gets n samples.
- Want this to be impossible using $I \ll \frac{1}{n\,\delta^2}$ *min-information cost.*

The detection problem

- ~~Distinguish between $\mu_0 = \mathcal{N}(0,1)$ and $\mu_1 = \mathcal{N}(\delta, 1)$.~~
- Distinguish between $\mu_0 = \mathcal{N}\left(0, \frac{1}{n}\right)$ and $\mu_1 = \mathcal{N}\left(\delta, \frac{1}{n}\right)$.
- Each player gets ~~n samples.~~ one sample.
- Want this to be impossible using $I \ll \frac{1}{n\,\delta^2}$ *min-information cost.*

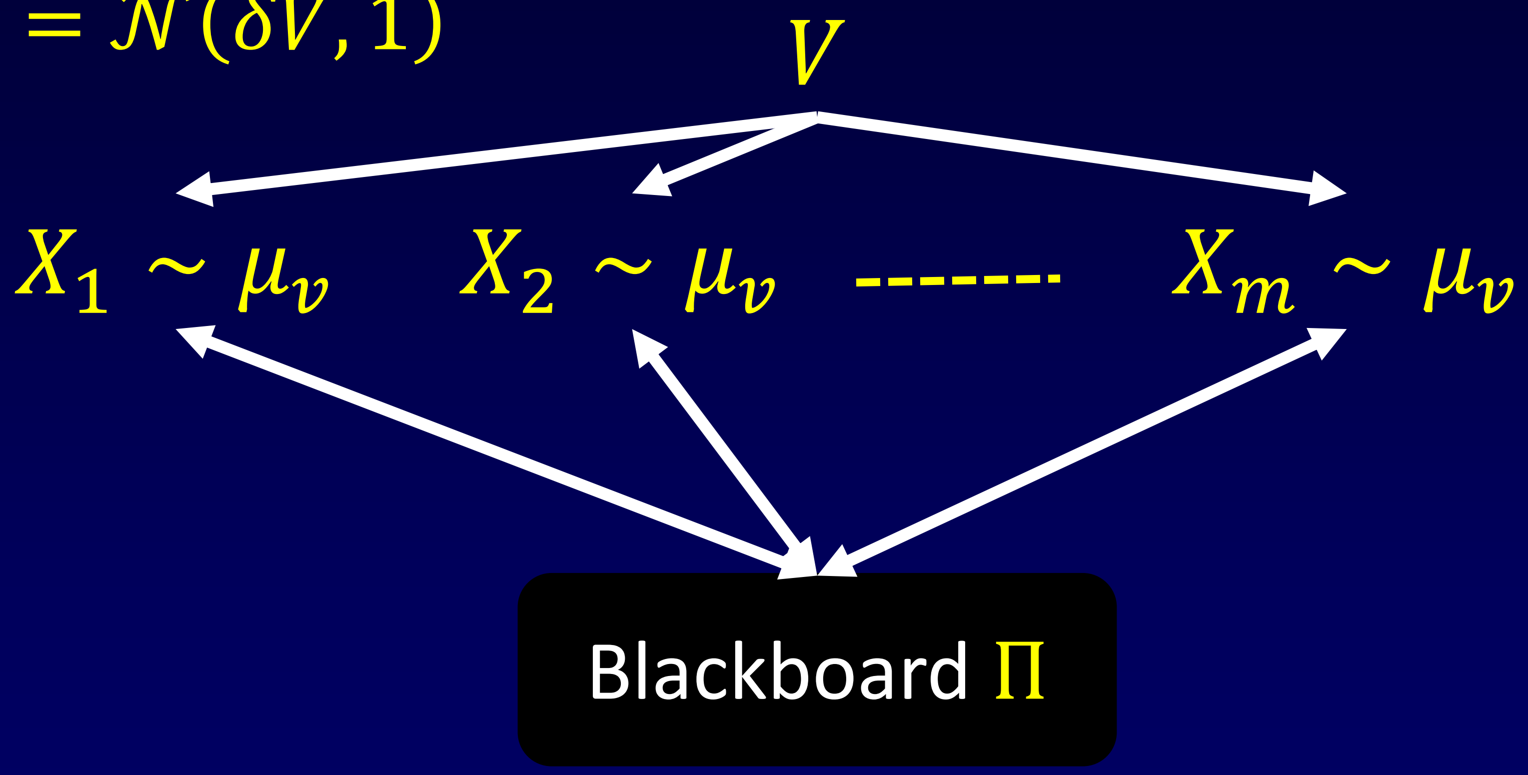
The detection problem

- By scaling everything by $\sqrt{n}$ (and replacing δ with $\delta\sqrt{n}$).
- Distinguish between $\mu_0 = \mathcal{N}(0,1)$ and $\mu_1 = \mathcal{N}(\delta, 1)$.
- Each player gets *one* sample.
- Want this to be impossible using $I \ll \frac{1}{\delta^2}$ *min-information cost.*

Tight (for m large enough, otherwise task impossible)

Information cost

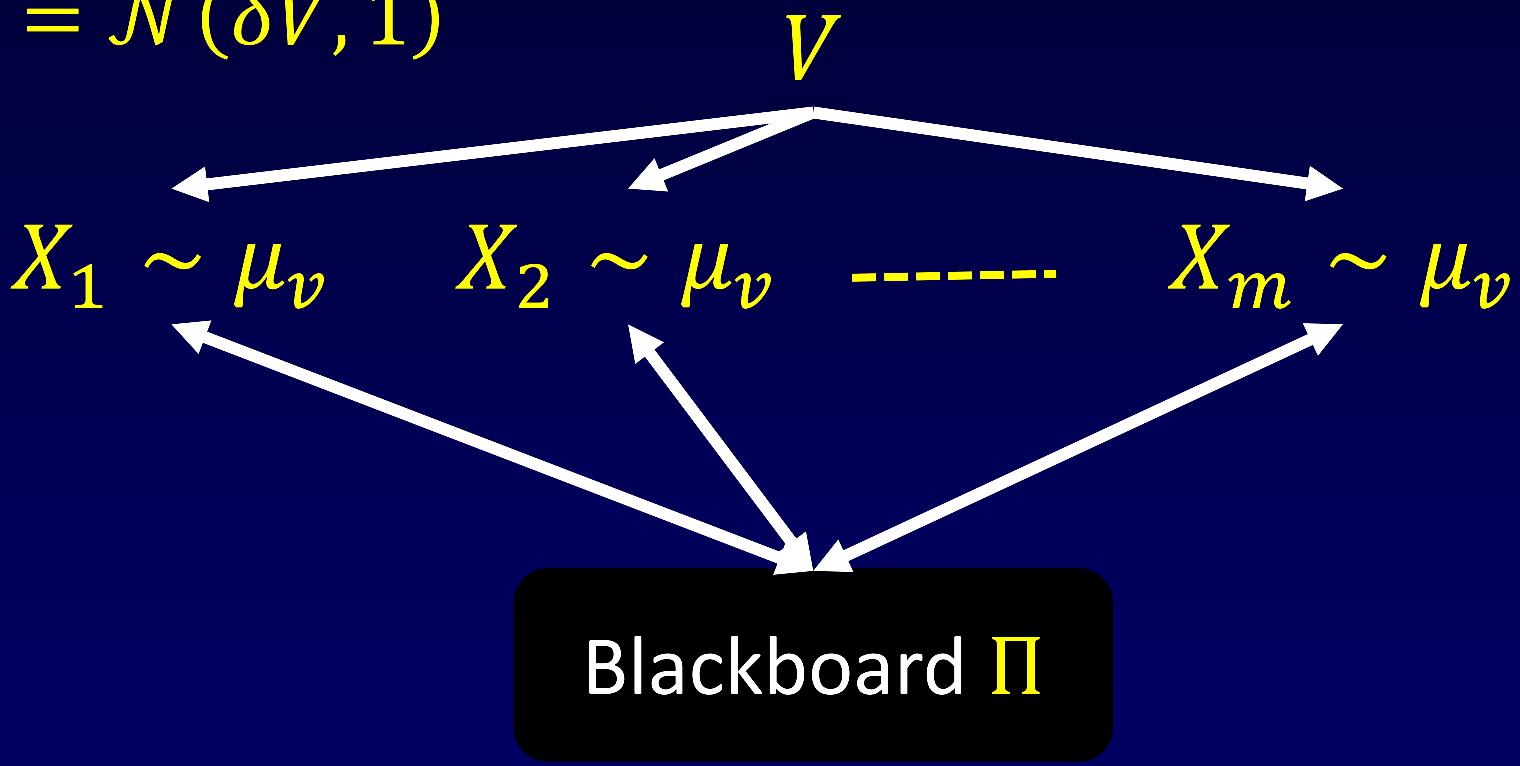
$\mu_v = \mathcal{N}(\delta V, 1)$



$$IC(\pi) := I(\Pi; X_1 X_2 \dots X_m)$$

Min-Information cost

$\mu_V = \mathcal{N}(\delta V, 1)$



$$minIC(\pi) := \min_{v \in \{0,1\}} I(\Pi; X_1 X_2 \dots X_m | V = v)$$

Min-Information cost

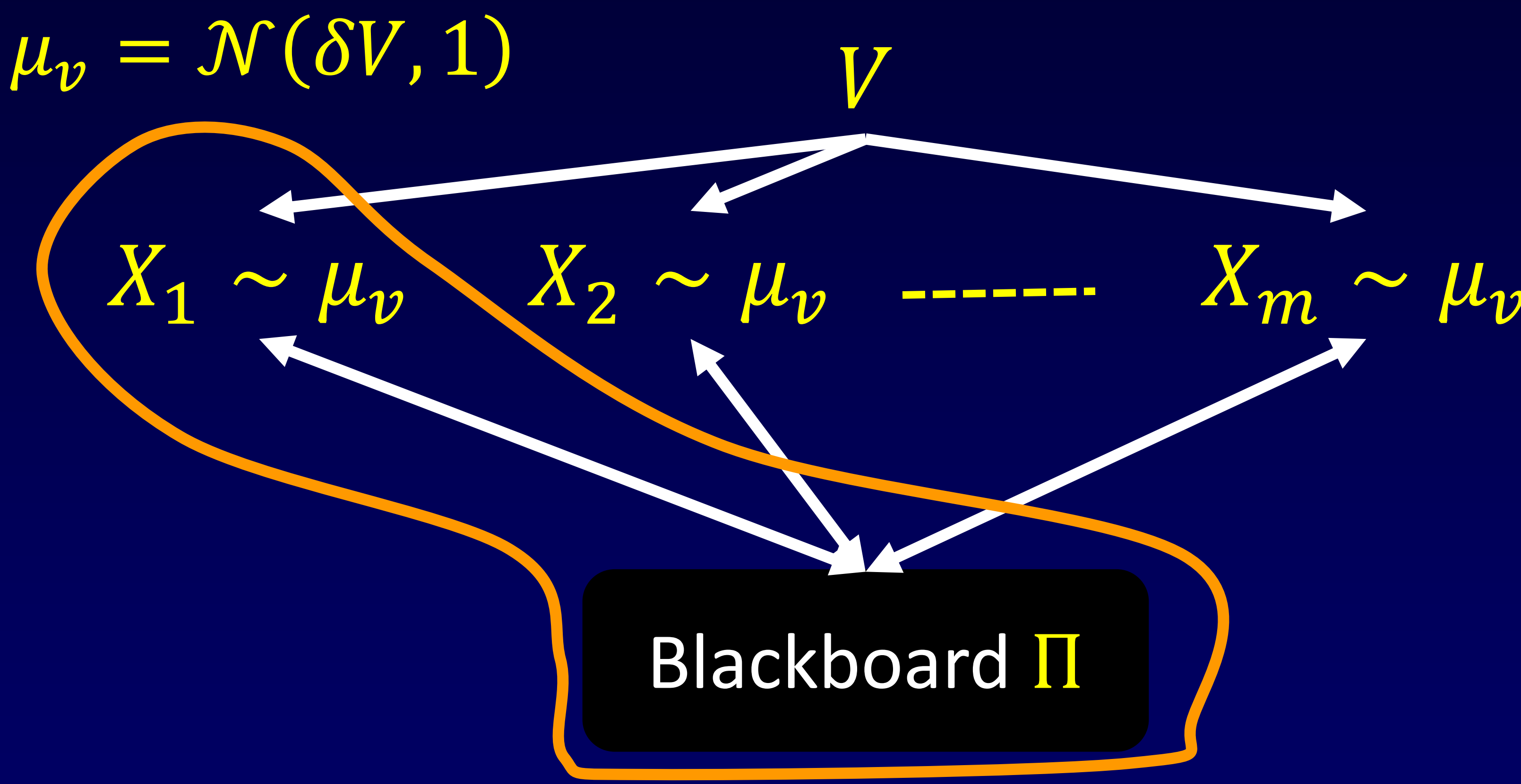
$$minIC(\pi) := \min_{v\in\{0,1\}} I(\Pi; X_1X_2\dots X_m|V=v)$$

- We will want this quantity to be $\Omega\left(\frac{1}{\delta^2}\right)$.
- Warning: it is not the same thing as $I(\Pi; X_1X_2\dots X_m|V) = \mathbb{E}_{v\sim V}\, I(\Pi; X_1X_2\dots X_m|V=v)$

because one case can be much smaller than the other.

- In our case, the need to use $minIC$ instead of IC happens because of the sparsity.

Strong data processing inequality



Fact: $|\Pi| \geq I(\Pi; X_1 X_2 \ldots X_m) = \sum_i I(\Pi; X_i | X_{<i})$

Strong data processing inequality

- $\mu_v = \mathcal{N}(\delta V, 1)$; suppose $V \sim B_{1/2}$.
- For each i, $V - X_i - \Pi$ is a Markov chain.
- Intuition: "X_i contains little information about V; no way to learn this information except by learning a lot about X_i".
- Data processing: $I(V; \Pi) \leq I(X_i; \Pi)$.
- *Strong* Data Processing: $I(V; \Pi) \leq \beta \cdot I(X_i; \Pi)$ for some $\beta = \beta(\mu_0, \mu_1) < 1$.

Strong data processing inequality

- $\mu_v = \mathcal{N}(\delta V, 1)$; suppose $V \sim B_{1/2}$.
- For each i, $V - X_i - \Pi$ is a Markov chain.
- *Strong* Data Processing: $I(V;\Pi) \le \beta \cdot I(X_i;\Pi)$ for some $\beta = \beta(\mu_0, \mu_1) < 1$.
- In this case ($\mu_0 = \mathcal{N}(0,1), \mu_1 = \mathcal{N}(\delta, 1)$):

$$\beta(\mu_0, \mu_1) \sim \frac{I\big(V; \mathrm{sign}(X_i)\big)}{I(X_i; \mathrm{sign}(X_i))} \sim \delta^2$$

“Proof”

- $\mu_v = \mathcal{N}(\delta V, 1)$; suppose $V \sim B_{1/2}$.
- *Strong* Data Processing: $I(V;\Pi) \leq \delta^2 \cdot I(X_i;\Pi)$
- We know $I(V;\Pi) = \Omega(1)$.

$$|\Pi| \geq I(\Pi; X_1 X_2 \ldots X_m) \gtrsim \sum_i I(\Pi; X_i) \geq \frac{1}{\delta^2} \ldots$$

$$\sum_i \text{"}Info\ \Pi\ conveys\ about\ V\ through\ player\ i\text{"} \gtrsim$$

$$\frac{1}{\delta^2} I(V;\Pi) = \Omega\left(\frac{1}{\delta^2}\right)$$

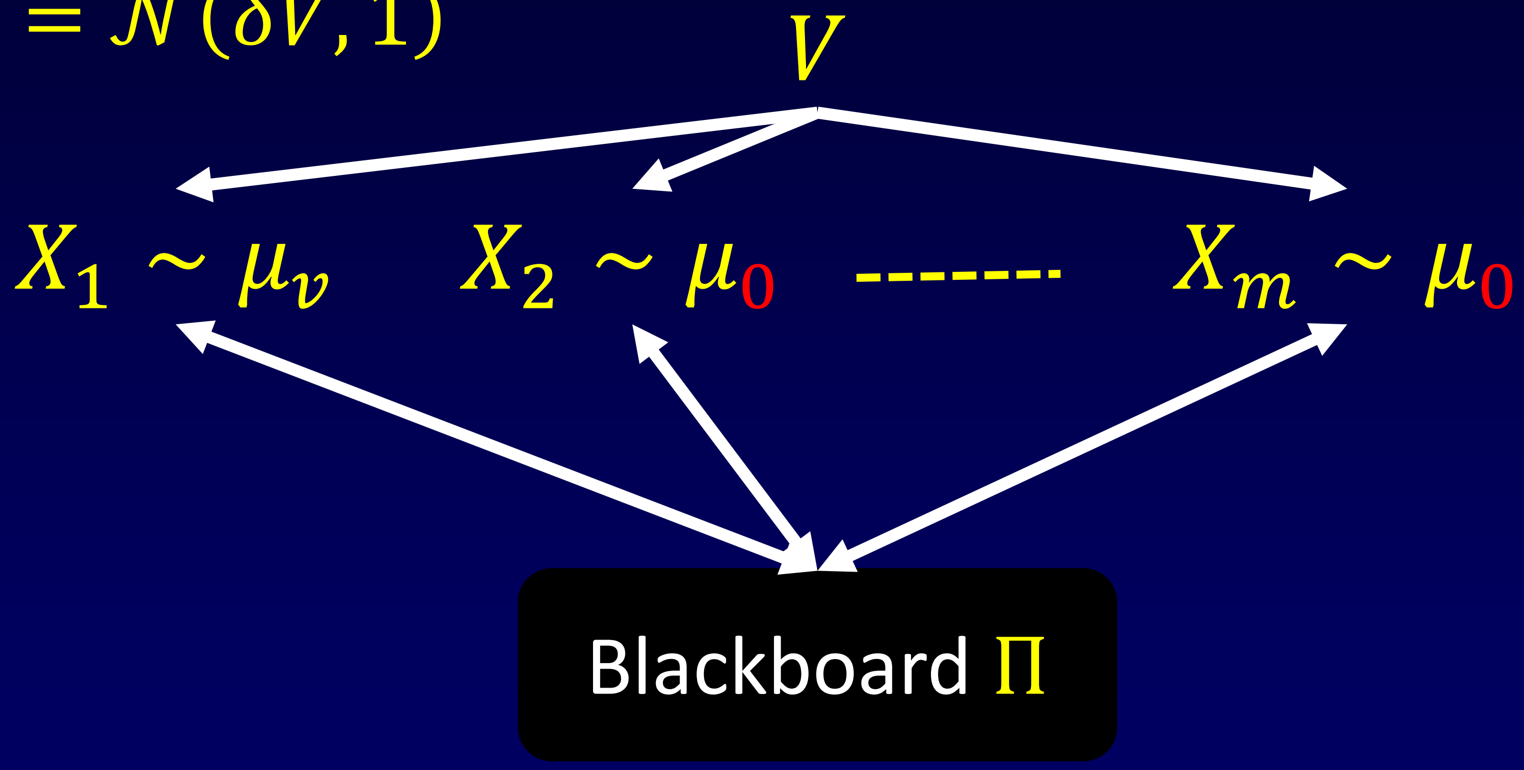
Q.E.D!

Issues with the proof

- The right high level idea.
- Two main issues:
  - Not clear how to deal with additivity over coordinates.
  - Dealing with $minIC$ instead of IC.

If the picture were this…

$\mu_v = \mathcal{N}(\delta V, 1)$



Then indeed $I(\Pi; V) \leq \delta^2 \cdot I(\Pi; X_1)$.

Hellinger distance

- Solution to additivity: using Hellinger distance $\int_\Omega \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx$
- Following from [Jayram'09].

$$h^2(\Pi_{V=0}, \Pi_{V=1}) \sim I(V; \Pi) = \Omega(1)$$

- $h^2(\Pi_{V=0}, \Pi_{V=1})$ decomposes into m scenarios as above using the fact that Π is a protocol.

$minIC$

- Dealing with $minIC$ is more technical. Recall:
- $minIC(\pi) := \min_{v \in \{0,1\}} I(\Pi; X_1 X_2 \ldots X_m | V = v)$
- Leads to our main technical statement: "Distributed Strong Data Processing Inequality"

Theorem: Suppose $\Omega(1) \cdot \mu_0 \leq \mu_1 \leq O(1) \cdot \mu_0$, and let $\beta(\mu_0, \mu_1)$ be the SDPI constant. Then

$$h^2(\Pi_{V=0}, \Pi_{V=1}) \leq O\big(\beta(\mu_0, \mu_1)\big) \cdot minIC(\pi)$$

Putting it together

Theorem: Suppose $\Omega(1) \cdot \mu_0 \leq \mu_1 \leq O(1) \cdot \mu_0$, and let $\beta(\mu_0, \mu_1)$ be the SDPI constant. Then

$$h^2(\Pi_{V=0}, \Pi_{V=1}) \leq O\big(\beta(\mu_0, \mu_1)\big) \cdot minIC(\pi)$$

- With $\mu_0 = \mathcal{N}(0,1)$, $\mu_1 = \mathcal{N}(\delta, 1)$, $\beta \sim \delta^2$, we get $\Omega(1) = h^2(\Pi_{V=0}, \Pi_{V=1}) \leq \delta^2 \cdot minIC(\pi)$
- Therefore, $minIC(\pi) = \Omega\left(\frac{1}{\delta^2}\right)$.

Putting it together

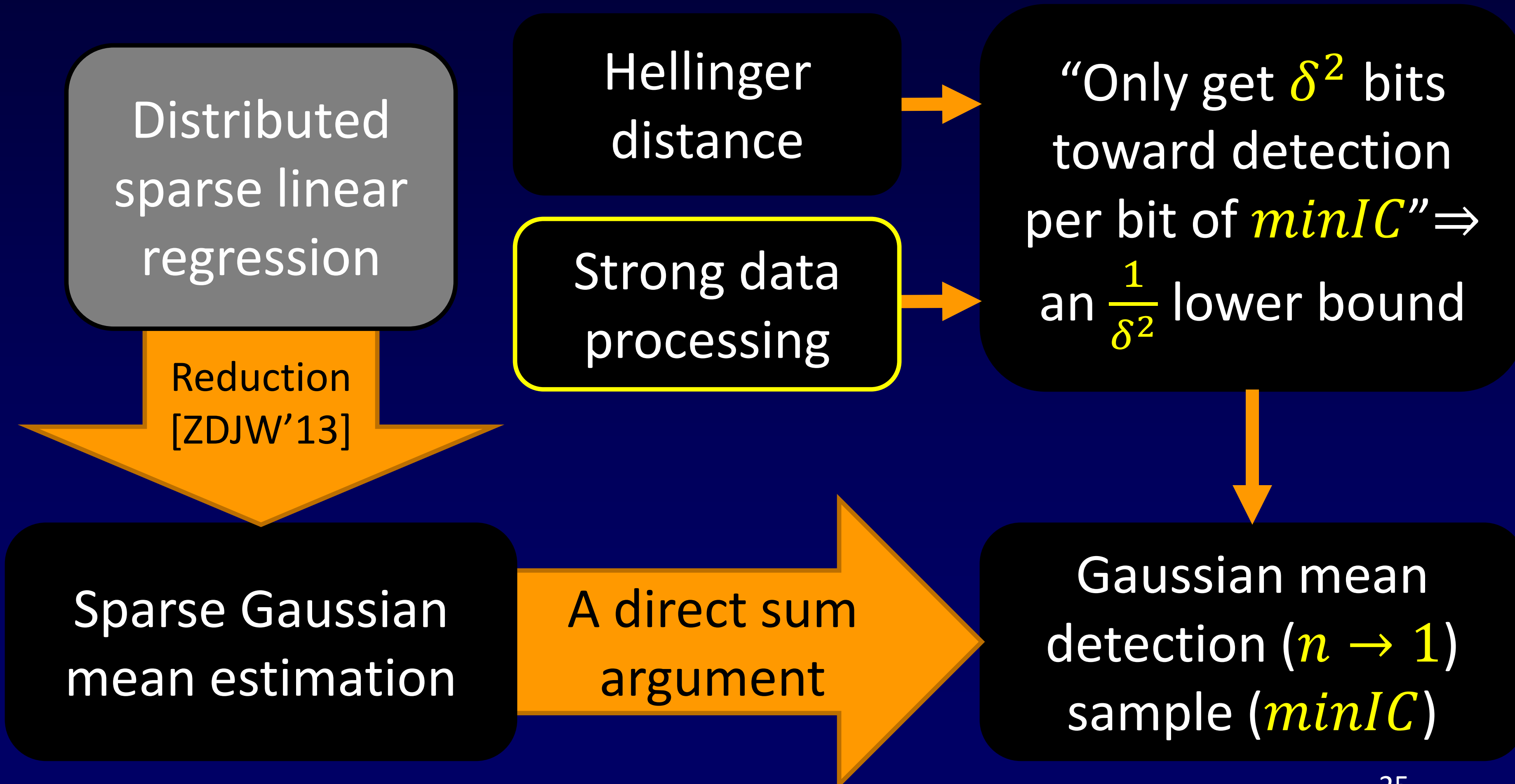
Essential!

Theorem: Suppose $\Omega(1) \cdot \mu_0 \leq \mu_1 \leq O(1) \cdot \mu_0$, and let $\beta(\mu_0, \mu_1)$ be the SDPI constant. Then

$$h^2(\Pi_{V=0}, \Pi_{V=1}) \leq O\big(\beta(\mu_0, \mu_1)\big) \cdot minIC(\pi)$$

- With $\mu_0 = \mathcal{N}(0,1)$, $\mu_1 = \mathcal{N}(\delta, 1)$
- $\Omega(1) \cdot \mu_0 \leq \mu_1 \leq O(1) \cdot \mu_0$ fails!!
- Need an additional truncation step. Fortunately, the failure happens far in the tails.

Summary


Distributed sparse linear regression
Reduction [ZDJW'13]
Sparse Gaussian mean estimation
A direct sum argument
Hellinger distance
Strong data processing
"Only get δ^2 bits toward detection per bit of $minIC$"⇒ an $\frac{1}{\delta^2}$ lower bound
Gaussian mean detection ($n \to 1$) sample ($minIC$)

Distributed sparse linear regression

- Each machine gets n data of the form (A^j, y^j), where $y^j = \langle A^j, \theta \rangle + w^j$, $w^j \sim \mathcal{N}(0, \sigma^2)$
- Promised that θ is k-sparse: $\|\theta\|_0 \leq k$.
- Ambient dimension d.
- Loss $R = \mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right]$.
- How much communication to achieve statistically optimal loss?

Distributed sparse linear regression

- Promised that θ is k-sparse: $\|\theta\|_0 \leq k$.
- Ambient dimension d. Loss $R = \mathbb{E}\left[\left\|\hat{\theta} - \theta\right\|^2\right]$.
- How much communication to achieve statistically optimal loss?
- We get: $C = \Omega(m \cdot \min(n, d))$ (small k doesn't help).
- [Lee-Sun-Liu-Taylor'15]: under some conditions $C = O(m \cdot d)$ suffice.

A new upper bound (time permitting)

- For the one-dimensional distributed Gaussian estimation (generalizes to d dimensions trivially).
- For optimal statistical performance, $\Omega(m)$ is the lower bound.
- We give a simple simultaneous-message upper bound of $O(m)$.
- Previously: multi-round $O(m)$ [GMN’14] or simultaneous $O(m \log n)$ [folklore].

A new upper bound (time permitting)

(Stylized) main idea:

- Each machine wants to send the empirical average $y_i \in [0,1]$ of its input.
- Then the average $\frac{1}{m}\sum_{i=1}^{m} y_i = \hat{y}$ is computed.
- Instead of y_i each machine sends b_i sampled from Bernoulli distribution B_{y_i}.
- Form the estimate $\hat{\hat{y}} = \frac{1}{m}\sum_{i=1}^{m} b_i$.
- "Good enough" if $\mathrm{var}(y_i) \sim 1$.

Open problems

- Closing the gap for the sparse linear regression problem.
- Other statistical questions in the distributed framework. More general theorems?
- Can Strong Data Processing be applied to the two-party Gap Hamming Distance problem?

- http://csnexus.info/

Organizers

- Mark Braverman (Princeton University)
- Bobak Nazer (Boston University)
- Anup Rao (University of Washington)
- Aslan Tchamkerten, General Chair (Telecom Paristech)

- http://csnexus.info/

Primary themes

- Distributed Computation and Communication
- Fundamental Inequalities and Lower Bounds
- Inference Problems
- Secrecy and Privacy

Institut Henri Poincaré

http://csnexus.info/

Thank You!