Information-theoretic Lower Bounds for Distributed Statistical Estimation with Communication Constraints

Yuchen Zhang    John Duchi
Michael I. Jordan    Martin J. Wainwright

University of California, Berkeley

NIPS 2013
A Modern Data Center

- Holds 10,000+ servers.
- Data storage and data processing highly distributed.
- Communication cost $\gg$ computation cost.
A Fundamental Trade-off

When learning from distributed data,

**Target 1:** maximize statistical accuracy.

**Target 2:** minimize communication cost.
A Fundamental Trade-off

When learning from distributed data,

**Target 1:** maximize statistical accuracy.

**Target 2:** minimize communication cost.

**This talk:** study the fundamental trade-off between these two targets.
Main Result

Communication-Accuracy trade-off:

![Graph showing the trade-off between communication and error in distributed statistical estimation.]
Statistical Estimation

**Given:** i.i.d. data drawn from unknown distribution $P$

**Goal:** estimate a parameter $\theta(P)$. 
Statistical Estimation

**Given:** i.i.d. data drawn from unknown distribution $P$

**Goal:** estimate a parameter $\theta(P)$.

**Example:**

- Gaussian location model.
- Linear Regression.
- Probit Regression.
Distributed Statistical Estimation

- Data is stored on $m$ separate machines.
- Each machine generates a message based on its local data.
- Output a message-based estimator.

![Diagram of distributed statistical estimation]

Output Estimator:

$$\hat{\theta}(Y_1, Y_2, \ldots, Y_m)$$
Distributed Statistical Estimation

- Data is stored on $m$ separate machines.
- Each machine generates a message based on its local data.
- Output a message-based estimator.

![Diagram of distributed statistical estimation](image)

- Statistical accuracy: $\mathbb{E}[\|\hat{\theta} - \theta\|^2_2]$
- Communication cost: $\sum_{i=1}^{m} \text{Length}(Y_i)$
Example: Gaussian Location Model

$m$ machines, each machine gets $X_i \sim \mathcal{N}(\theta, 1)$. Want to estimate $\theta$. 
Example: Gaussian Location Model

$m$ machines, each machine gets $X_i \sim \mathcal{N}(\theta, 1)$. Want to estimate $\theta$.

\[
\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} Y_i
\]
Example: Gaussian Location Model

$m$ machines, each machine gets $X_i \sim \mathcal{N}(\theta, 1)$. Want to estimate $\theta$.

\[
\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} Y_i
\]

Analysis:

- Estimation error: $\mathbb{E}[(\hat{\theta} - \theta)^2] \approx \frac{1}{m}$. (optimal rate)
- Communication cost $\approx m$. 
Example: Gaussian Location Model

$m$ machines, each machine gets $X_i \sim \mathcal{N}(\theta, 1)$. Want to estimate $\theta$.

\[
\begin{align*}
X_1 & \overset{\text{Quantize}}{\rightarrow} Y_1 \\
X_2 & \overset{\text{Quantize}}{\rightarrow} Y_2 \\
& \quad \vdots \\
X_m & \overset{\text{Quantize}}{\rightarrow} Y_m
\end{align*}
\]

\[
\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} Y_i
\]

Analysis:

- Estimation error: $\mathbb{E}[(\hat{\theta} - \theta)^2] \approx \frac{1}{m}$. (optimal rate)
- Communication cost $\approx m$.

Question: Is there a better estimator?
Minimum Possible Communication

Answer is: NO.
Minimum Possible Communication

Answer is: NO.

**Theorem**

If each of $m$ machines gets one i.i.d. sample from $N(\theta, 1)$, then any optimal estimator of $\theta$ must communicate $\tilde{\Omega}(m)$ bits.
Answer is: NO.

**Theorem**

If each of $m$ machines gets one i.i.d. sample from $N(\theta, 1)$, then any optimal estimator of $\theta$ must communicate $\tilde{\Omega}(m)$ bits.

<table>
<thead>
<tr>
<th>Centralized Estimation</th>
<th>Distributed Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Send $\Theta(\log m)$ bits</td>
<td>Send $\tilde{\Theta}(m)$ bits</td>
</tr>
</tbody>
</table>


**Gaussian Location Model** ($n \geq 1, \ d \geq 1$)

Given: $m$ machines, each machine gets $n$ i.i.d. samples from $\mathcal{N}(\theta, \sigma^2 I_d \times d)$.

Goal: find the Gaussian mean $\theta \in \mathbb{R}^d$. 

**Gaussian Location Model** \((n \geq 1, d \geq 1)\)

**Given:** \(m\) machines, each machine gets \(n\) i.i.d. samples from \(\mathcal{N}(\theta, \sigma^2 I_{d \times d})\).

**Goal:** find the Gaussian mean \(\theta \in \mathbb{R}^d\).

**Theorem**

If an estimator is allowed to communicate \(B\) bits, then

\[
\max_{\theta \in [-1,1]^d} \mathbb{E}[(\hat{\theta} - \theta)^2] \geq C \cdot \frac{d}{mn} \cdot \max \left\{ 1, \frac{dm}{B \log m} \right\}
\]
Gaussian Location Model \((n \geq 1, \ d \geq 1)\)

**Given:** \(m\) machines, each machine gets \(n\) i.i.d. samples from \(\mathcal{N}(\theta, \sigma^2 I_{d \times d})\).

**Goal:** find the Gaussian mean \(\theta \in \mathbb{R}^d\).

**Theorem**
If an estimator is allowed to communicate \(B\) bits, then

\[
\max_{\theta \in [-1,1]^d} \mathbb{E}[(\hat{\theta} - \theta)^2] \geq C \cdot \frac{d}{mn} \cdot \max \left\{ 1, \frac{dm}{B \log m} \right\}
\]

**Remark:**
- Optimal convergence rate is \(\mathcal{O}(\frac{d}{mn})\).
Gaussian Location Model \( (n \geq 1, \ d \geq 1) \)

**Given:** \( m \) machines, each machine gets \( n \) i.i.d. samples from \( \mathcal{N}(\theta, \sigma^2 I_{d \times d}) \).

**Goal:** find the Gaussian mean \( \theta \in \mathbb{R}^d \).

**Theorem**

If an estimator is allowed to communicate \( B \) bits, then

\[
\max_{\theta \in [-1,1]^d} \mathbb{E}[(\hat{\theta} - \theta)^2] \geq C \cdot \frac{d}{mn} \cdot \max \left\{ 1, \frac{dm}{B \log m} \right\}
\]

**Remark:**

- Optimal convergence rate is \( \mathcal{O}(\frac{d}{mn}) \).
- Any optimal estimator must communicate \( B = \Omega(\frac{dm}{\log m}) \) bits.
Lower Bound Curve

Communication Error

Distributed Statistical Estimation

Centralized Statistical Estimation

optimal rate

d log(m)

dm

Communication

Yuchen Zhang (UC Berkeley)
Achievability of Lower Bound

\[
\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} Y_i
\]
Achievability of Lower Bound

\[
\frac{\sum_{j=1}^{n} X_{1,j}}{n} \xrightarrow{\text{Quantize}} Y_1 \\
\frac{\sum_{j=1}^{n} X_{2,j}}{n} \xrightarrow{\text{Quantize}} Y_2 \\
\frac{\sum_{j=1}^{n} X_{m,j}}{n} \xrightarrow{\text{Quantize}} Y_m
\]

\[
\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} Y_i
\]

Analysis:

- Estimation error: \( \mathbb{E}[\| \hat{\theta} - \theta \|_2^2] = O\left(\frac{d}{mn}\right) \). (optimal rate)
- Communication cost: \( O(dm \log(mn)) \).
Achievability of Lower Bound

\[
\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} Y_i
\]

Analysis:

- Estimation error: \( \mathbb{E}[\|\hat{\theta} - \theta\|^2] = \mathcal{O}\left(\frac{d}{mn}\right) \). (optimal rate)
- Communication cost: \( \mathcal{O}(dm \log(mn)) \).

Conclusion: \( \tilde{\Theta}(dm) \) bits of communication are necessary and sufficient.
Consequence for Regression Problems

### Linear Regression

**Given:** $m$ machines, each machine gets $n$ i.i.d. inputs $(x_i, z_i)$ satisfying

$$x_i \in \mathbb{R}^d \quad \text{and} \quad z_i = \theta^T x_i + w_i$$

where $w_i \sim \mathcal{N}(0, \sigma^2)$.

**Goal:** find the regression coefficient $\theta \in \mathbb{R}^d$.

### Probit Regression

**Given:** $m$ machines, each machine gets $n$ i.i.d. inputs $(x_i, y_i)$ satisfying

$$x_i \in \mathbb{R}^d \quad \text{and} \quad z_i = \begin{cases} 
1 & \text{with probability } \Phi(\theta^T x_i) \\
0 & \text{with probability } 1 - \Phi(\theta^T x_i)
\end{cases}$$

where $\Phi$ is the CDF of standard normal distribution.

**Goal:** find the regression coefficient $\theta \in \mathbb{R}^d$. 
Consequence for Regression Problems

<table>
<thead>
<tr>
<th><strong>Lower Bound</strong></th>
</tr>
</thead>
</table>

For linear regression and probit regression, any optimal estimator of $\theta$ must communicates $\Omega(\frac{dm}{\log m})$ bits.
Consequence for Regression Problems

**Lower Bound**

For linear regression and probit regression, any optimal estimator of $\theta$ must communicates $\Omega(dm/\log m)$ bits.

**Upper Bound (Z, Duchi, Wainwright, NIPS’12)**

- Local Estimator $\hat{\theta}_1$ → Quantize → $Y_1$
- Local Estimator $\hat{\theta}_2$ → Quantize → $Y_2$
- Local Estimator $\hat{\theta}_m$ → Quantize → $Y_m$

Estimation error: $\mathbb{E}[\|\hat{\theta} - \theta\|^2] = \mathcal{O}(\frac{d}{mn})$. (optimal rate)

Communication cost: $\mathcal{O}(dm \log(mn))$. 
Multiple Rounds of Communication

- In each round, messages are generated by local data and old messages of previous rounds.
- Output a message-based estimator.

![Diagram]

Output Estimator: \( \hat{\theta}(\text{messages}) \)

- Unknown Distribution \( P \)
- Machine 1
- Machine 2
- Machine \( m \)
- Fusion Center

Send Message

Free Broadcast
Multiple Rounds of Communication

- In each round, messages are generated by local data and old messages of previous rounds.
- Output a message-based estimator.

\[ \hat{\theta}(\text{messages}) \]

\[ \mathbb{E}[\|\hat{\theta} - \theta\|^2_2] \]

\[ \sum \text{Length(message)} \]
Multiple Rounds of Communication: Lower Bound

**Theorem**

For \{Gaussian location model, linear regression, probit regression\} of dimension \(d = 1\), any optimal estimator of \(\theta\) must communicates \(\tilde{\Omega}(m)\) bits.

**Remark:**

- Interactivity doesn't help (communication cost linear in \(m\)).
- Open: generalization to \(d > 1\)?
Proof Ideas

1. Fix a communication budget $B \geq \text{Length}(messages)$. 
Proof Ideas

1. Fix a communication budget $B \geq \text{Length}(messages)$.
2. Data processing inequality:

$$I(\text{parameter, messages}) \leq I(\text{parameter, data}) \cdot I(\text{data, messages}) \leq B$$

message independent

parameter $\rightarrow$ data $\rightarrow$ messages
Proof Ideas

1. Fix a communication budget $B \geq \text{Length}(\text{messages})$.

2. Data processing inequality:

$$I(\text{parameter, messages}) \leq I(\text{parameter, data}) \cdot I(\text{data, messages})$$

message independent

parameter $\rightarrow$ data $\rightarrow$ messages

3. Lower bound $\mathbb{E}[\|\hat{\theta} - \theta\|^2]$ by the bound for $I(\text{parameter, messages})$. 
Proof Ideas

1. Fix a communication budget $B \geq \text{Length}(\text{messages})$.

2. Data processing inequality:

   $$I(\text{parameter, messages}) \leq I(\text{parameter, data}) \cdot I(\text{data, messages}) \leq B$$

   message independent

   parameter $\rightarrow$ data $\rightarrow$ messages

3. Lower bound $\mathbb{E}[\|\hat{\theta} - \theta\|_2^2]$ by the bound for $I(\text{parameter, messages})$.

   For $d$-dimension problem, a stronger inequality:

   $$I(\text{parameter, messages}) \leq \frac{I(\text{parameter, data})}{d} \cdot I(\text{data, messages})$$
Conclusion

Characterize trade-off between communication and accuracy:

- Single-round communication: Gaussian location model, linear regression, probit regression.
- Interactive communication: same problem set, $d = 1$. 
Conclusion

Characterize trade-off between communication and accuracy:

- Single-round communication: Gaussian location model, linear regression, probit regression.
- Interactive communication: same problem set, $d = 1$.

Future Works:

- Generalize the result to other statistical estimation problems.
- Tight lower bound for interactive communication in arbitrary dimension.