Statistical Inference in Distributed or Constrained Settings: Techniques and Recipes

Conference on Learning Theory 2021
What’s on the menu?

I. Appetizers  
   Jayadev

II. MC 1  
    Jayadev

III. MC 2  
     Himanshu

IV. DIY Desserts  
    Clément

Chefs: Jayadev Acharya, Clément Canonne, Himanshu Tyagi

COLT 2021
Appetizers

- Statistical Inference
- Distributed / constrained settings
- Problems and examples
- Related work and pointers
Main Course – I: Discrete distributions

• A puzzle to solve **all** problems under communication constraints

• Lower bounds for interactive estimation for arbitrary channels
  • Tight bounds under communication, privacy as application
Main Course – II: General distributions

Unified method to prove “interactive” lower bounds

• Discrete, high-dimensional, nonparametric, etc

• Communication, privacy, etc

• General plug-n-play methods
DIY desserts: Recitation

• How to apply the lower bounds
• Several exercises
Statistical Inference

\( \mathcal{P} \): family of distributions over \( \mathcal{X} \)

Given \( X^n := (X_1, \ldots, X_n) \): i.i.d. samples from an unknown \( p \)

Solve some inference task about \( p \)

**Sample complexity:** smallest \( n \) to solve the task

This is inference in central setting
Information Constraints
Distributed or Constrained Settings

No direct access to $X_i$'s
Statistical Inference under constraints

Local constraints
Statistical Inference

The messages are what we observe with constraints
Modeling the constraints

$n$ users, user $t$ observes $X_t$ and sends message $Y_t$

\[ W_t(y|x) := \Pr(Y_t = y | X_t = x) \]

$W_t \in \mathcal{W}$: a set of allowed (randomized) channels $\Leftrightarrow$ the constraints

The algorithm/protocol dictates how user $t$ chooses $W_t$ from $\mathcal{W}$
Modeling the local information constraints

When \( X_t \sim p \)

\[
p^{W_t}(Y_t = y) := \sum_x p(x)W_t(y|x) = \mathbb{E}[W_t(y|X)]
\]
Example 1: Communication constraints

\[ \mathcal{W}_\ell := \{ W : X \to \{0,1\}^\ell \} \]

Each \( X_t \) is mapped to \( \ell \) bits.

Bandwidth constraints
Example 2: Local Differential Privacy (LDP)

\[ W: \mathcal{X} \to \{0,1\}^* \text{ is } \varrho\text{-LDP if } \forall x, x' \in \mathcal{X}, \forall y, \]

\[
\frac{W(y|x)}{W(y|x')} \leq e^\varrho \approx 1 + \varrho
\]

\[ \mathcal{W}_\varrho = \{ \text{all } \varrho \text{-LDP channels} \} \]

Privacy guarantees even “against” the server

[Warner65, EPR03, KLNRS11]
The Protocols
Given $Y^n := Y_1, ..., Y_n$, solve the inference task
Distributed statistical inference

Once we decide $W^n := W_1, ..., W_n$,

$$p^{W^n}(Y^n) = \prod_{t} p^{W_t}(Y_t)$$

How to choose $W_1, W_2, ..., W_n \in \mathcal{W}$ to minimize $n$?
The protocols

Simultaneous Message Passing (SMP)/Non-interactive schemes

\( W_t \) s are chosen simultaneously

private-coin SMP (no shared randomness)

\( W_t \) s are chosen independently

\( Y_1, Y_2, \ldots, Y_n \) are independent

e.g., \( W_1, \ldots, W_n \) are fixed
Noninteractive ("simultaneous message-passing"),
no common randomness
The protocols

Simultaneous Message Passing (SMP)/Non-interactive schemes

$W_t$'s are chosen simultaneously

public-coin SMP (shared randomness)

$U$: common random string available to all users and referee

$W_t$ is a function of $U$

$Y_1, Y_2, ..., Y_n$ are independent given $U$
Noninteractive ("simultaneous message-passing"), but common random seed
The protocols

Interactive schemes

$W_t$s can depend on previous messages

sequentially interactive protocols

$U$: common random string available to all users and referee

for $t = 1, \ldots, n$

$W_t$ is a function of $(U, Y^{t-1})$
Sequentially Interactive protocols

Interactive ("one-pass, sequential"), and common random seed
Types of protocols

Blackboard protocols

Fully interactive (“many passes”), and common random seed
Types of protocols

Each of these models is at least as powerful as the previous

private-coin \leq public-coin \leq sequentially interactive \leq blackboard

Each has its pros and cons (both in theory and practice) and may require different techniques to analyze.
Questions about setting?
The Problems
Parameter/density estimation

Goodness-of-fit / Hypothesis testing

**Sample complexity:** smallest $n$ to solve the task
Example 1: Discrete distributions

\[ \mathcal{P} = \Delta_d : \text{distbs on } [d] := \{1 \ldots d\} \]

**Goal:** output \( \hat{p} \) such that

\[ \mathbb{E}[\text{TV}(\hat{p}, p)] \leq \varepsilon \]

Sample complexity = \( \Theta \left( \frac{d}{\varepsilon^2} \right) \)

(without constraints)

**q:** a reference distribution

**Goal:** Test

\[ p = q \text{ vs } \text{TV}(p, q) > \varepsilon \]

Sample complexity = \( \Theta \left( \frac{\sqrt{d}}{\varepsilon^2} \right) \)

(without constraints) [Paninski08]

\[ \text{TV}(p, q) := \sup_{S \subseteq [k]} (p(S) - q(S)) = \frac{1}{2} \ell_1(p, q) \]
Example 2: High dimensional distributions

\[ \mathcal{P} = \{ N(\mu, I_d) : \mu \in \mathbb{R}^d \} \]

**Goal:** output \( \hat{\mu} \) such that

\[ \mathbb{E}[|\hat{\mu} - \mu|^2] \leq \varepsilon^2 \]

Sample complexity = \( \Theta \left( \frac{d}{\varepsilon^2} \right) \)

(without constraints)

**Goal:** Test

\[ \mu = 0 \text{ vs } |\mu|_2 > \varepsilon \]

Sample complexity = \( \Theta \left( \frac{\sqrt{d}}{\varepsilon^2} \right) \)

(without constraints)

*detecting signal vs noise

Other families: product Bernoulli
Research goals

Establish sample complexity bounds for ...

- Different $\mathcal{W}$s
- Estimation/Testing/other properties
- Private-coin SMP/public-coin SMP/interactive
- Discrete/high-dimensional/non-parametric

Mix-n-match?

Already a bit too much ... each interesting in its own right ... !
For example ... discrete distribution testing

\( W_q, \) [AminJosephMao ’20, BerrettButucea’20, AcharyaCanonneLiuSunTyagi’20]:

Private-coin SMP \( \ll \) public-coin SMP \( \approx \) SMP/interactive

\( W_\ell, \) [AcharyaCanonneLiuSunTyagi’20]:

Private-coin SMP \( \ll \) public-coin SMP \( \approx \) SMP/interactive

General \( W, \) [AcharyaCanonneLiuSunTyagi’20]:

Private-coin SMP \( \ll \) public-coin SMP \( \ll \) SMP/interactive

Similarly for Gaussian mean testing ... [AcharyaCanonneTyagi’20, SzaboVuursteenVanZanten’20]
Parameter/density estimation

Goodness-of-fit / Hypothesis testing

Part 3 of tutorial (link)

Learn about Ingster’s method from HT!
Establishing tight results for SMP protocols generally easier ...

\( Y_1, \ldots, Y_n \) independent (given \( U \))

See general discussion in

Methods to establish interactive lower bounds

1. **Cramer-Rao/van Trees inequality** [BarnesHanOzgur19, BarnesChenOzgur20, SarbuZaidi21]
   - Unified results for $\Delta_d, \mathcal{B}_d, \mathcal{G}_d$
   - Results hold for $\ell_2$ loss

2. **Strong Data Processing + Assouad’s method**
   [BravermanGardMaNguyenWoodruff16, DuchiRogers19]
   - Lower bounds for $\mathcal{B}_d, \mathcal{G}_d$ under $\ell_2$ loss
   - Naturally extends to other $\ell_p$ loss functions

3. **Chi-squared contractions + Assouad’s method**
   [AcharysCanonneLiuSunTyagi20, AcharyaCanonneSunTyagi20]
   - Unified bounds for $\Delta_d, \mathcal{B}_d, \mathcal{G}_d$
   - Works under $\ell_p$ for $p \geq 1$
   - For arbitrary channels
Pointers

Part 2 of tutorial (link)

Cramer-Rao/van Trees inequality

Strong Data Processing + Assouad’s method
Next two parts ... 

MC1:
- Discrete distributions
  - Simulate and infer for upper bounds
  - Lower bounds

MC2:
- Unified approach for general distributions and channel families
MC 1: Discrete Distributions
Discrete distribution estimation

\( \mathcal{P} = \Delta_d : \text{distbs on } [d] := \{1 \ldots d\} \)

**Goal:** output \( \hat{\mathbf{p}} \) such that

\[ \mathbb{E}[\text{TV}(\hat{\mathbf{p}}, \mathbf{p})] \leq \varepsilon \]

Sample complexity = \( \Theta \left( \frac{d}{\varepsilon^2} \right) \) (without constraints)
Empirical distribution works - DIY

\[ X_1, \ldots, X_n \sim p, \quad N_x := \text{# times } x \text{ appears} \]

Empirical distribution: \( \hat{p}(x) = N_x/n \)

\[ N_x \sim \text{Bin}(n, p(x)) \]

\[
\mathbb{E} \left[ (\hat{p}(x) - p(x))^2 \right] = \frac{p(x)(1 - p(x))}{n} \Rightarrow \mathbb{E}[\ell_2^2(\hat{p}, p)] \leq \frac{1}{n}
\]

\[
\mathbb{E}[\ell_1(\hat{p}, p)]^2 \leq \mathbb{E}[\ell_1(\hat{p}, p)^2] \leq d \cdot \mathbb{E}[\ell_2^2(\hat{p}, p)] \leq \frac{d}{n} \quad \text{(Jensen)}
\]

\[
\mathbb{E}[\ell_1(\hat{p}, p)]^2 \leq d \cdot \mathbb{E}[\ell_2^2(\hat{p}, p)] \leq \frac{d}{n} \quad \text{(Cauchy Schwarz)}
\]
Under communication constraints

\[ \begin{align*}
X_1 \rightarrow W_1 \rightarrow Y_1 \\
X_2 \rightarrow W_2 \rightarrow Y_2 \\
\vdots \\
X_n \rightarrow W_n \rightarrow Y_n \\
\end{align*} \]

\[ p \]

\[ p \in \mathcal{W}_\ell \]

\[ \ell \in \{0,1\}^\ell \]
A simulation puzzle ...

**Goal:** To simulate a sample from messages

\[ X_1, X_2, \ldots, X_n \]
\[ W_1, W_2, \ldots, W_n \]
\[ Y_1, Y_2, \ldots, Y_n \]

\[ \mathcal{W}_\ell \]
\[ \{0,1\}^\ell \]

\[ X \sim \mathbf{p} \]
One simulation to solve them all ...

**Theorem.** Suppose simulation is possible with $f(d, \ell)$ samples.

Let $T$ be some task with sample complexity $T(d, \varepsilon)$.

Then $T$ can be solved with $f(d, \ell) \cdot T(d, \varepsilon)$ samples under $\mathcal{W}_\ell$.

What is $f(d, \log_2 d)$?
One simulation to solve them all ...

**Theorem.** There is a private-coin SMP protocol with

\[ f(d, \ell) \approx \max \left\{ \frac{d}{2^\ell}, 1 \right\}. \]

No protocol (even interactive) can do better!

Estimation with \( \Theta \left( \frac{d}{\varepsilon^2} \cdot \frac{d}{2^\ell} \right) \) and testing with \( \Theta \left( \frac{\sqrt{d}}{\varepsilon^2} \cdot \frac{d}{2^\ell} \right) \)
Algorithm for one-bit

Take $2d$ players and pair them into $d$ groups:

- First pair tell if their input is symbol 1
- Second tell if their input is symbol 2
- And so on ...
Algorithm for one-bit

\[ Y_{2i-1} = I\{X_{2i-1} = i\} \]
\[ Y_{2i} = I\{X_{2i} = i\} \]
Algorithm for one-bit

• Output $i \in [d]$ if:
  • Player $2i - 1$ is the only odd player sending 1
  • Player $2i$ sends 0
• If no such $i$, output $\bot$

Conditioned on not outputting $\bot$, output $\sim p$
Algorithm for one-bit

Player $2i - 1$ is the only odd player sending 1

\[
\Pr(Y_{2i-1} = 1, Y_{2i' - 1} = 0 \text{ for } i' \neq i) = p(i) \prod_{i' \neq i} (1 - p(i'))
\]

Player $2i$ sends 0

\[
\Pr(Y_{2i} = 0) = (1 - p(i))
\]

\[
\Pr(\text{output } i | \text{ not } \bot) = p(i) \cdot \prod_{i' \in [d]} (1 - p(i')) \propto p(i)
\]
### Corollary

<table>
<thead>
<tr>
<th>Inference Task</th>
<th>Centralized</th>
<th>One-bit private-SMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimation</td>
<td>$\Theta \left( \frac{d}{\varepsilon^2} \right)$</td>
<td>$\Theta \left( \frac{d^2}{\varepsilon^2} \right)$</td>
</tr>
<tr>
<td>Testing</td>
<td>$\Theta \left( \frac{\sqrt{d}}{\varepsilon^2} \right)$</td>
<td>$\Theta \left( \frac{d^{3/2}}{\varepsilon^2} \right)$</td>
</tr>
</tbody>
</table>
Corollary

<table>
<thead>
<tr>
<th>Inference Task</th>
<th>Centralized</th>
<th>One-bit private-SMP</th>
<th>One-bit public-SMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimation</td>
<td>$\Theta\left(\frac{d}{\varepsilon^2}\right)$</td>
<td>$\Theta\left(\frac{d^2}{\varepsilon^2}\right)$</td>
<td>$\Theta\left(\frac{d^2}{\varepsilon^2}\right)$</td>
</tr>
<tr>
<td>Testing:</td>
<td>$\Theta\left(\frac{\sqrt{d}}{\varepsilon^2}\right)$</td>
<td>$\Theta\left(\frac{d^{3/2}}{\varepsilon^2}\right)$</td>
<td>$\Theta\left(\frac{d}{\varepsilon^2}\right)$</td>
</tr>
</tbody>
</table>

Bounds are tight ... simulate and infer is optimal for private-coin SMP
Related work

Under SMP protocols these bounds are tight for communication constraints [HanMukherjeeOzgur19, AcharyaCanonneTyagi’19] and LDP [DuchiJordanWainwright14]

Sample complexity with interactivity and general channels?

Reminder of my time: prove lower bounds

Recipe:

• Design **hard instances** that has some structure
• Show that problem is hard within these
• Assouad’s method and reduction to testing
• Bound “information contraction” due to constraints
A hard instance
A hard instance

[Paninski’08] Let $\mathcal{Z} = \{-1,1\}^{d/2}$, and $\mathcal{P}_\mathcal{Z} = \{p_z : z \in \mathcal{Z}\}$, where

$$p_z(2i - 1) = \frac{1 + z_i \cdot 2\varepsilon}{d}, \quad p_z(2i) = \frac{1 - z_i \cdot 2\varepsilon}{d}, \quad i = 1, \ldots, d/2.$$
Learning lower bounds

\[ Z = (Z_1, \ldots, Z_{d/2}) \sim_{uar} \mathcal{Z}, \text{ i.e., each } Z_i \sim_{iid} \text{ Bern}(0.5) \]
Exercise: Let \( z \in \mathcal{Z} \) and \( \hat{p} \) satisfies \( d_{TV}(\hat{p}, p_z) < \frac{\varepsilon}{10} \).

Then,

\[
    z^* = \arg\min_{z'} d_{TV}(\hat{p}, p_{z'})
\]

satisfies

\[
    \text{Ham}(z, z^*) < \frac{d}{10}.
\]
From learning to testing
Assouad’s method

If we can estimate $p_Z \in_{uar} P_Z$, then we can estimate $Z$!

**Theorem.** Pick $Z \sim_{uar} Z$.

If

$$\mathbb{E}_Z \left[ \mathbb{E}_{p_Z} [d_{TV}(\hat{p}(Y^n, U), p_Z)] \right] < \frac{\varepsilon}{10}$$

then there exists an estimator $\hat{Z}(Y^n, U)$ such that

$$\sum_{1 \leq i \leq d/2} \Pr(\hat{Z}_i = Z_i) > 0.8 \times \frac{d}{2}.$$ 

- **Note:** We could write this bound as $\sum_i I(Z_i \land Y^n | U) = \Omega(d)$
Assouad’s method

Exercise. If

$$\sum_{1 \leq i \leq d/2} \Pr(\hat{Z}_i = Z_i) > 0.8 \times \frac{d}{2},$$

then there exists a subset $S \subseteq \{1, \ldots, d/2\}$ with $|S| > d/6$ s.t. if $i \in S$,

$$\Pr(\hat{Z}_i = Z_i) > 0.7.$$

Now we need a lower bound on $n$ for this to happen.
Information bound on one coordinate
Notation

Fix $i \in [d/2]$, when can we figure $Z_i$?

$p_{Z_i}^{Y^n}$: distribution of $Y^n$ when input distribution $p_z$
Information bound on one coordinate

average output distribution fixing $Z_i = \pm 1$:

When $Z_i = 1$:  
\[ p_{+i}^{y_n} := \frac{1}{2d/2-1} \sum_{z:z_i=+1} p_z^{y_n} \]

When $Z_i = -1$:  
\[ p_{-i}^{y_n} := \frac{1}{2d/2-1} \sum_{z:z_i=-1} p_z^{y_n} \]

If we can guess $Z_i$ from $Y^n$
\[ \iff \text{d}_{TV}(p_{+i}^{y_n}, p_{-i}^{y_n}) \text{ must be large} \]
\[ \Rightarrow \text{bound distance between } p_{+i}^{y_n} \text{ and } p_{-i}^{y_n} \]
Total variation and hypothesis testing

\( p_1, p_2 \) be any two distributions over \( \mathcal{Y} \)

\( j \in \{1, 2\} \) be picked at random

Given \( Y \sim p_j \), design a \( \hat{j}(Y) \) that is a guess for \( j \)

For any \( \hat{j}(Y) \):

\[
\Pr(\hat{j}(Y) = j) \leq \frac{1}{2} \left( 1 + d_{TV}(p_1, p_2) \right)
\]
Information bound on one coordinate

In our case, $p_1 = p_{+i}^n, p_2 = p_{-i}^n$, and

$$\Pr(\hat{Z}_i = Z_i) > 0.7 \Rightarrow d_{TV}(p_{+i}^n, p_{-i}^n) \geq 0.4$$

Since this holds for at least $d/6$ coordinates,

$$\sum_{i} d_{TV}(p_{+i}^n, p_{-i}^n)^2 \geq \frac{d}{6} \times 0.16.$$
Some ingredients

\[ D(p_1 || p_2) := \sum_y p_1(y) \log \frac{p_1(y)}{p_2(y)}, \chi^2(p_1, p_2) := \sum_y \left( \frac{p_1(y) - p_2(y)}{p_2(y)} \right)^2 \]

Pinsker’s inequality, convexity of logarithms:

\[ 2 \cdot d_{TV}(p_1, p_2)^2 \leq D(p_1 || p_2) \leq \chi^2(p_1, p_2) \]

Chain rule of KL divergence: If \( p_1 \) and \( p_2 \) are over \( Y_1 \times Y_2 \):

\[
D(p_1(Y_1, Y_2) || p_2(Y_1, Y_2)) \\
= D(p_1(Y_1) || p_2(Y_1)) + \mathbb{E}_{Y_1} [D(p_1(Y_2|Y_1) || p_2(Y_2|Y_1))] 
\]
KL $\leq$ chi-squared (DIY)

Since $\log(1 + x) \leq x$ (why?)

$$D(p||q) := \sum_x p(x) \log \left(1 + \frac{p(x) - q(x)}{q(x)}\right)$$

$$\leq \sum_x p(x) \frac{(p(x) - q(x))}{q(x)} = \chi^2(p, q)$$

Exercise: Prove the chain rule of KL.
Why go to KL?

By Pinsker’s inequality,

\[ 4 \cdot d_{TV}(p_{+i}^n, p_{-i}^n)^2 \leq \left( D(p_{+i}^n \| p_{-i}^n) + D(p_{-i}^n \| p_{+i}^n) \right) \]

Summing over \( i \),

\[
\sum_i \left( D(p_{+i}^n \| p_{-i}^n) + D(p_{-i}^n \| p_{+i}^n) \right) \\
\geq \sum_i 4 \cdot d_{TV}(p_{+i}^n, p_{-i}^n)^2 \geq 4 \cdot \frac{d}{6} \times 0.16 \geq \frac{d}{10}
\]

\( p_{+i}^n \) are mixture distributions!

Handling mixtures is painful, leads to issues to extend SMP lower bounds to interactive setting.
Exercise: KL divergence is convex.
For any distributions $\mathbf{p}_1, \mathbf{p}_2$ and $\mathbf{q}_1, \mathbf{q}_2$ and $\lambda \in [0,1],$

$$D(\lambda \mathbf{p}_1 + (1 - \lambda) \mathbf{q}_1 \parallel \lambda \mathbf{p}_2 + (1 - \lambda) \mathbf{q}_2)$$

$$\leq \lambda \cdot D(\mathbf{p}_1 \parallel \mathbf{p}_2) + (1 - \lambda) \cdot D(\mathbf{p}_1 \parallel \mathbf{p}_2)$$

Prove using concavity of logarithms
Convexity to handle mixtures

$z \in \{-1,1\}^{k/2}$, $z^\oplus_i$ obtained by flipping the $i$th coordinate of $z$

**Theorem.**

$$\frac{1}{2}(D(p_{+i}^n || p_{-i}^n) + D(p_{-i}^n || p_{+i}^n)) \leq \mathbb{E}_Z[D(p_Z^n || p_{Z^\oplus i}^n)]$$

**Proof.** Convexity of divergence to the definitions of $p_{+i}^n$ and $p_{-i}^n$ ■

Information about $Z_i$ bounded by average divergence in message distribution upon changing only $Z_i$ when all others are fixed!
Convexity to handle mixtures

Summing over $i$

\[
\frac{d}{20} \leq \mathbb{E}_Z \left[ \sum_i D(p^n_Z \| p^n_{Z \oplus i}) \right]
\]

- For given $Z$ the sum is divergences when changing one coordinate
Focus on one $z$

By expectation<max, and linearity of expectations,

$$\frac{d}{20} \leq \max_z \left[ \sum_i D(p_{z}^{y} \parallel p_{z \oplus i}^{y}) \right]$$

** the following is the original bound in terms of MI:

$$\sum_i I(Z_i \wedge Y^n) \leq \frac{1}{2} \cdot \max_z \left[ \sum_i D(p_{z}^{y} \parallel p_{z \oplus i}^{y}) \right]$$
Bounding $\sum_i D\left( p_{Z}^{Y_n} \| p_{Z \oplus i}^{Y_n} \right)$

By the chain rule of divergence

$$\sum_i D\left( p_{Z}^{Y_n} \| p_{Z \oplus i}^{Y_n} \right) = \sum_t E_{p_{Z}^{Y_t}} \left[ \sum_i D\left( p_{Z}^{Y_{t-1}|Y_t} \| p_{Z \oplus i}^{Y_{t-1}|Y_t} \right) \right].$$

- $p_{Z}^{Y_{t}|Y^{t-1}}$: Distribution of $Y_t$ with input $p_{Z}$ conditioned on $Y^{t-1}$
- Channel at player $t$ a function only of $Y^{t-1}$, denoted $W_{Y^{t-1}}$
Recall

For $z \in \{-1,1\}^{d/2}$,

$$p_z(2i - 1) = \frac{1 + z_i 2\varepsilon}{d}, \quad p_z(2i) = \frac{1 - z_i 2\varepsilon}{d}, \quad i = 1, \ldots, d/2.$$  

$p_z$ and $p_z \oplus i$ differ only on $2i - 1$ and $2i$.
Bounding $\sum_i D \left( \frac{p_{z|Y_{t-1}}^{Y_t} || p_{z+\Theta_i|Y_{t-1}}^{Y_t}}{\sum_i D (p_{z|Y_{t-1}}^{Y_t} || p_{z+\Theta_i|Y_{t-1}}^{Y_t})} \right)\n
p_z$ and $p_{z+\Theta_i}$ differ only on $2i - 1$ and $2i$ by $4\varepsilon z_i/d$

- Fix $Y_{t-1}$

$$p_{z|Y_{t-1}}^{Y_t}(y) = p_{z+\Theta_i|Y_{t-1}}^{Y_t}(y) + \frac{4\varepsilon z_i}{d} \left( W^{Y_{t-1}}(y|2i - 1) - W^{Y_{t-1}}(y|2i) \right)$$
Bounding $\sum_i D \left( p^Y_{Z|Y^{t-1}} \parallel p^Y_{Z\oplus i} \right)$

Since $KL \leq \chi^2$, plugging the expression above

$$\sum_i D \left( p^Y_{Z|Y^{t-1}} \parallel p^Y_{Z\oplus i} \right) \leq \sum_i \sum_y \frac{\left( p^Y_{t}(y) - p^Y_{t\oplus i}(y) \right)^2}{p^Y_{t\oplus i}(y)} \leq \frac{8\varepsilon^2}{d} \cdot \sum_i \sum_y \frac{(W(y|2i-1) - W(y|2i))^2}{\sum_x W(y|x)}$$

Recall

$$p^Y_{Z}(2i-1) = \frac{1 + Z_i \varepsilon}{d}, \quad p^Y_{Z}(2i) = \frac{1 - Z_i \varepsilon}{d}$$

$$|W(y|2i-1) - W(y|2i)| \text{ large} \iff \text{seeing } y \text{ tells about } Z_i$$
An average information contraction bound

**Theorem.** [ACLST20] Under any interactive protocol,

\[ \sum_i I(Z_i \land Y^n) \leq n \cdot \frac{8\varepsilon^2}{d} \cdot \sup_{w \in \mathcal{W}} \sum_i \sum_y \left( \frac{W(y|2i - 1) - W(y|2i)}{\sum_x W(y|x)} \right)^2 \]

**Theorem.** If there exists an estimator then

\[ \frac{d}{20} \leq n \cdot \frac{8\varepsilon^2}{d} \cdot \sup_{w \in \mathcal{W}} \sum_i \sum_y \left( \frac{W(y|2i - 1) - W(y|2i)}{\sum_x W(y|x)} \right)^2 \]
Applications
For any $W \in \mathcal{W}_\ell$

$$\sum_i \sum_y \frac{(W(y|2i - 1) - W(y|2i))^2}{\sum_x W(y|x)} \leq 2^\ell$$

For any $W \in \mathcal{W}_q, q \leq 1$

$$\sum_i \sum_y \frac{(W(y|2i - 1) - W(y|2i))^2}{\sum_x W(y|x)} = O(q^2)$$
Interactive lower bound for estimation

\[
\frac{d}{20} \leq n \cdot \frac{8\varepsilon^2}{d^2} \cdot 2^\ell
\]

\[
n = \Omega\left( \frac{d^2}{2^\ell \varepsilon^2} \right)
\]

\[
\frac{d}{20} \leq n \cdot \frac{8\varepsilon^2}{d} \cdot q^2
\]

\[
n = \Omega\left( \frac{d^2}{\varepsilon^2 q^2} \right)
\]
Plug-n-play bounds

$H(W)$ is a $\frac{d}{2} \times \frac{d}{2}$ PSD matrix:

\[
(H(W))_{ij} := \sum_{y \in Y} \frac{(W(y|2i - 1) - W(y|2i))(W(y|2j - 1) - W(y|2j))}{\sum_j W(y|j)}
\]

\[
\sum_i \sum_y \frac{(W(y|2i - 1) - W(y|2i))^2}{\sum_x W(y|x)} = \| H(W) \|_*
\]
## Plug-n-play bounds

\[ \| \mathcal{W} \| \overset{\text{def}}{=} \max_{W \in \mathcal{W}} \| H(W) \| \]

**Testing:**

<table>
<thead>
<tr>
<th>Classic SMP</th>
<th>Private-coin SMP</th>
<th>Public-coin SMP</th>
<th>Sequentially Interactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Omega \left( \frac{\sqrt{d}}{\varepsilon^2} \right) )</td>
<td>( \Omega \left( \frac{d^{3/2}}{\varepsilon^2 | \mathcal{W} |_*} \right) )</td>
<td>( \Omega \left( \frac{d}{\varepsilon^2 | \mathcal{W} |_F} \right) )</td>
<td>( \Omega \left( \frac{d}{\varepsilon^2 \sqrt{| \mathcal{W} |<em>{OP} | \mathcal{W} |</em>*}} \right) )</td>
</tr>
</tbody>
</table>

**Estimation**

<table>
<thead>
<tr>
<th>Classic</th>
<th>Sequentially Interactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Omega \left( \frac{d}{\varepsilon^2} \right) )</td>
<td>( \Omega \left( \frac{d^2}{\varepsilon^2 | \mathcal{W} |_*} \right) )</td>
</tr>
</tbody>
</table>
Next 45 minutes:

**Reinforcement Learning** by Himanshu Tyagi ...
# References

(Click to go)

<table>
<thead>
<tr>
<th>Year</th>
<th>Title</th>
<th>Authors</th>
<th>Conference/Source</th>
<th>Year</th>
</tr>
</thead>
</table>
References

2019 (5)


Fisher Information for Distributed Estimation under a Blackboard Communication Protocol. Leighton P. Barnes; Yanjun Han; and Ayfer Özgür. In ISIT, pages 2704–2708, 2019. IEEE


2018 (4)


2017 (1)


2016 (1)


2014 (2)


2013 (1)


2009 (1)

Some references and previous work
Some references and previous work

Too many for a single slide, or two. Starts, more or less, with Tsitsiklis’89, picks up again in the mid-2000’s with a slightly different focus: local privacy, various types of communication constraints, ML-related motivations...

For a detailed bibliography:

Now you all say ... Phew!