Statistical Inference in Distributed and Constrained Settings

Part 2: Unified Lower Bounds

What will we see?

- More general lower bounds that apply to high-dimensional models
- Bounds that allow interaction

The model: ②

- $\mathcal{P}_{\Theta} = \{ p_\theta, \ \theta \in \Theta \}$
  a parametric family of distributions on $\mathcal{X}$
- $X_1, \dots, X_n$
  i.i.d. samples from an unknown
  $p_\theta \in \mathcal{P}_{\Theta}$

- $Y_1, \ldots, Y_n$

Information constraints:

$$\begin{tikzcd} & (Y_1, Y_2, \ldots, Y_{t-1}) \arrow[d, dashed, "\pi"'] & \\ X_t \arrow[r] & \boxed{W_t} \arrow[r] & Y_t \end{tikzcd}$$

sequentially interactive

- $\hat{\Theta}(Y_1, \ldots, Y_n)$

Estimate for unknown Θ

- $$\max_{\Theta \in \Theta} \mathbb{E}_\Theta \left[\| \theta - \hat{\Theta} \|_p \right]$$

worst-case loss to capture the performance of your algorithm

estimator $\hat{\Theta}$

protocol π

How small can we we make $\max_{\theta} \mathbb{E}_{\theta}[\|\theta - \hat{\theta}\|_p]$ if we are allowed to use any algorithm in the world?

Heuristics from information theory:

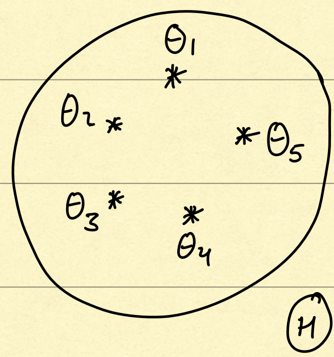
(1) Uncertainty about the unknown θ decreases as we observe $Y_1, \ldots, Y_n$

(2) This uncertainty can be quantified using measures of information

(3) Information is usually additive in n

(4) The presence of W_t's will reduce the information (data processing)

(1) Quantifying uncertainty



- Choose a finite set of M points in the parameter space

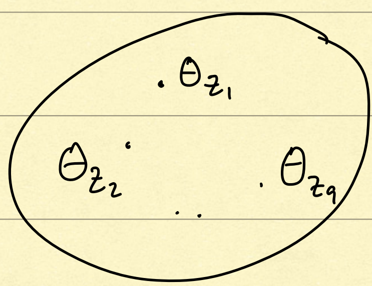
- Let one of them be chosen randomly and n samples be generated from it

- Uncertainty remaining upon observing Y^n
  - → Two points : $1 - d\left(P_{\theta_1}^{Y^n}, P_{\theta_2}^{Y^n}\right)$ ⇝ Le Cam
  - → Multiple points: $H(Z \mid Y^n)$ ⇝ Fano

6

→ Average uncertainty for each coordinate
⇝ Assouad's method

(Suited for high-dimensional + interactive)



$\Theta \subseteq \mathbb{R}^k$

Embedding: 7

$$|\theta_{z,i} - \theta_{z',i}| \approx r\, \mathbb{1}_{\{z_i \neq z_i'\}}$$

$$\|\theta_z - \theta_{z'}\|^2 \approx r^2 \sum_{i=1}^{k} \mathbb{1}_{\{z_i \neq z_i'\}}$$

8

$$\hat{Z} = \underset{Z \in \{-1,1\}^k}{\operatorname{argmin}} \|\hat{\theta} - \theta_Z\|$$

$$\Rightarrow \mathbb{E}_{\theta_Z}\left[\|\hat{\theta} - \theta_Z\|^2\right]$$

$$\geq \frac{1}{2} \mathbb{E}_{\theta_Z}\left[\|\theta_{\hat{Z}} - \theta_Z\|^2\right]$$

$$\approx r^2 \sum_{i=1}^{k} \mathbb{P}_{\theta_Z}\left[Z_i \neq \hat{Z}\right]$$

Recall:

For $Z \sim \text{unif}\{-1,1\}^k$, $(Z \to \theta_Z \to Y^n \to \hat{\theta} \to \hat{Z})$

9

$$\mathbb{E}_Z\left[\mathbb{E}_{\theta_Z}\left[\|\hat{\theta} - \theta_Z\|^2\right]\right]$$

$$\gtrsim r^2 \sum_{i=1}^{k} \mathbb{P}\left[Z_i \neq \hat{Z}_i\right]$$

$\mathbb{P}(Z_i \neq \hat{Z}_i) \rightsquigarrow$ probability of error for a binary hypothesis testing prob.

$$p_{+i}^{y^n} = p(y^n \mid Z_i = +1)$$
$$p_{-i}^{y^n} = p(y^n \mid Z_i = -1)$$
$$\mathbb{P}(Z_i \neq \hat{Z}_i) \gtrsim \frac{1}{2}\left(1 - d\left(p_{+i}^{y^n}, p_{-i}^{y^n}\right)\right)$$

$$\mathbb{E}_Z\left[\mathbb{E}_{\theta_Z}\left[\|\hat{\theta} - \theta_Z\|^2\right]\right] \tag{10}$$

$$\gtrsim r^2 \sum_{i=1}^{k}\left(1 - d\left(p_{+i}^{y^n}, p_{-i}^{y^n}\right)\right)$$

Thus,

loss $\gtrsim$ "uncertainty"

$$\simeq r^2 k \left(1 - \frac{1}{k}\sum_{i=1}^{k} d\left(p_{+i}^{y^n}, p_{-i}^{y^n}\right)\right)$$

(2) Uncertainty to information

$$\left(\frac{1}{k}\sum_{i=1}^{k} d\left(P_{+i}^{Y^n}, P_{-i}^{Y^n}\right)\right)^2$$

$$\leq \frac{1}{k}\sum_{i=1}^{k} d\left(P_{+i}^{Y^n}, P_{-i}^{Y^n}\right)^2$$

<u>Lemma</u>. P_+, P_- two distributions on $\mathcal{X}$

$U \sim \text{unif}\{-1, +1\}$

$P_{X|U=+1} = P_{+1}$

$P_{X|U=-1} = P_{-1}$

Then,

$$d(P_+, P_-)^2 \leq 2 I(U \wedge X)$$

<u>Proof</u>. $d(P_+, P_-)^2 \leq 2\left(d(P_+, Q)^2 + d(P_-, Q)^2\right)$

$Q = \frac{P_+ + P_-}{2}$

$$\leq D(P_+ \| Q) + D(P_- \| Q) = 2 I(U \wedge X) \quad \blacksquare$$

Lemma $\Rightarrow$

$$d\left(P_{+i}^{Y^n}, P_{-i}^{Y^n}\right)^2 \le 2 I(Z_i \wedge Y^n)$$

$=$

$$\le 2 I(Z_i \wedge Y^n, Z^{-i})$$

(independence of $Z_1, \ldots, Z_k$) $\le 2 I(Z_i \wedge Y^n \mid Z^{-i})$

$$\left(\frac{1}{k} \sum_{i=1}^{k} d\left(P_{+i}^{Y^n}, P_{-i}^{Y^n}\right)\right)^2$$

$$\le \frac{1}{k} \sum_{i=1}^{k} I(Z_i \wedge Y^n \mid Z^{-i}).$$

(13)

(1), (2) $\Rightarrow$

$$\mathbb{E}_Z\left[\mathbb{E}_{P_{\Theta_Z}}\left[\|\Theta_Z - \hat{\Theta}\|^2\right]\right] \quad \text{"loss"}$$

$$\gtrsim r^2 k \left(1 - \sqrt{\frac{2}{k}\sum_{i=1}^{k} I(Z_i \wedge Y^n \mid Z^{-i})}\right)$$

"information"

need to derive an upper bound for this information quantity

(3) Tensorization of information

(14)

$$I(Z_i \wedge Y^n \mid Z^{-i}) = \sum_{t=1}^{n} I(Z_i \wedge Y_t \mid Z^{-i}, Y^{t-1})$$

(4) How W_t s reduce information

Bounding $I\left(Z_i \wedge Y_t \mid Z^{-i} = z^{-i}, Y^{t-1} = y^{t-1}\right)$

Simplifying notation: (15)

$$W_t = W^{y^{t-1}}$$

$$z^{+i} = (z_1, \dots, z_{i-1}, 1, z_{i+1}, \dots, z_k)$$

$$z^{-i} = (z_1, \dots, z_{i-1}, -1, z_{i+1}, \dots, z_k)$$

$$P_{z^{+i}}^{W_t} = p\left(Y_t \mid Z_i = 1, z^{-i}, y^{t-1}\right)$$

$$X_t \sim P_{z+1} \rightarrow \boxed{W_t} \rightarrow Y_t$$

$$P_{z^{-i}}^{W_t} = p\left(Y_t \mid Z_i = -1, z^{-i}, y^{t-1}\right)$$

$$X_t \sim P_{z^{-i}} \rightarrow \boxed{W_t} \rightarrow Y_t$$

$$I\left(Z_i \wedge Y_t \mid Z^{-i} = z^{-i}, Y^{t-1} = y^{t-1}\right) \qquad (16)$$

$$= \frac{1}{2} D\left(P_{z^{+i}}^{W_t} \,\middle\|\, \frac{1}{2} P_{z^{+i}}^{W_t} + \frac{1}{2} P_{z^{-i}}^{W_t}\right)$$

$$+ \frac{1}{2} D\left(P_{z^{-i}}^{W_t} \,\middle\|\, \frac{1}{2} P_{z^{+i}}^{W_t} + \frac{1}{2} P_{z^{-i}}^{W_t}\right)$$

Conditioning on Z^{-i} helped — with this the distribution of X_t doesn't depend on Y^{t-1}:

$$X_t — (Z_i, Z^{-i}) — Y^{t-1}$$

$$\Rightarrow I(Z_i \wedge Y_t \mid Z^{i-1}, Y^{t-1})$$

$$= \mathbb{E}_{Z}\left[\mathbb{E}_{Y^{t-1}}\left[D\left(P_Z^{W_t} \,\middle\|\, \frac{1}{2}P_Z^{W_t} + \frac{1}{2}P_{Z\oplus i}^{W_t}\right)\right]\right]$$

- p, q two distributions on $\mathcal{X}$ (17)
- $W : \mathcal{X} \to \mathcal{Y}$

$$D\left(p^W \,\middle\|\, \frac{1}{2}p^W + \frac{1}{2}q^W\right)$$

$$\leq d_{\chi^2}\left(p^W \,\middle\|\, \frac{1}{2}p^W + \frac{1}{2}q^W\right)$$

$$= \int \frac{\left(\frac{1}{2}p^W(y) - \frac{1}{2}q^W(y)\right)^2}{\frac{1}{2}\left(p^W(y) + q^W(y)\right)} \, d\mu$$

$$\leq \frac{1}{2} \int \frac{\left(p^W(y) - q^W(y)\right)^2}{p^W(y)} \, d\mu$$

$$p^W(y) - q^W(y) = \mathbb{E}_p \left[W(y|X) \underbrace{\left(\frac{q(X)}{p(X)} - 1 \right)}_{=: \phi(X)} \right]$$

$$= \frac{1}{2} \int \frac{\mathbb{E}_p \left[W(y|X) \ \phi(X) \right]^2}{\mathbb{E}_p \left[W(y|X) \right]} \, d\mu$$

(18)

$$D\left(p^W \middle\| \frac{1}{2} p^W + \frac{1}{2} q^W\right) \leq$$

$$\frac{1}{2} \int \frac{\mathbb{E}_p \left[\phi(X) W(y|X) \right]^2}{\mathbb{E}_p \left[W(y|X) \right]} \, d\mu$$

where $\phi(x) = \dfrac{q(x)}{p(x)} - 1$.

b

$$\Rightarrow \sum_{i=1}^{k} \sum_{t=1}^{n} I(Z_i \wedge Y_t \mid Z^{-i}, Y^{t-1}) \qquad (19)$$

$$= \sum_{i=1}^{k} \sum_{t=1}^{n} \mathbb{E}_{Z, Y^{t-1}} \left[\underbrace{D\left(P_Z^{W_t} \,\middle\|\, \frac{1}{2} P_Z^{W_t} + \frac{1}{2} P_{Z \oplus i}^{W_t}\right)} \right]$$

$$\leq \frac{1}{2} \int \frac{\mathbb{E}_{P_Z}[\phi_{Z,i}(X) W_t(y|X)]^2}{\mathbb{E}_{P_Z}[W_t(y|X)]} d\mu$$

where $\phi_{z,i}(x) = \frac{dP_{z \oplus i}}{dP_z}(x) - 1$.

$$\leq \frac{1}{2} \cdot n \cdot \max_{z} \max_{W \in \mathcal{W}} \sum_{i=1}^{k} \int \frac{\mathbb{E}_{P_z}[\phi_{z,i}(X) W(y|X)]^2}{\mathbb{E}_{P_z}[W(y|X)]} d\mu$$

Theorem (Average information contraction bound)

$$|\theta_{z,i} - \theta_{z',i}| \approx r \, \mathbb{1}_{\{z_i \neq z_i'\}}$$

$$\frac{d p_{z \oplus i}}{d p_z} = 1 + \phi^r_{z,i}$$

$$\mathbb{E}_z \, \mathbb{E}_{p_{\theta_z}} \left[\|\theta_z - \hat{\theta}\|_2^2 \right]$$

$$\gtrsim r^2 k \left(1 - \sqrt{\frac{n}{2} \max_z \max_W \frac{1}{k} \sum_{i=1}^{k} \int \frac{\mathbb{E}_{p_z}[\phi^r_{z,i}(X) W(y|X)]^2}{\mathbb{E}_{p_z}[W(y|X)]} d\mu} \right)$$

Further bounds for (21)

$$\sum_{i=1}^{k} \frac{\mathbb{E}_{P_z}\left[W(y|X)\,\phi_{z,i}(x)\right]^2}{\mathbb{E}_{P_z}\left[W(y|X)\right]}$$

Additional conditions: (22)

- $\mathbb{E}_z\left[\phi_{z,i}^2\right] \leq \alpha^2$
- $\phi_{z,1}, \ldots, \phi_{z,k}$ are orthogonal

Then, since $\phi_{z,i}$ are zero-mean under P_z,

$$\sum_{i=1}^{k} \mathbb{E}_{P_z}\left[a(X)\,\phi_{z,i}(X)\right]^2 = \sum_{i=1}^{k} \langle a, \phi_{z,i} \rangle^2$$

$$\overset{?}{=} \sum_{i=1}^{k} \langle a - \mathbb{E}[a], \phi_{z,i} \rangle^2 \overset{?}{\leq} \|a - \mathbb{E}[a]\|^2$$

$$\Rightarrow \quad \sum_{i=1}^{k} \frac{\mathbb{E}_{P_Z}\left[W(y|X)\,\phi_{Z,i}\right]^2}{\mathbb{E}_{P_Z}\left[W(y|X)\right]} \leq \frac{\mathrm{Var}_{P_Z}\left(W(y|X)\right)}{\mathbb{E}_{P_Z}\left[W(y|X)\right]}$$

Additional conditions: 23

- $\mathbb{E}_Z\left[\phi_{Z,i}^2\right] \leq \alpha^2$
- $\phi_{Z,1}, \ldots, \phi_{Z,k}$ are independent and σ^2-subgaussian under P_Z

Under these assumptions, Gibbs variational formula can be used to show that

$$\sum_{i=1}^{k} \frac{\mathbb{E}_{P_Z}\left[\phi_{Z,i}(X)\,W(y|X)\right]^2}{\mathbb{E}_{P_Z}\left[W(y|X)\right]} \leq 2\alpha^2\sigma^2\, \mathbb{E}_{P_Z}\left[W(y|X) \ln \frac{W(y|X)}{\mathbb{E}[W(y|X)]}\right]$$

$\mathbb{E}_{p_z}[W(y|x)]$

$$\Rightarrow \sum_{i=1}^{k} \int \frac{\mathbb{E}_{p_z}[\phi_{z,i}(x) W(y|x)]^2}{\mathbb{E}_{p_z}[W(y|x)]} d\mu \le 2\alpha^2 \sigma^2 I(p_z; W)$$

Summary: (24)

$$\mathbb{E}_{z} \mathbb{E}_{p_{\theta_z}} \left[\|\theta_z - \hat{\theta}\|_2^2 \right]$$

$$\gtrsim r^2 k \left(1 - \sqrt{\frac{n}{2} \max_{z} \max_{W} \frac{1}{k} \sum_{i=1}^{k} \int \frac{\mathbb{E}_{p_z}[\phi_{z,i}^r(x) W(y|x)]^2}{\mathbb{E}_{p_z}[W(y|x)]} d\mu} \right)$$

Assumptions:

(i) $\mathbb{E}\left[\phi_{z,i}^{r\,2} \right] \le \alpha^2$

(ii) $\phi_{z,1}, \dots, \phi_{z,k}$ orthonormal

(iii) $\phi_{z,1}, \dots, \phi_{z,k}$ independent and σ^2-subgaussian (25)

$$\text{(i),(ii)} \Rightarrow \gtrsim r^2 k \left(1 - \sqrt{\frac{n\alpha^2}{} \max \left(\mathrm{Var}_{p_z}(W(y|x)) \right) d\mu} \right)$$

$$1 \quad \sqrt{2k} \quad z, W \int \frac{\cdots}{\mathbb{E}_{p_z}[W(y|X)]} \Big)$$

$$(i), (iii) \Rightarrow \gtrsim r^2 k \left(1 - \sqrt{\frac{n\alpha^2}{2k} \max_{z,W} I(p_z ; W)} \right)$$

So far we have made heuristics (1), (2), (3) concrete

(4) Information reduces due to W

- Communication constraints:

$$I(p_z ; W) \le \ell \quad \text{and} \quad \sum_y \frac{Var(W(y|X))}{\mathbb{E}[W(y|X)]} \le 2^\ell$$

- Privacy constraints: $\displaystyle\sum_y \frac{Var(W(y|X))}{\mathbb{E}[W(y|X)]} \le (e^\varepsilon - 1)^2$

References and related work:

Our presentation was based on:

"Unified lower bounds for interactive high-dimensional estimation under information constraints", Acharya, Canonne, Sun, and T.

Strong data processing approach:

- Zhang, Duchi, Jordan, Wainwright, 2013
- Xu and Raginsky, 2017
- Braverman, Garg, Ma, Woodruff, 2016

Remarks:

(i) Works only for the high-dimensional setting; may not work for discrete dist.

(ii) Can be recovered from the general bound we presented

Fisher information based approach:

- Barnes, Chen, Ozgur 2020
- Barnes, Han, Ozgur 2019, 2020.

Remarks:

(1) Works only for ℓ_2 loss

(2) Can be recovered from the general bound when:

$$\phi^r_{z,i}(x) \approx r\,\tilde{\phi}^r_{z,i}(x)$$

with

$$\|\tilde{\phi}^r_{z,i}(x)\|_2 \leq \text{constant}$$

for all r small

(Basically

$$\lim_{r\to 0} \sum_{i=1}^{k} \int \frac{\mathbb{E}_{P_z}[\tilde{\phi}^r_{z,i}(X)\,W(y|X)]^2}{\mathbb{E}_{P_z}[W(y|X)]}\,d\mu$$

$$\approx \mathrm{Tr}\,(J^W(\theta))$$

"Trace of the Fisher Info. Matrix")

Lots of other works:

Similar to what we presented:

- "Geometric lower bounds for distributed parameter estimation under Communication Constraints", Han, Ozgur, Weissman (2020 version).
- Ohad Shamir 2014 also develops a related approach

Extending the Fisher information based approach to a general ℓ_p loss:

- Sarbu and Zaidi, ISIT 2021

A very different approach: Duchi, Rogers 2019