Statistical Inference in Distributed and Constrained Settings

Part 2: Unified Lower Bounds

What will we see?

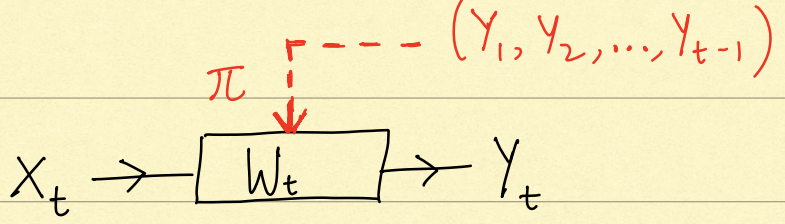
- More general lower bounds that apply to high-dimensional models
- Bounds that allow interaction

The model: (2)

- $\mathcal{P}_{(H)} = \{P_\theta, \theta \in (H)\}$
  a parametric family of distributions on $\mathcal{X}$
- $X_1, \ldots, X_n$
  i.i.d. samples from an unknown $P_\theta \in \mathcal{P}_{(H)}$

- $Y_1, \ldots, Y_n$

  Information constraints:



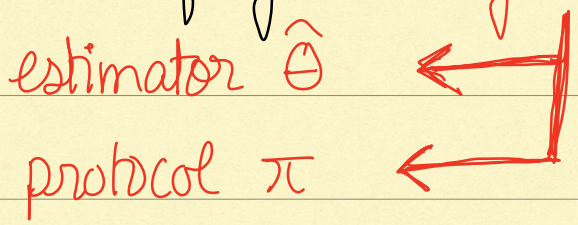
  sequentially interactive

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

  Estimate for unknown θ

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

  worst-case loss to capture the performance of your algorithm



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

Heuristics from information theory:

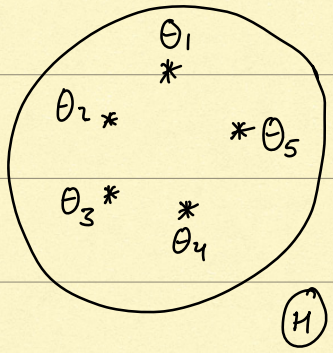
(1) Uncertainty about the unknown θ decreases as we observe $Y_1, \ldots, Y_n$

(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

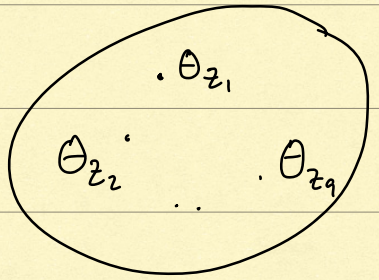
- Let one of them be chosen randomly and n samples be generated from it

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

→ Average uncertainty for each coordinate

⇝ Assouad's method

(Suited for high-dimensional + interactive)



Embedding:

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

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

$$\hat{Z} = \underset{z \in \{-1,1\}^K}{\operatorname{argmin}} \|\hat{\theta} - \theta_z\|$$

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

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

$$\approx r^2 \sum_{i=1}^{K} \mathbb{P}_{\theta_z}\left[z_i \neq \hat{z}\right]$$

For $Z \sim \text{unif}\{-1,1\}^K$, (Recall: $z \to \theta_z \to Y^n \to \hat{\theta} \to \hat{z}$)

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

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

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

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

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

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

Thus,

loss $\geq$ "uncertainty"

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

(2) Uncertainty to information

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

Lemma. 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,

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

Proof. $d(P_+, P_-)^2 \le 2\left(d(P_+, Q)^2 + d(P_-, Q)^2\right)$

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

$\le D(P_+\|Q) + D(P_-\|Q) = 2I(U \wedge X)$ ∎

Lemma $\Rightarrow$

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

$=$

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

(independence of $Z_1, \ldots, Z_k$)

$$\leq 2\, I\left(Z_i \wedge Y^n \mid Z^{-i}\right)$$

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

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

(1), (2) $\Rightarrow$

$$\mathbb{E}_Z\left[\mathbb{E}_{P_{\theta_2}}\left[\|\theta_2 - \hat{\theta}\|^2\right]\right] \quad \text{"loss"}$$

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

"information"

need to derive an upper bound for this information quantity

(3) Tensorization of information

(14)

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

(4) How W_ts reduce information

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

Simplifying notation: (15)

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

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

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

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

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

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

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

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

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

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

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

$$X_t - (Z_i, Z^{-i}) - Y^{t-1}$$

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

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

(17)

- P, Q two distributions on $\mathcal{X}$
- $W: \mathcal{X} \to \mathcal{Y}$

$$D\left(P^W \,\middle\|\, \tfrac{1}{2}P^W + \tfrac{1}{2}Q^W\right)$$

$$\leq d_{\chi^2}\left(P^W \,\middle\|\, \tfrac{1}{2}P^W + \tfrac{1}{2}Q^W\right)$$

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

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

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

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

(18)

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

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

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

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

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

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

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)

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

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

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

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

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)

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

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

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

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

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

Additional conditions:

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

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

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

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

Summary: (24)

$$\mathbb{E}_Z \mathbb{E}_{P_{\theta_2}}\left[\|\theta_2 - \hat{\theta}\|_2^2 \right]$$

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

Assumptions:

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

(ii) $\phi_{Z,1}, \ldots, \phi_{Z,k}$ orthonormal

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

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

$$\left(1 - \sqrt{2k \max_{z,W} \frac{I(p_z; W)}{\mathbb{E}_{p_z}[W(y|x)]}}\right)$$

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

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

(4) Information reduces due to W

- Communication constraints:

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

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