Infotheory for Statistics and Learning

Lecture 10

- Minimax lower bounds[1]
  - Le Cam's method
  - Assouad's method
  - Mutual information method

[1] based on notes by J. Duchi and Y. Wu and book by M. Wainwright

Recap - Minimax Risk Problem

- $\mathcal{P}$ denotes class of distributions defined on sample space $\mathcal{X}$.
- $\theta : \mathcal{P} \to \Theta$ denotes function that maps distribution P on $\theta(P)$
- **IID data:** $X^n = (X_1, \ldots, X_n)$ are n iid observations $X_i \sim P$
- **Estimator:** measurable function $\hat{\theta} : \mathcal{X}^n \to \Theta$

Minimax risk: Let $\rho : \Theta \times \Theta \to \mathbb{R}_+$ be a metric and $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ a non-decreasing function (e.g. $\rho(\theta, \theta')) = |\theta - \theta'|$ and $\Phi(t) = t^2$). The minimax risk $\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho)$ is defined as

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\Phi(\rho(\hat{\theta}(X^n), \theta(P))) \right]$$

From estimation to testing: Let $\{P_v\}_{v \in \mathcal{V}}$ be a 2δ-**packing**, then

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\Psi} \mathbb{P}[\Psi(X^n) \neq V] \tag{1}$$

Overview and outlook

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\Psi} \mathbb{P}[\Psi(X^n) \neq V]$$

Remaining challenges for minimax lower bound:

1. Find a good 2δ-packing
   - larger δ results in larger factor $\Phi(\delta)$
2. Find a good lower bound on the error probability
   - packing with uniform error probability seems desirable

Outlook

- Packing: metric entropy and packing numbers (lect 9)
- Fano method: $|\mathcal{V}| \geq 2$ and multiple hypothesis test (lect 9)
- Le Cam method: $|\mathcal{V}| = 2$ and binary hypothesis test
- Assouad method:$|\mathcal{V}| = 2^d$ and multiple binary hypothesis tests

Le Cam's method - Recap binary hypothesis test

Binary hypothesis test setup

- Let P_1 and P_2 be two distributions on set $\mathcal{X}$. Nature choses one distribution at random where $V \in \{1,2\}$ denotes the index. We then observe X drawn according P_V and try to guess realization of V with test $\Psi : \mathcal{X} \to \{1,2\}$.

Lemma (Lower bound on error probability)

$$\inf_{\Psi}\{P_1(\Psi(X) \neq 1) + P_2(\Psi(X) \neq 2)\} = 1 - \|P_1 - P_2\|_{TV} \quad (2)$$

Proof: Let $\mathcal{A} \subseteq \mathcal{X}$ denote decision region to output 1, then

$$P_1(\Psi(X) \neq 1)+P_2(\Psi(X) \neq 2) = P_1(\mathcal{A}^c)+P_2(\mathcal{A}) = 1-P_1(\mathcal{A})+P_2(\mathcal{A}).$$

Taking infimum, we have $\inf_{\Psi}\{P_1(\Psi(X) \neq 1) + P_2(\Psi(X) \neq 2)\} = \inf_{\mathcal{A}\subseteq\mathcal{X}}\{1 - P_1(\mathcal{A}) + P_2(\mathcal{A})\}. = 1 - \sup_{\mathcal{A}\subseteq\mathcal{X}}\{P_1(\mathcal{A}) - P_2(\mathcal{A})\}.$ $\square$

Le Cam's method

Q: How can we utilize a binary (packing with $|\mathcal{V}| = 2$) hypothesis test, i.e. lower bound on error probability, to lower bound the minimax error?

- (2) for iid observations $X_i \sim P_V$ with uniformly distributed V:

$$\inf_{\Psi}\{\mathbb{P}(\Psi(X^n) \neq V)\} = \frac{1}{2} - \frac{1}{2}\|P_1^n - P_2^n\|_{TV} \qquad (3)$$

- Le Cam's method: For any $\mathcal{P}$ for which there exists $P_1, P_2 \in \mathcal{P}$ such that $\rho(\theta(P_1), \theta(P_2)) \geq 2\delta$ we obtain from (1) and (3)

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta)\left[\frac{1}{2} - \frac{1}{2}\|P_1^n - P_2^n\|_{TV}\right]$$

  - Remaining task: Find *good* P_1, P_2 with large $\rho(\theta(P_1), \theta(P_2))$ and small $\|P_1^n - P_2^n\|_{TV}$, poss. upper bound[2] $\|P_1^n - P_2^n\|_{TV}$.

[2] e.g. using Pinsker ineq. $2\|P^n - Q^n\|_{TV}^2 \leq D(P^n||Q^n) \overset{iid}{=} nD(P_1||Q_1)$.

Example: Mean estimation of Gaussian distribution family

- Consider $\mathcal{P} = \{\mathcal{N}(\theta, \sigma^2) : \theta \in \mathbb{R}\}$ and $\|\cdot\|_1$ or $\|\cdot\|_2^2$ as loss.
- Let $X_1, \ldots, X_n$ be iid samples of $\mathcal{N}(\theta, \sigma^2)$ denoted as P_θ^n.
- Pick following two distributions: P_0^n and $P_{\theta'}^n$ with $\theta' = 2\delta$
  - $P_0, P_{\theta'} \in \mathcal{P}$ with $\rho(\theta(P_0), \theta(P_{\theta'})) = |0 - 2\delta| \geq 2\delta$
  - HW: $\|P_{\theta'}^n - P_0^n\|_{TV}^2 \leq \frac{1}{4}(\exp(n\frac{\theta'^2}{\sigma^2}) - 1) = \frac{1}{4}(\exp(4n\frac{\delta^2}{\sigma^2}) - 1)$
- Le Cam's lower bounds[3] with $\delta = \frac{\sigma}{2\sqrt{n}}$:
  - $\Phi \circ \rho = \|\cdot\|_1$:

$$\inf_{\hat{\theta}} \sup_{P_\theta \in \mathcal{P}} E_{P_\theta}\left[|\hat{\theta}(X_1^n) - \theta|\right] \geq \frac{\delta}{2}(1 - \frac{1}{2}\sqrt{\mathrm{e} - 1}) \geq \frac{\delta}{6} = \frac{\sigma}{12\sqrt{n}}$$

  - $\Phi \circ \rho = \|\cdot\|_2^2$:

$$\inf_{\hat{\theta}} \sup_{P_\theta \in \mathcal{P}} E_{P_\theta}\left[|\hat{\theta}(X_1^n) - \theta|^2\right] \geq \frac{\delta^2}{2}(1 - \frac{1}{2}\sqrt{\mathrm{e} - 1}) \geq \frac{\delta^2}{6} = \frac{\sigma^2}{24n}$$

[3] Pre-factors are not optimal, but scalings $\frac{\sigma}{\sqrt{n}}$ and $\frac{\sigma^2}{n}$ are sharp.

Squared Hellinger distance

- The **squared Hellinger distance** for $P_1, P_2 \ll \lambda$ is given by

$$H^2(P_1||P_2) = \int \left(\sqrt{\frac{\mathrm{d}P_1}{\mathrm{d}\lambda}} - \sqrt{\frac{\mathrm{d}P_2}{\mathrm{d}\lambda}} \right)^2 \mathrm{d}\lambda$$

  - we also write $H^2(p_1||p_2)$ with p_i pdf of P_i, $i = 1, 2$ (if exists).
- **Le Cam's ineq.**: Upper bound[4] on TV distance (proofs HW):

$$||P_1 - P_2||_{TV} \leq H^2(P_1||P_2)\sqrt{1 - \frac{H^2(P_1||P_2)}{4}} \qquad (4)$$

- For $P_1, \ldots, P_n$ and $P^{1:n} = \otimes_{i=1}^n P_i$ and likewise $Q^{1:n}$ we have
  - $\frac{1}{2}H^2(P^{1:n}||Q^{1:n}) = 1 - \prod_{i=1}^n (1 - \frac{1}{2}H^2(P_i||Q_i))$
  - and in the iid case $\frac{1}{2}H^2(P^{1:n}||Q^{1:n}) = 1 - (1 - \frac{1}{2}H^2(P_1||Q_1))^n \leq n\frac{1}{2}H^2(P_1||Q_1)$

[4] Upper bound is decreasing with increasing $H^2(P_1||P_2)$.

Example: Mean estimation of Uniform distribution family

- Consider $\mathcal{P} = \{\mathcal{U}([\theta, \theta+1]) : \theta \in \mathbb{R}\}$ and $\|\cdot\|_2^2$ as loss.
- Let $X_1, \ldots, X_n$ be iid samples of $\mathcal{U}([\theta, \theta+1])$ denoted as U_θ^n.
- Hellinger distance $H^2(U_\theta || U_{\theta'}) = H^2(U_{\theta'} || U_\theta)$ for $\theta, \theta' \in \mathbb{R}$
    - for $|\theta' - \theta| > 1$ we have $H^2(U_\theta || U_{\theta'}) = 2$
    - for $\theta' \in (\theta, \theta+1]$: $H^2(U_\theta || U_{\theta'}) = \int_\theta^{\theta'} \mathrm{d}t + \int_{\theta+1}^{\theta'+1} \mathrm{d}t = 2|\theta' - \theta|$
- Pick distributions U_θ and $U_{\theta'}$ such that $|\theta' - \theta| = 2\delta = \frac{1}{4n}$,
    - thus $\frac{1}{2} H^2(U_\theta^n || U_{\theta'}^n) \le \frac{n}{2} H^2(U_\theta || U_{\theta'}) = \frac{n}{2} 2|\theta' - \theta| = \frac{1}{4}$,
    - Le Cam's ineq.: $\|U_\theta^n - U_{\theta'}^n\|_{TV}^2 \le H^2(U_\theta^n || U_{\theta'}^n) \le \frac{1}{2}$.
- Le Cam's lower bounds[5]

$$\inf_{\hat{\theta}} \sup_{U_\theta \in \mathcal{P}} E_{U_\theta}\left[|\hat{\theta}(X_1^n) - \theta|^2\right] \ge \underbrace{\delta^2/2}_{=\frac{1}{2}(\frac{1}{8n})^2} \Big(\underbrace{1 - \|U_\theta^n - U_{\theta'}^n\|_{TV}}_{\ge 1 - \frac{1}{\sqrt{2}}}\Big)$$

[5] n^{-2} rate is optimal, e.g. achieved with estimator $\hat{\theta}(X_1^n) = \min_{1 \le i \le n} X_i$.

Le Cam's convex hull method

- **Idea:** Instead two distributions, consider two classes of distributions!
  - Separation over the convex hull can be (significantly) smaller than the point-wise separation.
- Subsets $\mathcal{P}_0, \mathcal{P}_1 \subset \mathcal{P}$ are 2δ-separated if $\rho(\theta(P_1), \theta(P_2)) \geq 2\delta$ for all $P_0 \in \mathcal{P}_0$ and $P_1 \in \mathcal{P}_1$.
- Le Cam's convex hull method: For any 2δ-separated classes $\mathcal{P}_0, \mathcal{P}_1 \subset \mathcal{P}$ we have

$$\mathfrak{M}(\theta(\mathcal{P}), \rho) \geq \frac{\delta}{2} \sup_{P_i \in \mathrm{ConvexHull}(\mathcal{P}_i)_{i=1,2}} [1 - \|P_1^n - P_2^n\|_{TV}]$$

  - Proof can be found in [Wainwright, Lemma 15.9]
- Improved bound for mean estimation of Gaussian dist. family: $\mathcal{P}_0 = \{P_0^n\}$ and $\mathcal{P}_1 = \{P_\theta^n, P_{-\theta}^n\} \rightarrow$ pre-factor $\frac{3}{20}$ instead $\frac{1}{12}$.

Le Cam for functionals

- Estimate functional $\theta : \mathcal{F} \to \mathbb{R}$ defined on set of densities $\mathcal{F}$
  - e.g. evaluational functional $\theta(f) = f(0)$ (density at point 0)
- **Lipschitz constant** of functionals w.r.t. Hellinger norm:

$$\omega\big(\epsilon; \theta, \mathcal{F}\big) = \sup_{f,g \in \mathcal{F}} \left\{ |\theta(f) - \theta(g)| : H^2(f \| g) \leq \epsilon^2 \right\}$$

  - measure of fluctuations of $\theta(f)$ when f is perturbed
- Le Cam's for functionals:

$$\inf_{\hat{\theta}} \sup_{f \in \mathcal{F}} E\left[\Phi\big(\hat{\theta} - \theta(f)\big)\right] \geq \frac{1}{4}\Phi\Bigg(\frac{\omega\big(\frac{1}{2\sqrt{n}}; \theta, \mathcal{F}\big)}{2}\Bigg)$$

- Proof: Set $\epsilon^2 = \frac{1}{4n}$, pick f, g that achieves[6] $\omega(\epsilon; \theta, \mathcal{F})$. Apply Le Cam method with $\delta = \frac{1}{2}\omega(\epsilon; \theta, \mathcal{F})$ and bound $\|P_f^n - P_g^n\|_{TV}^2 \leq H^2(P_f^n \| P_g^n) \leq nH^2(P_f \| P_g) \leq \frac{1}{4}$. □

[6] A sequence that comes arbitrary closely if supremum is not achieved.

Example: Point-wise estimation of Lipschitz densities

- Consider set of Lipschitz densities defined on $[-\frac{1}{2}, \frac{1}{2}]$ bounded away from zero & linear functional $\theta(f) = f(0)$.
- Approach: Apply lower bound on $\omega\big(\epsilon; \theta, \mathcal{F}\big)$ in Le Cam bound
  - To this end, pick $f_0, g \in \mathcal{F}$ with $H^2(f_0 \| g) = \frac{1}{4n} = \epsilon^2$, then $|f_0(0) - g(0)|$ provides lower bound to $\omega\big(\frac{1}{2\sqrt{n}}; \theta, \mathcal{F}\big)$
- Let f_0 uniform density and $g = f_0 + \phi(x)$ with perturbation $\phi(x)$ that is $\delta - |x|$ for $|x| \leq \delta$ and $|x - 2\delta| - \delta$ for $x \in [\delta, 3\delta]$.
  - It can be shown[7] that $H^2(f_0 \| g) \leq \frac{1}{3}\delta^3$.
  - Thus, setting $\delta^3 = \frac{3}{4n}$ results in $H^2(f_0 \| g) \leq \frac{1}{4n}$,
  - which gives lower bound $\omega\big(\frac{1}{2\sqrt{n}}; \theta, \mathcal{F}\big) \geq |f_0(0) - g(0)| = \delta$.
  - Considering $\Phi(t) = t^2$ gives then

$$\inf_{\hat{\theta}} \sup_{f \in \mathcal{F}} E\left[\big(\hat{\theta} - \theta(f)\big)^2\right] \geq \frac{1}{4}\Big(\frac{\omega\big(\frac{1}{2\sqrt{n}}; \theta, \mathcal{F}\big)}{2}\Big)^2 \geq \frac{1}{16}\left(\frac{3}{4n}\right)^{\frac{2}{3}}$$

[7] For details see Wainwright, Example 15.7.

Assoud's method - 2δ-Hamming separation

- **Idea:** Transform the estimation problem in multiple binary hypothesis testing problems exploiting the *problem structure*

Definition (2δ-Hamming separation)

The set $\{P_v\}_{v\in\mathcal{V}} \subset \mathcal{P}$ with $\mathcal{V} = \{-1,+1\}^d, d \in \mathbb{N}$ induces a 2δ-Hamming separation for $\Phi \circ \rho$ if there exists a function $\hat{v} : \theta(\mathcal{P}) \to \{-1,+1\}^d$ such that

$$\Phi(\rho(\theta, \theta(P_v))) \geq 2\delta \sum_{j=1}^{d} \mathbb{1}\{[\hat{v}(\theta)]_j \neq v_j\}$$

- index $v = (v_1, \ldots, v_d) \in \mathcal{V} = \{-1,+1\}^d$ is a binary sequence
- loss function of parameter θ can be component-wisely lower bounded

2δ-Hamming separation example

- Consider Laplace distributed data $x \in \mathbb{R}^d$ with $p(x) = \frac{1}{2^d} \exp(-\|x - \mu\|_1)$
- want to estimate mean μ in $\|\cdot\|_1$-distance, i.e. $\theta(p) = \mu$
- for some $\delta > 0$ define set $\{p_v\}_{v \in \mathcal{V}}$ with $v \in \mathcal{V} = \{-1, +1\}^d$ with $p_v(x) = \frac{1}{2^d} \exp(-\|x - \delta v\|_1)$ with mean $\theta(p_v) = \delta v$
- for any $\theta \in \mathbb{R}^d$ we have for the $\|\cdot\|_1$-error

$$\|\theta - \theta(p_v)\|_1 = \sum_{j=1}^{d} |\theta_j - \delta v_j| \geq \delta \sum_{j=1}^{d} \mathbb{1}\{\text{sign}(\theta_j) \neq v_j\}$$

since $|\frac{\theta_j}{\delta} - v_j| \geq \mathbb{1}\{\text{sign}(\theta_j) \neq v_j\} \geq 0$

$\Rightarrow$ $\{p_v\}_{v \in \mathcal{V}}$ is a 2δ-Hamming separation for $\|\cdot\|_1$-error because $[\hat{v}(\theta)]_j = \text{sign}(\theta_j)$ for all j satisfies the condition

Assouad's method

Lemma (Sharper version of Assouad's lemma)

Let $\{P_v\}_{v \in \mathcal{V}}$ be a 2δ-Hamming separation for loss $\Phi \circ \rho$, then

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \delta \sum_{j=1}^{d} \inf_{\Psi} \left[\mathbb{P}_{+j}([\Psi(X)]_j \neq +1) + \mathbb{P}_{-j}([\Psi(X)]_j \neq -1) \right]$$

with $\mathbb{P}_{+j}$ (resp. $\mathbb{P}_{-j}$) denotes the joint probability over X and random index V that is uniformly distributed over $\{+1, -1\}^d$ conditioned that the j-th coordinate $V_j = +1$ (resp. $V_j = -1$).

- Let $P_{+j}(x) = \frac{1}{2^{d-1}} \sum_{v \in \mathcal{V}: v_j = +1} P_v$ denote the marginal distribution conditioned $V_j = +1$ (resp. $P_{-j}(x)$ for $V_j = -1$), then the bound can be equivalently written as (see Le Cam)

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \delta \sum_{j=1}^{d} \left[1 - \| P_{+j} - P_{-j} \|_{TV} \right]$$

Proof of Lemma

- Since an average over $\{P_v\}_{v\in\mathcal{V}} \subset \mathcal{P}$ (which also satisfies the 2δ-Hamming separation condition) is smaller than the supremum over $\mathcal{P}$ we have for an estimator $\hat{\theta} : \mathcal{X} \to \Theta$

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P\left[\Phi(\rho(\hat{\theta}(X), \theta(P)))\right] \geq \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathbb{E}_{P_v}\Big[\underbrace{\Phi(\rho(\hat{\theta}(X), \theta_v))}_{\geq 2\delta \sum_{j=1}^{d} \mathbb{1}\{[\hat{v}(\hat{\theta}(X))]_j \neq v_j\}} \Big]$$

- with $\Psi(X) = \hat{v}(\hat{\theta}(X))$ we can rewrite the sum as follows

$$\frac{2}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathbb{E}_{P_v} \mathbb{1}\{[\Psi(X)]_j \neq v_j\} = \frac{2}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} P_v([\Psi(X)]_j \neq v_j)$$

$$= \underbrace{\frac{2}{|\mathcal{V}|}\sum_{v:v_j=+1} P_v([\Psi(X)]_j \neq v_j)}_{=\mathbb{P}_{+j}([\Psi(X)]_j \neq +1)} + \underbrace{\frac{2}{|\mathcal{V}|}\sum_{v:v_j=-1} P_v([\Psi(X)]_j \neq v_j)}_{=\mathbb{P}_{-j}([\Psi(X)]_j \neq -1)}$$

- result follows taking infima over all tests Ψ and estimators $\hat{\theta}$ $\square$

Standard version of Assouad's lemma

- Let $\mathcal{V}^{\setminus j}$ denote the index set excluding index v_j and let $P_{v^{\setminus j},+}$ (resp. $P_{v^{\setminus j},-}$) with $v^{\setminus j} \in \mathcal{V}^{\setminus j}$ denote P_v with $v \in \mathcal{V}$ where $v_j = +1$ (resp. $v_j = -1$) & $d_H(\cdot,\cdot)$ be the Hamming distance.
- Using triangle inequality of total variation (due to convexity)

$$
\begin{aligned}
\|P_{+j} - P_{-j}\|_{TV} &= \left\| \frac{1}{2^{d-1}} \sum_{v^{\setminus j} \in \mathcal{V}^{\setminus j}} P_{v^{\setminus j},+} - \frac{1}{2^{d-1}} \sum_{v^{\setminus j} \in \mathcal{V}^{\setminus j}} P_{v^{\setminus j},-} \right\|_{TV} \\
&\leq \frac{1}{2^{d-1}} \sum_{v^{\setminus j} \in \mathcal{V}^{\setminus j}} \left\| P_{v^{\setminus j},+} - P_{v^{\setminus j},-} \right\|_{TV} \\
&\leq \max_{\substack{v,v' \in \mathcal{V} \\ d_H(v,v')=1}} \|P_v - P_{v'}\|_{TV} \qquad \forall j \in \{1, 2, \ldots, d\}
\end{aligned}
$$

- Ineq. above leads to standard version of Assouad's lemma:

$$
\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \delta d \Big[1 - \max_{\substack{v,v' \in \mathcal{V} \\ d_H(v,v')=1}} \|P_v - P_{v'}\|_{TV} \Big]
$$

Example: Mean estimation of Normal distribution

- Consider $\mathcal{P} = \{\mathcal{N}(\theta, \sigma^2 I_d) : \theta \in \mathbb{R}^d\}$ and $\|\cdot\|_2^2$ as loss
- Construction of δ^2-Hamming separation with $\mathcal{V} = \{-1, +1\}^d$
  - Fix $\delta > 0$ and define $\theta_v = \delta v$ and $P_v = \mathcal{N}(\delta v, \sigma^2)\ \forall v \in \mathcal{V}$.
  - Family $\{P_v\}_{v\in\mathcal{V}}$ satisfies the condition since for any θ we have $\|\theta - \theta(P_v)\|_2^2 = \sum_{j=1}^d |\theta_j - \delta v_j|^2 \geq \delta^2 \sum_{j=1}^d \mathbb{1}\{\text{sign}(\theta_j) \neq v_j\}$.
- Assouad's lemma for n iid observations

$$\mathfrak{M}_n(\theta(\mathcal{P}), \|\cdot\|_2^2) \geq \frac{\delta^2}{2} \sum_{j=1}^d \left[1 - \|P_{+j}^n - P_{-j}^n\|_{TV}\right] \qquad (5)$$

- $\|\cdot\|_{TV}$ can be bounded using Pinsker inequality

$$\|P_{+j}^n - P_{-j}^n\|_{TV}^2 \leq \max_{\substack{v,v' \in \mathcal{V} \\ d_H(v,v')=1}} \|P_v - P_{v'}\|_{TV}^2 \leq \frac{1}{2} \max_{\substack{v,v' \in \mathcal{V} \\ d_H(v,v')=1}} D(P_v \| P_{v'})$$

  - with $D(P_v\|P_{v'}) = \frac{n}{2\sigma^2}\|\theta_v - \theta_{v'}\|_2^2 = \frac{2n}{\sigma^2}\delta^2$ since $d_H(v, v') = 1$.
- For $\delta^2 = \frac{\sigma^2}{8n}$ this gives lower bound[8] $\mathfrak{M}_n(\theta(\mathcal{P}), \|\cdot\|_2^2) \geq \frac{d\sigma^2}{8n}$.

[8]Bound asymptotically sharp since sample mean has mean square error $\frac{d\sigma^2}{n}$.

Example: Model fitting for logistic regression

- **Logistic regression:** Estimation of probabilities of binary (categorical) variable $Y_i \in \{-1, +1\}$ and dependent variable $X_i \in \mathbb{R}^d$ using a *logistic function*.
  - Bernoulli distribution: $P(Y_i = y | X_i, \theta) = \frac{1}{1+\exp(-yX_i^T\theta)}$
- **Task:** Estimation of parameter $\theta \in \mathbb{R}^d$ after observing a sequence of (Y_i, X_i), $1 \le i \le n$, that fits logistic model best!
  - maximization over set of Bernoulli distributions $\mathcal{P}$
  - $\theta(P)$ denotes estimation of parameter for $P \in \mathcal{P}$
  - use squared ℓ_2 error as metric
- Use Assouad method to **lower bound minimax risk**[9]
  - Same δ^2-Hamming separation as before: $\mathcal{V} = \{-1, +1\}^d$; for some $\delta > 0$ define $\theta_v = \delta v$ and $P_v = P(Y_i = y | X_i, \theta_v)$, [10] then

$$\mathfrak{M}_n(\theta(\mathcal{P}), \|\cdot\|_2^2) \ge \frac{d\delta^2}{2}\left[1 - \sqrt{\frac{\delta^2}{d}\|X\|_F^2}\right] \overset{\delta^2 = \frac{d}{4\|X\|_F^2}}{=} \frac{d^2}{16\|X\|_F^2}$$

[9] Frobenius norm $\|X\|_F^2 = \sum_{i-1}^n \sum_{j=1}^d X_{i,j}$, $X_{i,j}$ j-th element of X_i.

[10] As before, family $\{P_v\}_{v \in \mathcal{V}}$ satisfies the condition since for any θ we have $\|\theta - \theta(P_v)\|_2^2 = \sum_{j=1}^d |\theta_j - \delta v_j|^2 \ge \delta^2 \sum_{j=1}^d \mathbb{1}\{\text{sign}(\theta_j) \neq v_j\}$.

- For n iid observations using P_v^m we use Assoud's lemma (5)

$$\mathfrak{M}_n(\theta(\mathcal{P}), \|\cdot\|_2^2) \geq \frac{\delta^2 d}{2}\left(1 - \frac{1}{d}\sum_{j=1}^{d} \|P_{+j}^n - P_{-j}^n\|_{TV}\right)$$

- using C.S. and Jensen ineq. ($\|\cdot\|_{TV}$ is convex) we end up with a weaker version of Assoud's lemma used here [11]

$$\sum_j \|P_{+j}^n - P_{-j}^n\|_{TV} \overset{C.S.}{\leq} \sqrt{d}\left(\sum_j \|P_{+j}^n - P_{-j}^n\|_{TV}^2\right)^{1/2}$$
$$\overset{J.ineq.}{\leq} \sqrt{d}\left(\sum_j \frac{1}{2^d}\sum_v \|P_{v,+j} - P_{v,-j}\|_{TV}^2\right)^{1/2}$$

- It remains to bound $\|P_{v,+j} - P_{v,-j}\|_{TV}^2$ for Bernoulli distributions using Pinsker ineq. etc (HW). $\square$

[11] $P_{v,+j}$ is distribution where j-th element takes $+1$; $P_{+j}^n = \frac{1}{2^d}\sum_v P_{v,+j}$.

Mutual information method

- **Task**: Estimate parameter $\theta \in \Theta$ distributed by some prior π using estimator $\hat{\theta}$ observing data X.
- Accordingly, we have Markov chain $\theta - X - \hat{\theta}$ so that with data processing inequality we get

$$\inf_{P_{\hat{\theta}|\theta} : El(\theta,\hat{\theta}) \leq R_\pi^*} I(\theta; \hat{\theta}) \leq I(\theta; \hat{\theta}) \leq I(\theta, X) \leq \sup_\pi I(\theta; X)$$

  - Note, only lower bound involves R_π^* and loss.
  - Lower (upper) bound relate to rate-distortion-like (capacity-like) bounds.
- **Approach:** Derive lower and upper bounds and solve for R_π^*
  - Derivation of bounds can be difficult, see [W] for more approaches and discussion.

Example: Gaussian location model

- Upper bound: $Y = X + Z$, $Z \sim \mathcal{N}(0, I_p)$, $Z \perp X$
  - $\max_{P_X \in \{P_X : E\|X\|_2^2 \leq ps\}} I(X; X+Z) = \frac{p}{2}\log(1+s)$
- Lower bound: $X \sim \mathcal{N}(0, sI_p)$ and squared distortion
  - $\min\limits_{P_{Y|X} : E[\|Y-X\|^2] \leq p\epsilon} I(X;Y) = \begin{cases} \frac{p}{2}\log(\frac{s}{\epsilon}), & \epsilon < s \\ 0, & \text{otherwise} \end{cases}$
- Let $\theta \sim \mathcal{N}(0, S \cdot I_p)$ and $P_{X|\theta} \sim \mathcal{N}(\theta, \frac{1}{n}I_p)$ then
  - from the upper bound: $I(\theta, \hat{\theta}) \leq I(\theta, X) = \frac{p}{2}\log(1 + S \cdot n)$
  - from the lower bound
    $I(\theta, \hat{\theta}) \geq \min_{P_{\hat{\theta}|\theta} : E\|\theta - \hat{\theta}\|_2^2 \leq R_\pi^*} I(\theta, \hat{\theta}) = \frac{p}{2}\log\frac{S}{R_\pi^*/p}$

$\Rightarrow$ Combining both bounds and solve for R_π^*

$$R_\pi^* \geq \frac{S \cdot p}{1 + S \cdot n}$$