36-789: Topics in High Dimensional Statistics II **Fall 2015**

Lecture 2: October 29

Lecturer: Alessandro Rinaldo *Scribes: Jisu KIM*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

2.1 Recap[1]

As discussed in Lecture 1(Oct 27), general strategy to obtain minimax rates yields

$$\inf_{\hat{\theta}} \sup_{P\in\mathcal{P}} \mathbb{E}_P \left[w(d(\hat{\theta},\, \theta(P))\right] \geq w(\delta) \inf_{\psi} \max_{j=0,\cdots,M} \mathbb{P}_{\theta_j}\left(\psi(X)\neq j\right),$$

[2] where $\psi : X \to \{0, \cdots, M\}$ is a test function, and $d(\theta_i,\, \theta_j) \geq 2\delta$ for all $i \neq j$ (2δ-packing[3]). Denote

$$p_e(\theta_0, \cdots, \theta_M) = \inf_{\psi} \max_{j} \mathbb{P}_{\theta_j}(\psi(X)\neq j).$$

Now next job is to lower bound p_e by constant. If

$$p_e \geq c \geq 0,$$

then $w(\delta)c$ is a lower bound and $w(\delta) = w(\delta_n) \to 0$ will give you a rate, when $\delta = \delta_n \to 0$ as $n \to \infty$.[4] This rate is optimal if you can find a $\hat{\theta}(X)$ such that xxxx

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P w\left(d(\hat{\theta}(X),\, \theta)\right) = O\left(w(\delta_n)\right).$$

2.2 Distance between probability distributions[5]

Let P, Q be two probability measures on $(\Omega, \mathcal{A})$, having densities p and q with respect to some dominating measure (i.e. Lebesgue measure on $\mathbb{R}^d$). e.g., $\mu = P + Q$.

2.2.1 Total variation distance

Definition 2.1 [6]

$$d_{TV}(P,\, Q) = ||P - Q||_{TV} := \sup_{A\in\mathcal{A}} |P(A) - Q(A)|.$$

[1]See Section 2.2 in [T2008], p.79-80
[2]Proposition 2.3 in [D2014], p.13
[3]Section 2.2.1 in [D2014], p.13
[4]Equation (2.3) in [T2008], p.80
[5]See Section 2.4 in [T2008], p.83-91
[6]Definition 2.4 in [T2008], p.83 and Equation (1.2.4) in [D2014], p.7

Then following holds:[7]

- d_{TV} is a metric.
- $0 \le d_{TV} \le 1$.
- $d_{TV} = 0$ if and only if $P = Q$.
- $d_{TV} = 1$ if and only if P and Q are singular, i.e. there exists A such that $P(A) = 1$ and $Q(A) = 0$.
- d_{TV} is a very strong distance.

Lemma 2.2 *Scheffe lemma.*[8]

$$d_{TV}(P, Q) = \frac{1}{2} \int_{\mathscr{X}} |p(x) - q(x)| d\mu(x)$$

Proof: Take $A = \{x \in \mathscr{X} : \ q(x) \ge p(x)\}$. ■

2.2.1.1 Interpretation of d_{TV}

Suppose we observe X coming from either P or Q. And we have hypothesis test as

$$H_0 : X \sim P \text{ vs } H_a : X \sim Q.$$

Now for any test $\phi(X) \to \{0, 1\}$ with interpretation as $\phi(X) = \begin{cases} 1 & X \text{ comes from } Q \\ 0 & X \text{ comes from } P \end{cases}$, Type I error is $\mathbb{E}_P[\phi(X)]$, and Type II error is $\mathbb{E}_Q[1 - \phi(X)]$. Then

$$1 - d_{TV}(P, Q) = \inf_{\phi} \{\mathbb{E}_P[\phi(X)] + \mathbb{E}_Q[1 - \phi(X)]\}$$

for all tests(measurable functions) $\phi : \mathscr{X} \to \{0, 1\}$: exercise. **Proof:** Use Neyman-Pearson Lemma, the optimal test is $\phi(x) = \begin{cases} 1 & q(x) \ge p(x) \\ 0 & q(x) < p(x) \end{cases}$. ■

More generally,

$$\frac{1}{2} \int |p - q| = d_{TV}(P, Q) = 1 - \int_{\mathscr{X}} \min\{p(x), q(x) dx\}$$

follows from Scheffe lemma. Then following holds:

$$\inf_{0 \le f \le 1} \mathbb{E}_P[f] + \mathbb{E}_Q[1 - f] = \int_{\mathscr{X}} \min\{p(x), q(x)\} dx$$

$$\inf_{f, g \ge 0, f + g \ge 1} \{\mathbb{E}_P[f] + \mathbb{E}_Q[g]\} \ge \int \min\{p(x), q(x)\} = 1 - d_{TV}(p, q).$$

[7] See Properties of the total variance distance in Section 2.4 in [T2008], p. 84

[8] Lemma 2.1 in [T2008], p. 84

2.2.2 Hellinger

Definition 2.3 [9]

$$H(P, Q) = \sqrt{\int_{\mathscr{X}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 d\mu(x)}$$

Then following holds:[10]

- $H(P, Q)$ is a $L2$ distance between $\sqrt{p}$ and $\sqrt{q}$.
- $H(P, Q)$ gives canonical notion of regularity for statistical model: when $\sqrt{p(x)}$ is Hadamard differentiable.
- $H(P, Q)$ is a metric.
- $0 \leq H^2(P, Q) \leq 2$.
- $H^2(P, Q) = 2\left[1 - \underbrace{\int_{\mathscr{X}} \sqrt{p(x)}\sqrt{q(x)} d\mu(x)}_{\text{Hellinger Affinity}}\right]$.
- Tensorization[11]: if $P = \bigotimes_{i=1}^{n} P_i$ and $Q = \bigotimes_{i=1}^{n} Q_i$, then

$$H^2(P, Q) = 2\left[1 - \prod_{i=1}^{n}\left(1 - \frac{H^2(P_i, Q_i)}{2}\right)\right].$$

2.2.3 KL Divergence

Definition 2.4 [12]

$$KL(P, Q) = \begin{cases} \int_{\mathscr{X}} \log \frac{p(x)}{q(x)} p(x) d\mu(x) & P \ll Q \\ \infty & \textit{otherwise} \end{cases}$$

Then following holds:[13]

- $KL(P, Q) \geq 0$.
- $KL(P, Q) = 0$ if and only if $P = Q$.
- It is not symmetric and does not satisfy triangle inequality.
- Tensorization[14]: if $P = \bigotimes_{i=1}^{n} P_i$ and $Q = \bigotimes_{i=1}^{n} Q_i$, then

$$KL(P, Q) = \sum_{i=1}^{n} KL(P_i, Q_i).$$

[9] Definition 2.3 in [T2008], p.83, and Equation (2.2.2) in [D2014], p.15
[10] See Properties of the Hellinger distance in Section 2.4 in [T2008], p.83
[11] Equation (2.2.5) in [D2014], p.15
[12] Definition 2.5 in [T2008], p.84, and Section 1.2.2 in [D2014], p.5-7
[13] See Properties of the Kullback divergence in Section 2.4 in [T2008], p.83
[14] Equation (2.2.4) in [D2014], p.15

2.2.4 χ^2-divergence

Definition 2.5

$$\chi^2(P, Q) = \begin{cases} \int \left(\frac{p(x)}{q(x)} - 1\right)^2 q(x) d\mu(x) & \textit{if } P \ll Q \\ \infty & \textit{otherwise.} \end{cases}$$

Then following holds::[15]

- $\chi^2(P, Q) = \int \left(\frac{p(x)}{q(x)}\right)^2 q(x) d\mu(x) - 1$.
- $\chi^2(P, Q)$ equals f-divergence[16], with $f(x) = (x-1)^2$.
- Tensorization: if $P = \bigotimes_{i=1}^n P_i$ and $Q = \bigotimes_{i=1}^n Q_i$, then

$$\chi^2(P, Q) = \prod_{i=1}^n \left[1 - \chi^2(P_i, Q_i)\right].$$

2.2.5 Relationships among d_{TV}, H, KL, and χ^2

- $1 - d_{TV}(P, Q) = \int \min\{p, q\} dx \geq \frac{1}{2}\left[\int \sqrt{pq} dx\right]^2 = \frac{1}{2}\left[1 - \frac{H^2(P, Q)}{2}\right]^2$.[17]
- $\frac{1}{2} H^2(P, Q) \leq d_{TV}(P, Q) \leq H(P, Q)\sqrt{1 - \frac{H^2(P, Q)}{4}} \leq H(P, Q)$.[18]

Lemma 2.6 *(Donoho, Liu, 91) ("tensorization of d_{TV}")*

If $d_{TV}(P, Q) \leq 1 - \left(\frac{1-\delta^2}{2}\right)^{1/n}$ for some $\delta \in (0, 1)$, then $d_{TV}(P^n, Q^n) \leq \delta$.

Proof:

$$\begin{aligned} d_{TV}(P^n, Q^n) &\leq H(P^n, Q^n) \\ &= \sqrt{2\left[1 - \prod_{i=1}^n \left(1 - \frac{H^2(P, Q)}{2}\right)\right]} \\ &\leq \sqrt{2\left[1 - (1 - d_{TV}(P, Q))^2\right]} \\ &\leq \delta. \end{aligned}$$

■

Theorem 2.7 *(Pinsker inequality)*[19]

$$d_{TV}(P, Q) \leq \sqrt{\frac{KL(P, Q)}{2}}.$$

[15] See Properties of the $\chi 2$ divergence in Section 2.4 in [T2008], p.83
[16] For any function f, f-divergence is defined as $D_f(P, Q) = \int_{\mathscr{X}} f\left(\frac{p(x)}{q(x)}\right) q(x) d\mu(x)$. Refer to Section 1.2.3 in [D2014], p.7
[17] Lemma 2.3 in [T2008], p.86
[18] Lemma 2.3 in [T2008], p.86, and Proposition 2.4.(a) and Section 2.6.1 in [D2014], p.15, 30
[19] Lemma 2.5 in [T2008], p.88, and Proposition 2.4.(b) and Section 2.6.1 in [D2014], p.15, 30-31

- $KL(P, Q) \leq \chi^2(P, Q)$.[20]
- $d_{TV}(P, Q) \leq H(P, Q) \leq \sqrt{KL(P, Q)} \leq \sqrt{\chi^2(P, Q)}$. [21]

2.2.6 Minimax lower bounds based on 2 hypotheses

Recall that $p_e(P_0, P_1) = \inf_{\psi} \max_{i=0,1} P_i\left(\psi(X) \neq i\right)$. Now we want to lower bound it $[d(\theta(P_0), \theta(P_1)) \geq 2\delta]$

Theorem 2.8 [22] *1) If $d_{TV}(P_0, P_1) \leq \alpha (\leq 1)$, then $p_e(P_0, P_1) \geq \frac{1-\alpha}{2}$ (total variation version).*

2) $H^2(P_0, P_1) \leq \alpha (\leq 2)$, then $p_e \geq \frac{1}{2}\left[1 - \sqrt{\alpha\left(1 - \frac{\alpha}{2}\right)}\right]$ (Hellinger version).

3) If $KL(P_0, P_1) \leq \alpha < \infty$, $\chi^2(P_0, P_1) \leq \alpha < \infty$, then $p_e \geq \max\left\{\frac{1}{4}e^{-x}, \frac{1-\sqrt{\alpha/2}}{2}\right\}$ (Kullback/χ^2 version).

Proof: 2) and 3) are based on 1).

1):

$$
\begin{aligned}
p_e &= \inf_{\psi} \max_{i=0,1} P_i\left(\psi(X) \neq i\right) \\
&\geq \inf_{\psi}\left[\frac{1}{2} P_0\left(\psi(X) \neq 0\right) + \frac{1}{2} P_1\left(\psi(X) = 1\right)\right] \\
&= \frac{1}{2} \inf_{\psi} \left[\text{type I error} + \text{type II error}\right] \\
&= \frac{1}{2}\left[1 - d_{TV}(P, Q)\right].
\end{aligned}
$$

■

2.3 Le Cam's Lemma

Lemma 2.9 *Le Cam Lemma (Bin Yu's paper[23])*

$\Theta = \{\theta(P),\ P \in \mathcal{P}\}$. Suppose $\exists \Theta_1, \Theta_2 \subset \Theta$ such that $d(\theta_1, \theta_2) \geq 2\delta$, $\forall \theta_1 \in \Theta_1$, $\forall \theta_2 \in \Theta_2$. Let $\mathcal{P}_i \subset \mathcal{P}$ consisting of all $P \in \mathcal{P}$ such thath $\theta(P) \in \Theta_i$. Let $co(\mathcal{P}_i)$ be convex hall of $\mathcal{P}_i$, $i = 1, 2$. Then

$$
\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\left[w\left(d(\hat{\theta}, \theta(P))\right)\right] \geq w(\delta) \sup_{P_1 \in co(\mathcal{P}_1), P_2 \in co(\mathcal{P}_2)} \underbrace{\left[1 - d_{TV}(P_1, P_2)\right]}_{\int \min\{p_1, p_2\}}.
$$

Proof: Take $w(x) = x$

[20] Lemma 2.7 in [T2008], p.90
[21] Lemma 2.4 and Equation (2.27) in [T2008], p.90
[22] Theorem 2.2 in [T2008], p.90
[23] Lemma 1 in [Y1997], p.424-425

$$M = 2\sup_{P\in\mathcal{P}} \mathbb{E}_P\left[d(\hat{\theta},\theta(P))\right] \geq \mathbb{E}_{P_1}\left[d(\hat{\theta},\Theta_1)\right] + \mathbb{E}_{P_2}\left[d(\hat{\theta},\Theta_2)\right]$$

for any $P_i \in co(\mathcal{P}_i)$. Since

$$d(\hat{\theta},\Theta_1) + d(\hat{\theta},\Theta_2) \geq d(\Theta_1,\Theta_2) \geq 2\delta,$$

by hypothesis

$$\begin{aligned} M &\geq 2\delta\left(\mathbb{E}_{P_1}\left[\underbrace{\frac{d(\hat{\theta},\Theta_1)}{2\delta}}_{0\leq f}\right] + \mathbb{E}_{P_2}\left[\underbrace{\frac{d(\hat{\theta},\Theta_2)}{2\delta}}_{0\leq g}\right]\right) \\ &\geq 2\delta \inf_{f,g\geq 0,\ f+g\geq 1} \mathbb{E}_{P_1}\left[f(X)\right] + \mathbb{E}_{P_2}\left[g(X)\right] \\ &\geq 2\delta\left[1 - d_{TV}(P_1,P_2)\right] \end{aligned}$$

■

Example. Taking mixtures may help.

Suppose $\mathcal{P} = \{N(\theta, 1) :\ \theta \in \mathbb{R}\}$ and $\theta(N(\theta, 1)) = \theta$, and you want to lower bound minimax rate for $\hat{\theta}(P)$. If we consider $P_1 \sim N(\theta, 1)$ and $P_2 \sim N(0, 1)$, then

$$d_{TV}(N(\theta, 1),\, N(0, 1)) \approx \sqrt{\frac{2}{\pi}}|\theta| + o(\theta^2) \text{ as } \theta \to 0.$$

However, if we consider $\mathcal{P}_1 = \{N(\theta, 1),\, N(-\theta, 1)\}$ and $P_1 = \frac{1}{2}\left[N(-\theta, 1) + N(\theta, 1)\right] \in co(\mathcal{P}_1)$, then

$$d_{TV}\left(\frac{1}{2}\left[N(-\theta,1) + N(\theta,1)\right],\, N(0,1)\right) \approx \theta^2\Phi(1) + O(\theta^4) \text{ as } \theta \to 0,$$

where Φ is pdf of $N(0, 1)$. Hence taking mixtures gives better lower bound.

Reference

[T2008] Tsybakov, A. (2008). Introduction to Nonparametric Estimation, Springer.

[Y1997] Yu. B. (1997). Assuad, Fano, and Le Cam, Festschrift for Lucien Le Cam

[D2014] Duchi. J. (2014). John Duchi's notes on minimaxity from his class