10-704: Information Processing and Learning **Spring 2015**

Lecture 21: Strong Data Processing Inequalities

Lecturer: Akshay Krishnamurthy *Scribes: Che Zheng*

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

21.1 Review: Minimax Theory

We have been talking about techniques to lower bound the Minimax Risk, i.e.

$$\inf_T \sup_{\theta\in\Theta} \mathbb{E}_\theta[\Psi \circ \rho(T, \theta)], \tag{21.1}$$

where T is an estimator for a parameter θ that belongs to some family Θ, Ψ is a non-decreasing function with $\Psi(0) = 0$ and ρ is a semi-metric on $\Theta \times \Theta$.

Examples:

1. Min square error in normal mean problems.
2. MISE in non-parametric problems.
3. Adaptive Compressive Sensing.

Techniques:

1. Le Cam's method, which uses single vs single testing.
2. Fano's method, which uses multiple hypothesis testing.
3. Assouad's method, which uses multiple single vs single testing.

21.2 Review: Assouad's Method

In Assouad's Method, we want to find a packing V $= \{-1, 1\}^d$, s.t.

$$\forall\theta, \Phi \circ \rho(\theta, \theta_v) \geq 2\delta \sum_{j=1}^{d} \mathbb{1}\{\hat{v}(\theta) \neq v_j\} \tag{21.2}$$

where $\hat{v} : \Theta \to \{-1, 1\}^d$ is a function mapping from the parameter space Θ to the hypercube. Through this inequality, we are able to derive a lower bound for the original Minimax problem.

Example: Laplace mean estimator in l_1.
Suppose $p(x) \propto \exp(-||x-\mu||_1)$, where $x \in \mathbb{R}^d$. Let $v \in \{-1,1\}^d$ and $p_v(x) \propto \exp(-||x-\delta v||_1)$.

$$\|\theta - \theta(v)\|_1 = \sum_{j=1}^{d} |\theta_j - \delta v_j| \geq \delta \sum_{j=1}^{d} \mathbb{1}\{\text{sign}(\theta_j) \neq v_j\} \tag{21.3}$$

So $\hat{v}$ is sign function.

Setup: Pick v uniformly at random and let P_{+j} be the joint distribution on v, X conditioned on $v_j = +1$, similarly for P_{-j}. Then

$$R_n(\Theta, \Psi \circ \rho) \geq \delta \sum_{j=1}^{d} \inf_{\psi} [P_{+j}(\psi(x) \neq 1) + P_{-j}(\psi(x) \neq -1)] \tag{21.4}$$

where $\psi(x)$ is a testing function.

Example: Normal mean estimation in l_2^2 loss.
We consider the d-dimensional normal distribution with identity covariance, i.e. the distribution family is $P_\theta = N(\theta, I_{d\times d})$. Let $\theta_v = \delta v$ for $v \in \{-1,1\}^d$, then

$$||\theta - \theta_v||_2^2 \geq \delta^2 \sum_{j=1}^{d} \mathbb{1}\{\text{sign}(\theta_j) \neq v_j\} \tag{21.5}$$

Thus our subset $\{\theta_v\}_{v\in\{-1,+1\}^d}$ satisfies the conditions to use Assouad's method.

$$R(\Theta, ||\cdot||_2^2) \geq \frac{\delta^2}{2} \sum_{j=1}^{d} [1 - ||P_{+j} - P_{-j}||_{TV}] \tag{21.6}$$

$$||P_{+j} - P_{-j}||_{TV}^2 \leq \max_{\substack{v,v' \\ ||v-v'|| \leq 2}} ||P_v^n - P_{v'}^n||_{TV}^2 \leq \frac{1}{2} \max_{v,v'} KL(P_v^n || P_{v'}^n) \tag{21.7}$$

$$\tag{21.8}$$

The first inequality holds since total variance $||\cdot||_{TV}$ is convex. And for $||v - v'||_1 \leq 2$, we have

$$\frac{1}{2} KL(P_v^2 || P_{v'}^2) = \frac{n}{2} ||\theta_v - \theta_{v'}||_2^2 \leq 2n\delta^2 \tag{21.9}$$

Then

$$R(\Theta|| \cdot ||_2^2) \geq \frac{\delta^2}{2} \sum (1 - \sqrt{n\delta^2}) \geq \frac{\delta^2}{2} d(1 - \sqrt{n\delta^2}) \tag{21.10}$$

Set $\delta^2 = \frac{1}{4n}$, we have,

$$R(\Theta, ||\cdot||_2^2) \geq c\frac{d}{n} \tag{21.11}$$

21.3 Strong data processing inequalities

How can we leverage these lower bound techniques to new settings that arise in modern learning problems? One approach is to use *strong data processing inequalities*, as modern learning settings can be thought of as a classical problem with some transformation to the data, i.e.

$$\text{parameter } \rightarrow \text{classical data } \rightarrow \text{new data} \tag{21.12}$$

$$\theta \rightarrow X \rightarrow Z \tag{21.13}$$

Example: Local Differentially private channel: Channel $X \rightarrow Z$ must be differentially private for each data point, i.e. for each data point X_i we have distribution $Q(Z|X)$ s.t.

$$\sup_S \sup_{x,x' \in \mathcal{X}} \frac{Q(Z_i \in S|X_i = x)}{Q(Z_i \in S|X_i = x')} \leq exp(\alpha). \tag{21.14}$$

Example: Compression: Channel $X \rightarrow Z$ heavily compresses the input. For each input $X_i \in \mathbb{R}^d$, we pick uniformly a random subspace of m-dimension, and let $Z_i = (V_i, V_i X_i)$ where $V_i \in R^{m \times d}$ is a basis for subspace.

We would like to leverage existing technology to get lower bound in these settings for learning with Z. Clearly we can use data processing inequality, where we get $I(\theta, X) \geq I(\theta, Z)$ and indicates $R(Z^n, \theta) \geq R(X^n, \theta)$. But this bound is quite loose. Thus we are interested in strong data processing inequalities, where suppose we have channel $\theta \rightarrow X \rightarrow Z$, and $Q(Z|X)$ is the distribution of $Z|X$ with certain property, we want to show that $I(\theta; Z) \leq f(Q) I(\theta; X)$, where $f(Q) \ll 1$, which yields a much tighter lower bound.

21.4 Strong data processing inequality for α-local differential private channel

Suppose we have a α-local differential privacy channel $\theta \rightarrow X \in \mathcal{X} \rightarrow Z \in \mathcal{Z}$ and we get n samples X_1^n. For privacy reasons we use each X_i to create a new sample Z_i via channel $Q(Z_i|X_i)$. We require a per-example privacy, which is much more stringent than previous definition of differential privacy, that

$$\sup_S \sup_{x,x' \in \mathcal{X}} \frac{Q(Z_i \in S|X_i = x)}{Q(Z_i \in S|X_i = x')} \leq exp(\alpha) \tag{21.15}$$

The high-level claim is that if $\theta \rightarrow X \rightarrow Z$ is a α-locally differentially private channel, then $I(\theta, X) \leq \alpha^2 I(\theta, Z)$. More formally,

Theorem 21.1 *Let P_1, P_2 be distribution of $\mathcal{X}$ and let Q be a channel distribution that guarantees α-differential privacy ($\alpha \geq 0$). Define $M_i(S) = \int Q(S|x) dP_i(X)$, $i = 1, 2$ to be the marginal distribution. Then*

$$KL(M_1||M_2) + KL(M_2||M_1) \leq min\{4, e^{2\alpha}\}(e^\alpha - 1)^2 \|P_1 - P_2\|_{TV}^2. \tag{21.16}$$

Note for α small, where $e^\alpha - 1 \leq 2\alpha$ so we can write the rhs like

$$\leq c\alpha^2 ||P_1 - P_2||_{TV}^2 \tag{21.17}$$

The above theorem gives us an α^2 contraction in KL divergence, which means the effective sample size goes from n to $n\alpha^2$. This means that if we had n samples in the differentially private setting, it is as if we only had $n\alpha^2$ samples in the classical setting. So we need more samples in the new setting to learn well.

Proof: Let $m_1(z)$ be the density function of M_1, and similarly for m_2. We know

$$KL(M_1||M_2) + KL(M_2||M_1) = \int m_1(z) \log \frac{m_1(z)}{m_2(z)} d\mu(z) + \int m_2(z) \log \frac{m_2(z)}{m_1(z)} d\mu(z) \tag{21.18}$$

$$= \int (m_1(z) - m_2(z)) \log \frac{m_1(z)}{m_2(z)} d\mu(z) \tag{21.19}$$

Claim 1: For α differentially private channel Q with conditional density $q(\cdot|x)$:

$$|m_1(z) - m_2(z)| \leq c_\alpha \inf_x q(z|x)(e^\alpha - 1)||D_1 - D_2||_{TV}, c\alpha = \min\{2, e^\alpha\}. \tag{21.20}$$

Claim 2:

$$a, b \in R, |\log \frac{a}{b}| \leq \frac{|a-b|}{\min\{a,b\}} \tag{21.21}$$

If Claim 1 and Claim 2 are true, we have

$$|\log \frac{m_1(z)}{m_2(z)}| \leq \frac{|m_1(z) - m_2(z)|}{\min\{m_1(z), m_2(z)\}} \leq \frac{c_\alpha(e^\alpha - 1)\|P_1 - P_2\|_{TV} \inf_X q(z|x)}{\min\{m_1(z), m_2(z)\}} \leq c_\alpha(e^\alpha - 1)\|P_1 - P_2\|_{TV} \tag{21.22}$$

Similarly

$$|m_1(z) - m_2(z)| \leq c_\alpha(e^\alpha - 1)\|P_1 - P_2\|_{TV} \inf_x q(z|x) \tag{21.23}$$

Thus

$$KL(M_1||M_2) + KL(M_2||M_1) \leq c_\alpha^2(e^\alpha - 1)^2||P_1 - P_2||_{TV}^2 \int \inf_x q(z|x) d\mu(z) \tag{21.24}$$

And the integral is bounded by $\inf_x \int q(z|x)d\mu(z) = 1$.

Proof of Claim 1:

$$m_1(z) - m_2(z) = \int_\mathcal{X} q(z|x)(p_1(x) - p_2(x))d\mu(x) \tag{21.25}$$

$$= \int_\mathcal{X} q(z|x)\mathbb{1}\{P_1(x) \geq P_2(x)\}(P_1(x) - P_2(x))d\mu(x) \tag{21.26}$$

$$+ \int_\mathcal{X} q(z|x)\mathbb{1}\{P_1(x) < P_2(x)\}(P_1(x) - P_2(x))d\mu(x) \tag{21.27}$$

$$\leq \sup_{x\in\mathcal{X}} q(z|x) \int_{\mathcal{X}_+} (P_1(x) - P_2(x)) + \inf_{x\in\mathcal{X}} q(z|x) \int_{\mathcal{X}_-} (P_1(x) - P_2(x)) \tag{21.28}$$

$$=(\sup_x q(z|x) - \inf_x q(z|x)) \int_{\mathcal{X}_+} P_1(x) - P_2(x) \tag{21.29}$$

We know the second term is smaller than the total variance $\|P_1 - P_2\|_{TV}$ by definition. And for the first term

$$\sup_x q(z|x) - \inf_x q(z|x) \tag{21.30}$$

$$\leq \sup_{x,x'} |q(z|x) - q(z|x')| \tag{21.31}$$

$$= \inf_{\hat{x}} \sup_{x,x'} |q(z|x) - q(z|\hat{x}) + q(z|\hat{x}) - q(z|x')| \tag{21.32}$$

$$\leq 2 \inf_{\hat{x}} \sup_x |q(z|x) - q(z|\hat{x})| \tag{21.33}$$

$$= 2 \inf_{\hat{x}} q(z|\hat{x}) \sup_x |\frac{q(z|x)}{q(z|\hat{x})} - 1| \tag{21.34}$$

this gives

$$\leq 2|e^{\alpha} - 1| \inf_x q(z|x) \tag{21.35}$$

Since from α differentially privacy property $\frac{q(z|x)}{q(z|\hat{x})} \in [e^{-\alpha}, e^{\alpha}]$ and $|e^{\alpha} - 1| \geq |e^{-\alpha} - 1|$

$$\tag{21.36}$$

Proof of Claim 2: Since $\log(x) \leq x - 1$:

$$\log \frac{a}{b} \leq \frac{a}{b} - 1 = \frac{a-b}{b} \quad \text{If } a > b \tag{21.37}$$

$$\log \frac{b}{a} \leq \frac{b}{a} - 1 = \frac{b-a}{a} \quad \text{If } a \leq b \tag{21.38}$$

Then we get $|\log \frac{a}{b}| \leq \frac{|a-b|}{min\{a,b\}}$. ■

21.5 Strong data processing inequality for compressive sensing

Suppose we have $X_1, \ldots, X_n \sim N(0, \Sigma) \in \mathbb{R}^d$, and $Z = (U^T X, U)$, where $U \in \mathbb{R}^{d \times m}$ is an orthonormal basis for a random m-dimensional subspace, forms a channel as:

$$\Sigma \to X \to Z \tag{21.39}$$

Now instead of seeing $\{X_i\}_{i=1}^n$, we get $\{Z_i\} = \{(U_i^T X_i, U_i)\}_{i=1}^n$. We are interested in estimating Σ and how much information can we reveal about Σ.

Theorem 21.2 *Let D_0 be a distribution of (Z, U) where $X \sim N(0, \eta I)$, $U \sim unif$ and $Z = U^T X$. Let D_1 be the same distribution but $X \sim N(0, \eta I + \gamma vv^T)$, for $||v||_2 = 1$. Then:*

$$KL(D_1^n || D_0^n) \leq \frac{3}{2} \frac{\gamma^2}{\eta^2} \frac{nm^2}{d^2} \approx \frac{m^2}{d^2} KL(N^n(0, \eta I + \gamma vv^T) || N^n(0, \eta I)) \tag{21.40}$$

Similar to local differential privacy case, compression induces a contraction in KL divergence for Gaussian distributions, which can be used for lower bounds in covariance estimation problems, and the effective sample size is $\frac{nm^2}{d^2}$ rather than $\frac{nm}{d}$. But this result is far more specific than the previous one.

From the above theorem, we can show that:

$$\inf \sup \mathbb{E}[||\hat{\Sigma} - \Sigma||_2] \sim \sqrt{\frac{d^3}{nm^2} \log(d)} \tag{21.41}$$

while the uncompressed rate for covariance estimation in spectral norm is $\sqrt{\frac{d \log(d)}{n}}$.