10-704: Information Processing and Learning **Fall 2016**

Lecture 22: Nov 16

Lecturer: Aarti Singh

Note: *These notes are based on scribed notes from Spring15 offering of this course. LaTeX template courtesy of UC Berkeley EECS dept.*

22.1 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.

$$\text{parameter } \rightarrow \text{classical data } \rightarrow \text{new data} \tag{22.1}$$

$$\theta \rightarrow X \rightarrow Z \tag{22.2}$$

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.

$$\sup_S \sup_{x,x' \in \mathcal{X}} \frac{Q(Z_i \in S|X_i = x)}{Q(Z_i \in S|X_i = x')} \leq exp(\alpha). \tag{22.3}$$

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 $I(\theta, X) \geq I(\theta, Z)$. 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.

22.2 Strong data processing inequality for α-local differentially private channel

The following result is from [DJW13a] (Theorem 1). 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 and hence "local" privacy, which is much more stringent than previous definition of differential privacy, that

$$\sup_S \sup_{x,x' \in \mathcal{X}} \frac{Q(Z_i \in S|X_i = x)}{Q(Z_i \in S|X_i = x')} \leq exp(\alpha) \tag{22.4}$$

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

Theorem 22.1 *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*

$$KL(M_1||M_2) + KL(M_2||M_1) \leq (e^\alpha - 1)^2 \|P_1 - P_2\|_{TV}^2. \tag{22.5}$$

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

$$\leq c\alpha^2 ||P_1 - P_2||_{TV}^2 \tag{22.6}$$

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 m_2 be the density function of M_2 with respect to measure μ. We know

$$KL(M_1||M_2) + KL(M_2||M_1) = \int m_1(z) \log \frac{m_1(z)}{m_2(z)} d\mu(z) + \int m_2(z) \log \frac{m_2(z)}{m_1(z)} d\mu(z) \tag{22.7}$$

$$= \int (m_1(z) - m_2(z)) \log \frac{m_1(z)}{m_2(z)} d\mu(z) \tag{22.8}$$

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

$$|m_1(z) - m_2(z)| \leq \inf_x q(z|x)(e^\alpha - 1)||P_1 - P_2||_{TV}. \tag{22.9}$$

Claim 2:

$$a, b \in R, |\log \frac{a}{b}| \leq \frac{|a-b|}{\min\{a,b\}} \tag{22.10}$$

If Claim 1 and Claim 2 are true, we have

$$|\log \frac{m_1(z)}{m_2(z)}| \leq \frac{|m_1(z) - m_2(z)|}{\min\{m_1(z), m_2(z)\}} \leq \frac{(e^\alpha - 1)\|P_1 - P_2\|_{TV} \inf_x q(z|x)}{\min\{m_1(z), m_2(z)\}} \leq (e^\alpha - 1)\|P_1 - P_2\|_{TV} \tag{22.11}$$

since $\min\{m_1(z), m_2(z)\} \geq \inf_x q(z|x)$ (by Fatou's lemma). Similarly

$$|m_1(z) - m_2(z)| \leq (e^\alpha - 1)\|P_1 - P_2\|_{TV} \inf_x q(z|x) \tag{22.12}$$

Thus

$$KL(M_1||M_2) + KL(M_2||M_1) \leq (e^\alpha - 1)^2 ||P_1 - P_2||_{TV}^2 \int \inf_x q(z|x) d\mu(z) \tag{22.13}$$

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

Proof of Claim 1:

$$m_1(z) - m_2(z) = \int_{\mathcal{X}} q(z|x)(p_1(x) - p_2(x))d\mu(x) \tag{22.14}$$

$$= \int_{\mathcal{X}} q(z|x)\mathbb{1}\{p_1(x) \geq p_2(x)\}(p_1(x) - p_2(x))d\mu(x) \tag{22.15}$$

$$+ \int_{\mathcal{X}} q(z|x)\mathbb{1}\{p_1(x) < p_2(x)\}(p_1(x) - p_2(x))d\mu(x) \tag{22.16}$$

$$\leq \sup_{x\in\mathcal{X}} q(z|x) \int_{\mathcal{X}_+} |p_1(x) - p_2(x)| - \inf_{x\in\mathcal{X}} q(z|x) \int_{\mathcal{X}_-} |p_1(x) - p_2(x)| \tag{22.17}$$

$$=(\sup_x q(z|x) - \inf_x q(z|x)) \int_{\mathcal{X}} |p_1(x) - p_2(x)| \tag{22.18}$$

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

$$\sup_x q(z|x) - \inf_x q(z|x) \tag{22.19}$$

$$= \inf_{x'} q(z|x') \left[\frac{\sup_x q(z|x)}{\inf_{x'} q(z|x')} - 1\right] \tag{22.20}$$

$$\leq (e^\alpha - 1) \inf_{x'} q(z|x') \tag{22.21}$$

where the last step follows due to α-local differential privacy.

Proof of Claim 2: Since $\log(x) \leq x - 1$ for $x > 0$:

$$\text{If } a > b: \quad \log\frac{a}{b} \leq \frac{a}{b} - 1 = \frac{a-b}{b} \tag{22.22}$$

$$\text{If } a \leq b: \quad \log\frac{b}{a} \leq \frac{b}{a} - 1 = \frac{b-a}{a} \tag{22.23}$$

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

22.3 Strong data processing inequality for compressive sensing

The following result is from [AKS15] (Theorem 8). We consider the specific setting of estimating the covariance from compressed data. 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:

$$\Sigma \to X \to Z \tag{22.24}$$

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 compressed data reveal about Σ.

Theorem 22.2 *Let D_0 be a distribution of Z where $X \sim N(0, \eta I)$, $U \sim$ unif on the unit-sphere and $Z = U^T X$. Let D_1 be the same distribution but $X \sim N(0, \eta I + \gamma vv^T)$, for $||v||_2 = 1$, i.e. its covariance is a rank-1 perturbation of the covariance under D_0. Then:*

$$KL(D_1^n||D_0^n) \leq \frac{3}{2}\frac{\gamma^2}{\eta^2}\frac{nm^2}{d^2} \approx \frac{m^2}{d^2} KL(N^n(0, \eta I + \gamma vv^T)||N^n(0, \eta I)) \tag{22.25}$$

Similar to local differential privacy case, compression induces a contraction in KL divergence for Gaussian distributions, which can be used for lower bounds on covariance estimation for any distribution, 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 since it applies for only covariance estimation from compressed data.

From the above theorem, we can show that:

$$\inf_{\hat\Sigma} \sup_{\Sigma} \mathbb{E}[||\hat\Sigma - \Sigma||_2] = \Omega\left(\sqrt{\frac{d^3}{nm^2}}\right) \tag{22.26}$$

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

22.4 Strong data processing inequality for communication constrained mean estimation

The following result is from [DJW13b] (Proposition 2). We consider the specific setting of estimating the mean θ of a distribution supported on $[-1,1]^d$ under an independent communication-constrained protocol where there are m machines, each with a communication budget of $B_i, i = 1, \ldots, m$ for each of the machines. Under the independent protocol, each machine has n/m fraction of datapoints X_i and is allowed to transmit Y_i which is no more than B_i bits to a central server which combines the information received from all machines to generate an estimate $\hat{\theta}$. There is no further exchange of information between the server and machines, or between them machines themselves[1]. Also, for simplicity, we focus on the setting when $n = m$, i.e. only 1 data point per machine (see [DJW13b] for extension to general setting]. The goal is to lower bound the minimax communication constrained mean square error in estimating the mean:

$$\inf_{\text{ind protocols}(B_1,\ldots,B_m)} \inf_{\hat{\theta}} \sup_{\theta} \mathbb{E}[\|\theta - \hat{\theta}\|^2]$$

To lower bound the error in estimating mean, we follow the recipe we discussed last time of (1) finding a good discretization $\mathcal{P}'$ of the distributions under considerations $\mathcal{P}_{\theta\in\Theta}$ that are supported on $[-1,1]^d$ and have mean θ, (2) reducing the problem to testing between the distributions in $\mathcal{P}'$ and (3) lower bounding the testing error using (a variant of) Fano's inequality. Along the way, we will establish a strong data processing inequality for communication constrained mean estimation/testing.

- **Discretization:** Consider the subset $\Theta' = \{\theta : \theta = \delta v, v \in \{-1,+1\}^d\}$ and define the corresponding distributions by $P_\theta(x_j = v_j) = \frac{1+\delta v_j}{2}$ and $P_\theta(x_j = -v_j) = \frac{1-\delta v_j}{2}$. Note that by construction the mean $\mathbb{E}_{X\sim P_\theta}[X] = \delta v = \theta$.
- **Reduction to testing:** Consider a slightly stronger reduction to testing [DJW13b,DW13] where we can lower bound the estimation error in terms of the probability of error of a test that is allowed to make mistakes: Let V be uniformly sampled from $\{-1,1\}^d$. For any $t \geq 0$,

  $$\sup_{P_\theta \in \mathcal{P}'} \mathbb{E}_{X\sim P_\theta}[\|\theta - \hat{\theta}\|^2] \geq \delta^2(\lfloor t \rfloor + 1) \inf_{\hat{v}} P(d_H(\hat{v}, V) > t)$$

  where $d_H(\hat{v}, V)$ denotes the hamming distance between the binary vectors V and $\hat{v}$. Notice that for $t = 0$, we get the standard reduction we discussed in last class.
- **Lower bounding testing error:** The probability of error of such tests can be lower-bounded by a stronger Fano's lemma [DJW13b,DW13]: Let $V \to Y_{1:m} \to \hat{v}$ be a Markov chain, where v is uniform on $\mathcal{V}\{-1,+1\}^d$. For any $t \geq 0$

  $$P(d_H(\hat{v}, V) > t) \geq 1 - \frac{I(V, Y_{1:m}) + \log 2}{\log \frac{|\mathcal{V}|}{N_t}}$$

  where $N_t = \max_{v\in\mathcal{V}} |\{v' \in \mathcal{V} : d_H(v, v') \leq t|$, i.e. size of largest set of binary vectors that are within hamming distance t from any of the binary vectors in $\mathcal{V}$.

Now, we just need to upper bound the $I(V, Y_{1:m})$. Notice that, for each machine, we have the following Markov chain: $V \to X_i \to Y_i$. Data processing inequality tells us that $I(V, Y_i) \leq I(X, Y_i)$, however we will use a Sronger Data processing inequality by realizing that

$$\sup_{x_j} \sup_{v,v'} \frac{P(x_j|v)}{P(x_j|v')} \leq \frac{1+\delta}{1-\delta} = e^\alpha \quad \text{where} \quad \alpha = \log \frac{1+\delta}{1-\delta}$$

[1] The paper [DJW13b] also analyzes the interactive exchange setting, but for simplicity, in class we only focus on independent protocols.

This is a similar likelihood control as we had for α-local Differential Privacy. And hence, we have a similar data processing inequality (Lemma 3 in [DJW13b,DW13]):

$$I(V, Y_i) \le 2(e^{2\alpha} - 1)^2 I(X_i, Y_i)$$

Now we can bound $I(X_i, Y_i) \le \min H(X_i), H(Y_i) \le \min(d, B_i)$ since X_i is a d-dimensional binary vector and Y_i can only be represented using B_i bits due to communication constraint. We can now bound

$$I(V, Y_{1:m}) \le \sum_{i=1}^{m} I(V, Y_i) \le \sum_{i=1}^{m} 2(e^{2\alpha} - 1)^2 I(X_i, Y_i) = O\left(\delta^2 \sum_{i=1}^{m} \min(B_i, d)\right)$$

We can now complete the lower bound choosing $t = cd$ for some small constant c

$$\sup_{P_\theta \in \mathcal{P}'} \mathbb{E}_{X \sim P_\theta}[\|\theta - \hat{\theta}\|^2] = \Omega\left(\delta^2 d\left(1 - \frac{\delta^2 \sum_{i=1}^{m} \min(B_i, d) + \log 2}{d}\right)\right)$$

Choosing $\delta^2 \asymp \frac{d}{\sum_{i=1}^{m} \min(B_i, d)}$, we have

$$\sup_{P_\theta \in \mathcal{P}'} \mathbb{E}_{X \sim P_\theta}[\|\theta - \hat{\theta}\|^2] = \Omega\left(\frac{d}{m} \frac{m}{\sum_{i=1}^{m} \min(B_i/d, 1)}\right)$$

This expression tells us the tradeoff between communication and statistical efficiency, since the error rate for estimating the mean without communication constraint from $n = m$ samples is d/m. If $B_i \ge d$, then a similar error rate is possible, but if $B_i < d$, then there is a statistical price for communication constraint specified by B_i.

Remark: The technique of lower-bounding error using a construction based on hypothesis corresponding to the unit hypercube $\{-1, +1\}^d$ is also at the heart of Assouad's method (another method for proving lower bounds that we did not cover in class). Essentially, if the error metric is decomposable such that a packing $\mathcal{V} \in \{-1, +1\}^d$ can be found for which

$$\Phi(\rho(\hat{\theta}, \theta_v)) \ge \Phi(\delta) d_H(\hat{v}, v) = \Phi(\delta) \sum_{j=1}^{d} 1_{\hat{v}_j \ne v_j}$$

which many error metrics such as ℓ_1 and ℓ_2 loss satisfy, then the problem of testing between multiple hypothesis reduces to total error of multiple binary hypothesis tests.

References

[DJW13a] John C. Duchi, Michael I. Jordan, and Martin Wainwright, "Local Privacy and Statistical Minimax Rates", 54th Annual Symposium on Foundations of Computer Science (FOCS 2013).

[AKS15] Martin Azizyan, Akshay Krishnamurthy, and Aarti Singh, "Extreme Compressive Sampling for Covariance Estimation", https://arxiv.org/abs/1506.00898.

[DJW13b] John C. Duchi, Michael I. Jordan, and Martin Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints", Neural Information Processing Systems (NIPS 2013).

[DW13] J. C. Duchi and M. J. Wainwright, "Distance-based and continuum Fano inequalities with applications to statistical estimation", 2013, http://arxiv.org/abs/1311.2669.