Course: E2 209 Date **January 10th, 2019**

Lecture 2

Instructor: Himanshu Tyagi *Scribe: Avi Mohan*

Contents

- ⋆ Bounds on the growth of $d(P^n, Q^n)$ with the number of samples, n.
- ⋆ KL Divergence and its properties
  - Data processing inequality
  - Chain rule
  - Convexity
- ⋆ Pinsker's inequality and its proof
- ⋆ Preview of Lecture 3.

1 Subadditivity of TV distance

We ended the last lecture with an example involving binary hypothesis testing, where we hinted at how multiple samples could be used to improve the performance of the tests. This relied upon the assumption that the distance between the two distributions being tested would scale sufficiently with multiple samples, so that a "reliable" decision could eventually be taken in favor of one of the two.We now make precise the above statement and also give some estimates of how fast the distance between the two distributions increases.

Lemma 1.1. Let $(P_i, Q_i,\ 1 \le i \le n)$ be probability distributions defined on some measurable space $(\mathcal{X}, \mathcal{F})$. Define two new product measures P^n and Q^n as $P^n := P_1 \times P_2 \times \cdots P_n$ and $Q^n := Q_1 \times Q_2 \times \cdots Q_n$. Then,

$$d(P^n, Q^n) \le \sum_{i=1}^{n} d(P_i, Q_i) \tag{1}$$

Proof. We will show (1) for $n = 2$ and use an induction procedure to extend that to all $n \in \mathbb{N}$. Since TV distance is a true metric, it satisfies the triangle inequality, i.e.,

$$\begin{aligned}
d(P_1P_2, Q_1Q_2) &\le \underbrace{d(P_1P_2, Q_1P_2)}_{\text{term 1}} + \underbrace{d(Q_1P_2, Q_1Q_2)}_{\text{term 2}} \\
\text{term 1} &= \frac{1}{2}\sum_{x_1\in\mathcal{X}}\sum_{x_2\in\mathcal{X}}\Big|P_1(x_1)P_2(x_2) - Q_1(x_1)P_2(x_2)\Big| \\
&= \frac{1}{2}\sum_{x_1\in\mathcal{X}}\sum_{x_2\in\mathcal{X}}P_2(x_2)\Big|P_1(x_1) - Q_1(x_1)\Big| \\
&= d(P_1, Q_1) \\
\text{similarly, term 2} &= d(P_2, Q_2), \\
\Rightarrow d(P_1P_2, Q_1Q_2) &\le d(P_1, Q_1) + d(P_2, Q_2)
\end{aligned} \tag{2}$$

which proves (1) for $n = 2$. Assume that (1) holds for $n-1$. Now, invoking (2),

$$d(P^{n-1} \times P_n, Q^{n-1} \times Q_n) \leq d(P^{n-1}, Q^{n-1}) + d(P_n, Q_n) \leq \sum_{i=1}^{n} d(P_i, Q_i), \tag{3}$$

which completes the proof. □

Remark 1.1. The above lemma, under IID sampling, gives us

$$d(P^n, Q^n) \leq nd(P, Q). \tag{4}$$

Consider a binary hypothesis testing problem, where we wish to detect if a given set of samples of coin tosses was generated using a fair coin (bias $= \frac{1}{2}$) or one with bias $= \left(\frac{1}{2} + \epsilon\right)$ (recall Example 1 in the previous lecture). Denote the uniform probability mass function (pmf) by $P = \left\{\frac{1}{2}, \frac{1}{2}\right\}$ and the biased pmf by $P = \left\{\frac{1}{2} + \epsilon, \frac{1}{2} - \epsilon\right\}$. Now, suppose we wish to achieve a probability of error $P_e^* \leq 1/3$, (4) provides us with a guiding principle, viz.,

$$\begin{aligned} nd(P, Q) &\geq \frac{1}{3}, \\ \text{but } d(P, Q) &= \epsilon, \\ \Rightarrow n &\geq \frac{1}{3\epsilon}. \end{aligned} \tag{5}$$

This suggests that the sample complexity scales as $\frac{1}{\epsilon}$. However, at the end of Example 1 in Lecture 1 we had mentioned that this is actually of the order $\frac{1}{\epsilon^2}$. As we will now see, the bound in (4) is actually quite loose and provides a deceptively optimistic scaling. $d(\cdot, \cdot)$ really scales as $\sqrt{n}$, which gives rise to the requisite scaling.

2 Kullback-Leibler divergence

Another measure of distance between probability distributions is the Kullback-Leibler (KL) divergence. This is not called is distance because it does not satisfy two of the three axioms that metrics are required to satisfy, namely, symmetry and the triangle inequality. Given two distributions P and Q, their KL-divergence is defined as

$$D(P \parallel Q) = \begin{cases} \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}, \text{ if } supp(P) \subset supp(Q) \\ \infty, \text{ otherwise.} \end{cases} \tag{6}$$

In (6), $supp(P) := \{x \in \mathcal{X} : P(x) > 0\}$, the support of P. However, if P and Q have densities f and g resp., wrt some measure ν, then

$$D(P \parallel Q) = \begin{cases} \int_{x \in \mathcal{X}} f(x) \log \frac{f(x)}{g(x)} d\nu(x), \text{ if } supp(P) \subset supp(Q) \\ \infty, \text{ otherwise.} \end{cases} \tag{7}$$

$$= \begin{cases} \mathbb{E}_Q \left[\frac{dP}{dQ} \log \frac{dP}{dQ}\right], \text{ if } P \ll Q \\ \infty, \text{ otherwise.} \end{cases} \tag{8}$$

where, in (8), $P \ll Q$ means that P is absolutely continuous wrt Q, i.e., $Q(A) = 0 \Rightarrow P(A) = 0, \ \forall A \in \mathcal{F}$, in which case $\frac{dP}{dQ}$ is the Radon-Nikodym derivative of P wrt Q.

3 Properties of KL-divergence

3.1 The data processing inequality

Let $W : \mathcal{X} \to \mathcal{Y}$ be a fixed "channel," by which we mean that $W(\cdot|x)$ is a measure on $\mathcal{Y}$ for every $x \in \mathcal{X}$. Given any measure P on $\mathcal{X}$, denote by PW the distribution $\sum_{x \in \mathcal{X}} P(x) W(\cdot|x)$ on $\mathcal{Y}$. Using the definition

of TV distance and the triangle inequality, we immediately see that

$$d(PW, QW) \leq d(P, Q) \tag{9}$$

for any two measures P and Q. This result is intuitively quite satisfying, given the fact that TV distance is a measure of the ease with which two measures can be differentiated. Passing the measure through a channel only adds to the "randomness" (noise) in the samples, which should make it harder to distinguish between the resulting distributions. We now show that this is also true for the KL distance.

Theorem 3.1. $D(PW \parallel QW) \leq D(P \parallel Q)$

Proof. We will make use of an inequality called the log-sum inequality for the proof.

Lemma 3.2. For any non-negative reals $a_1, \cdots, a_n, b_1, \cdots, b_n$, we have

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \sum_{i=1}^{n} a_i \log \frac{\sum_{j=1}^{n} a_j}{\sum_{j=1}^{n} b_j}. \tag{10}$$

Proof. The proof is a simple application of Jensen's inequality. □

Now, getting back to the proof of Thm. 3.1, we see that

$$\begin{aligned}
D(PW \parallel QW) &= \sum_{y \in \mathcal{Y}} PW(y) \log \frac{PW(y)}{QW(y)} \\
&= \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} P(x)W(y|x) \log \frac{\sum_{z \in \mathcal{X}} P(z)W(y|z)}{\sum_{z \in \mathcal{X}} Q(z)W(y|z)} \right] \\
&\overset{\text{Lem. 3.2}}{\leq} \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} P(x)W(y|x) \log \frac{P(x)W(y|x)}{Q(x)W(y|x)} \right] \\
&= \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \underbrace{\sum_{y \in \mathcal{Y}} W(y|x)}_{=1 \ \forall x} \\
&= D(P \parallel Q).
\end{aligned} \tag{11}$$

□

We state the next two properties without proof.

3.2 Chain rule for KL-divergence

Lemma 3.3. Let $(P_i, Q_i,\ 1 \leq i \leq n)$ be probability distributions defined on some measure space $(\mathcal{X}, \mathcal{F})$. Define two new product measures P^n and Q^n as $P^n := P_1 \times P_2 \times \cdots P_n$ and $Q^n := Q_1 \times Q_2 \times \cdots Q_n$. Also, for $1 \leq i \leq n$, sample random variables $X_i \sim P_i$ and $Y_i \sim Q_i$. Then,

$$\begin{aligned}
D(P^n \parallel Q^n) &= D(P_{X_1, \cdots, X_n} \parallel Q_{Y_1, \cdots, Y_n}) \\
&= \sum_{i=1}^{n} \mathbb{E} D(P_{X_i|X^{i-1}} \parallel Q_{Y_i|Y^{i-1}}),
\end{aligned} \tag{12}$$

where $X^i := X_1, \cdots, X_i$ and the expectation in the i^{th} term in the above sum is computed wrt $P_{X^{i-1}}$. As a special case, when the $X_i's$ and the $Y_i's$ are all sampled IID P and Q respectively, $P_{X_i|X^{i-1}} = P$, $Q_{X_i|X^{i-1}} = Q$ and

$$D(P^n \parallel Q^n) = nD(P \parallel Q) \tag{13}$$

3.3 Convexity of KL-divergence

Lemma 3.4. KL-divergence is convex in its argument, i.e., if P_1, P_2, Q_1, Q_2 are measures on $(\mathcal{X}, \mathcal{F})$ and $\lambda \in [0,1]$, then

$$D(\lambda P_1 + (1-\lambda)P_2 \parallel \lambda Q_1 + (1-\lambda)Q_2) \leq \lambda D(P_1 \parallel Q_1) + (1-\lambda)D(P_2 \parallel Q_2) \tag{14}$$

4 Pinsker's inequality

Pinsker's inequality provides a relation between the two distance measures we have been studying. As a consequence, we will also see how the lower bound in the binary hypothesis testing example (5) can be tightened to get the ϵ^{-2} scaling mentioned in Lecture 1.

Theorem 4.1. Let P and Q be probability distributions defined on some measure space $(\mathcal{X}, \mathcal{F})$. Then,

$$d^2(P,Q) \leq \frac{1}{2} D(P \parallel Q). \tag{15}$$

Remark 4.1. Before we provide the proof, let us first solve the case of the missing $1/\epsilon$. Returning to Rem. 1.1, and invoking Pinkser's inequality for n IID samples, we see that

$$\begin{aligned} d^2(P^n, Q^n) &\leq \frac{1}{2} D(P^n \parallel Q^n) \\ &\overset{(13)}{=} \frac{n}{2} D(P \parallel Q). \end{aligned} \tag{16}$$

In the example discussed in Rem. 1.1, $P \equiv Ber(0.5)$ and $Q \equiv Ber(0.5+\epsilon)$.

$$\begin{aligned} D(P \parallel Q) &= 0.5 \log \frac{0.5}{0.5+\epsilon} + 0.5 \log \frac{0.5}{0.5-\epsilon} \\ &= 0.5 \log \frac{1}{4\,(0.25-\epsilon^2)} = -0.5 \log\,(1-4\epsilon^2) \end{aligned} \tag{17}$$

For $\epsilon \leq \frac{1}{3}$, one can use the above to show that $D(P \parallel Q) \leq \frac{18\epsilon^2}{5}$, which, when substituted in (16) gives us (recall that we needed $P_e^* \leq 1/3$)

$$\frac{1}{9} \leq n D(P \parallel Q) \leq \frac{18 n \epsilon^2}{5}, \tag{18}$$

whereby we get $n \geq \frac{c}{\epsilon^2}$ for some constant c, as promised. It now remains to be seen if there exists some hypothesis test that will actually hit this lower bound.

Proof. The proof proceeds in three steps.

- ⋆ Step 1: For any set $A \in \mathcal{F}$

$$D(P \parallel Q) \geq D(Ber(P(A)) \parallel Ber(Q(A))). \tag{19}$$

  This follows from the data processing inequality, with the "channel" defined as follows. For $y \in \{0,1\}$,

$$W(y|x) = y^{\mathbb{I}_{\{x \in A\}}}, \tag{20}$$

  which means that the channel outputs a "1" whenever the observation/input is in A. Clearly, with the above definition, if $X \sim P$, then $Y \sim WP(\cdot) \equiv Ber(P(A))$ and similarly if $X \sim Q$.

⋆ Step 2: Suppose (15) holds for $|\mathcal{X}| = 2$, i.e., when both the distributions are defined over a space with 2 elements. What this means, setting $P(A) = p$ and $Q(A) = q$, is that

$$\begin{aligned} (p-q)^2 &= d^2(Ber(p), Ber(q)) \\ &\leq \frac{1}{2} D(Ber(p) \parallel Ber(q)) \\ \Rightarrow \frac{1}{2} D(Ber(P(A)) \parallel Ber(Q(A))) &\geq (P(A), Q(A))^2, \ \forall A \in \mathcal{F}. \\ \Rightarrow \frac{1}{2} D(Ber(P(A)) \parallel Ber(Q(A))) &\geq \sup_{A \in \mathcal{F}} (P(A), Q(A))^2, \\ &= d(P, Q). \end{aligned} \tag{21}$$

The above, along with (4) proves Pinsker's inequality.

⋆ For $|\mathcal{X}| = 2$, define $f : [0,1] \times [0,1] \to \mathbb{R}$ as

$$f(p,q) = \underbrace{p \log \frac{p}{q} + (1-p) \log \frac{(1-p)}{(1-q)}}_{\text{KL-divergence} \times 2} - 2 \underbrace{(p-q)^2}_{\text{TV distance}}. \tag{22}$$

It suffices to show that $f(\cdot, \cdot) \geq 0$ to prove Pinsker's inequality for $|\mathcal{X}| = 2$. To do so, first notice that

- $f(q,q) = 0 \ \forall q \in [0,1]$, which means that one only needs to prove that for a fixed q, $f(p,q) \geq f(q,q), \ \forall p \in [0,1]$.
- $f(1-p, 1-q) = f(p,q)$. So, it further suffices to show that $f(p,q) \geq f(q,q), \ \forall p \geq q$.

But this latter statement is trivially true, since

$$\begin{aligned} \frac{\partial f}{\partial q} &= (p-q) \left(4 - \frac{1}{q(1-q)} \right) \\ &\geq 0, \text{ as long as } p \geq q \text{ and } q \in [0,1]. \end{aligned} \tag{23}$$

This completes the proof of Pinsker's inequality. □

Preview of Lecture 3:

⋆ Having studied binary testing, we will move on to the M-ary case and study Fano's inequality that lower bounds the performance of hypothesis tests in this case.

⋆ We will study the performance of the empirical pmf estimator for an M-ary distribution learning problem.

⋆ We will look at two formulations for performance evaluation, namely the minimax formulation (that was introduced in Lecture 1) and a Probably approximately correct (PAC) formulation, and how Fano's inequality can be used to bound minimax risk.