Teaching Note: Statistical Lower Bounds via Data-Processing

Daniel Russo

April 6, 2017

Warning: These are rough notes and typos are likely.

I'll present a tool for deriving lower bounds that can be used to establish any lower bound I will refer to in this course. A preliminary section reviews properties of of the Kullback-Leibler divergence. We will then show how various lower bounds follow from the so-called data-processing inequality of KL-divergence.

These techniques are closely related to traditional change of measure arguments. In particular, our end results are closely related to the general change of measure lemma in Kaufmann [2014].

1 Preliminaries: Facts about KL Divergence

Here I will review some basic properties of KL divergence. To keep everyhing simple, I will restrict focus to discrete random variables. More details can be found in chapter 2 of the beautiful textbook Cover and Thomas [2012]. A rigorous measure theoretic treatment is developed in Gray [2011].

Remark 1. *we will follow the the convention in information theory that* $0\log(0) = 0$*, which is consitent with the limit* $\lim_{x\to 0} x\log(x) = 0$.

Definition 1. *For two probability distributions* p *and* q *over* $\mathcal{X}$ *the Kullback-Leibler divergence is*

$$D(p(x)||q(y)) := \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right).$$

This can roughly be thought of as the extent to which observations from p "diverge" from what we would have expected under q. In general, the KL-divergence is not symetric ($D(p||q) \neq D(q||p)$); we may observe draws under p that are nearly impossible under q, even if all draws under q are fairly plausible under p. We next consider the definition of conditional KL divergence.

Definition 2. *(Conditional KL Divergence)*

$$D(p(y|x)||q(y|x)) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\left(\frac{p(y|x)}{q(y|x)}\right) = \sum_{x\in\mathcal{X}} p(x)\left(\sum_{y\in\mathcal{Y}} p(y|x)\log\left(\frac{p(y|x)}{q(y|x)}\right)\right).$$

Example 1 (Lemonade stand demand in uncertain whether)**.** *Suppose we are uncertain about the rate of demand at a lemonade stand. We know that demand is either* 0 *or* 1*. Let* y *denote the realized demand on a given day, and* $x \in \{0,1\}$ *denote whether it was rainy or sunny on that day (with* $x = 1$ *denoting "sunny".) It is sunny with probability* $3/4$ *under either* p *or* q*. We know that* *demand is always 0 when it is rainy. When it is sunny, demand is* 1 *with probability* θ_p *under* p *and is* 1 *with probability* θ_q *under* q*. Then*

$$D(p(y|x)||q(y|x)) = (3/4)D(\text{Bern}(\theta_p)||\text{Bern}(\theta_q))$$

where Bern(θ) *denotes the Bernoulli distribution with parameter* $\theta \in (0,1)$*. The main feature to take from this example is that we learn about demand only when it is sunny, and the probability of this event is precisely reflected in the conditional KL-divergence.*

Fact 1 (non-negativity)**.** *For any pmfs* p *and* q*,* $D(p||q) \geq 0$*.*

Proof. We show $-D(p||q) \leq 0$.

$$-D(p||q) = -\sum_x p(x)\log\left(\frac{p(x)}{q(x)}\right) = \sum_x p(x)\log\left(\frac{q(x)}{p(x)}\right) \leq \log\left(\sum_x p(x)\frac{q(x)}{p(x)}\right) = \log(1) = 0$$

where the inequality follows from Jensen's inequality. □

The next fact is important as it allows us to calculate the KL divergence in fairly complicated models. It basically follows from the factorization $p(x,y) = p(x)p(y|x)$ together with the fact that log takes products to sums $(\log(xy) = \log(x) + \log(y)$.

Fact 2 (Chain Rule)**.**

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$$

Proof. Using that $p(x,y) = p(x)p(y|x)$ and $q(x,y) = q(x)q(y|x)$ we have

$$\begin{aligned}\sum_{x,y} p(x,y)\log\left(\frac{p(x,y)}{q(x,y)}\right) &= \sum_{x,y} p(x,y)\left(\log\left(\frac{p(x)}{q(x)}\right) + \log\left(\frac{p(y|x)}{q(y|x)}\right)\right). \\ &= \sum_x p(x)\log\left(\frac{p(x)}{q(x)}\right) + \sum_{x,y} p(x,y)\log\left(\frac{p(y|x)}{q(y|x)}\right) \\ &= D(p(x)||q(x)) + D(p(y|x)||q(y|x)).\end{aligned}$$

□

Corollary 1 (Agreement on one variable)**.** *If* $p(x) = q(x)\ \forall x$ *then*

$$D(p(x,y)||q(x,y)) = D(p(y|x)||q(y|x)).$$

Corollary 2. *If* $(x_1, ...x_n)$ *are independent under both* p *and* q*, then*

$$D(p(x_1, x_2, ...x_n)||q(x_1, x_2, ..., x_n)) = \sum_{i=1}^{n} D(p(x_i)||q(x_i))$$

If the Kullback-Leibler divergence $D(p(x)||q(x))$ quantifies the degree to which observations from p allow us to rule out q, then it makes sense that we learn no more from observing a function $f(x)$ than from observing the raw data x. That is, processing the data-set cannot increase its information content. This is made formal in the following fact.

Fact 3 (Data Processing Inequality). *If $y = f(x)$ is a function of x, then*

$$D(p(y)||q(y)) \leq D(p(x)||q(x)).$$

Proof. We apply the chain rule of KL-divergence twice. First,

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + \underbrace{D(p(y|x)||q(y|x))}_{0} = D(p(x)||q(x)).$$

Then,

$$D(p(x)||q(x)) = D(p(x,y)||q(x,y)) = D(p(y)||q(y)) + D(p(x|y)||q(x|y)) \geq D(p(y)||q(y)).$$

□

2 Hypothesis testing lower bounds via data-processing

2.1 Notation

In the next subsection, we will consider problems in which $X_1, X_2, ... \sim f_\theta$ are drawn from a i.i.d from a density f_θ. The parameter θ is unknown, and the goal is to perform well (e.g. identify the mean) under every θ. I will write $\mathbb{P}_\theta(\cdot)$ to denote the probability measure under θ, so when writing $\mathbb{P}_\theta(X_1 \geq X_2 + X_3)$, we are implicitly integrating over draws of (X_1, X_2, X_3) from f_θ. We will study random variables that are functions of the sequence of X's, e.g. $Y = f(X_1, X_2, ..X_n)$. I'll denote the law of Y under θ by $P_\theta(Y) = \mathbb{P}_\theta(Y \in \cdot)$, and use

$$D(P_\theta(Y)||P_{\theta'}(Y))$$

to denote the KL-divergence between the distribution of Y under different parameters.

The notation for KL divergence is overloaded. I'll use

$$D(f_\theta||f_\theta') = \int \log\left(\frac{f_\theta(x)}{f_{\theta'}(x)}\right) f_\theta(x) dx$$

to denote the KL-divergence between f_θ and $f_{\theta'}$. For $p, q \in (0,1)$ let $d_{\mathrm{B}}(p||q) = p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)$ denote the KL divergence between Bernoulli distributions with parameters p and q.

2.2 Fixed Sample size

Consider an agent who observes

$$X_1, X_2, X_3, ... \sim f_\theta$$

for $\theta \in \{0,1\}$ They use a decision rule $\psi = (\psi_1, .\psi_2, ...)$ where for each $n \in \mathbb{N}$, $\psi_n : (X_1, ..., X_n) \mapsto \{0,1\}$ specifies a decision as a function of the random observations up to time n.

Theorem 3. *For any decision rule , $\mathbb{P}_0(\psi_n \neq 0) \leq \alpha$ and $\mathbb{P}_1(\psi_n \neq 1) \leq \beta$ implies*

$$n \geq \frac{d_{\mathrm{B}}(\alpha\,||\,1-\beta)}{D(f_0||f_1)}.$$

and

$$n \geq \frac{d_{\mathrm{B}}(\beta\,||\,1-\alpha)}{D(f_1||f_0)}.$$

Example 2. *Suppose under f_0, $X_1, X_2, ... \sim N(0, \sigma^2)$ and under f_1, $X_1, X_2, ... \sim N(\mu, \sigma^2)$. Then $D(f_1||f_0) = D(f_0||f_1) = \frac{\mu^2}{2\sigma^2}$. Suppose we would like the symmetric error guarantee $\mathbb{P}_\theta(\psi_n \neq \theta) \leq \delta$ for each of $\theta = 0, 1$. Then*

$$d_{\mathrm{B}}(\delta||1-\delta) = \delta \log\left(\frac{\delta}{1-\delta}\right) + (1-\delta)\log\left(\frac{1-\delta}{\delta}\right) \sim \log(1/\delta) \qquad \text{as } \delta \to 0.$$

This requires that

$$n \geq \frac{2 d_{\mathrm{B}}(\delta||1-\delta)}{\mu^2/\sigma^2} \underset{\delta\to 0}{\sim} \frac{2\log(1/\delta)}{\mathrm{SNR}^2}$$

where $\mathrm{SNR} = \mu/\sigma$ *is the signal to noise ratio.*

Proof. Set $H_n = (X_1, ..., X_n)$ to be the history of observations up to time n. We have

$$\begin{aligned} d_{\mathrm{B}}(\alpha, 1-\beta) &\leq D(P_0(\psi_n)||P_1(\psi_n)) && \text{(error constraint)} \\ &\leq D(P_0(H_n)||P_1(H_n)) && \text{(data processing)} \\ &= nD(f_0||f_1). && \text{(chain rule)} \end{aligned}$$

□

2.3 Adding action: optional sample collection

Let us extend the problem in the previous section. The agent still wants to predict whether $\theta = 0$ or $\theta = 1$, but now they can choose not to collect a sample at any particular time step. Can we formalize that the expected number of samples collected by the agent must be large in order to guarantee a low probability of error?

Formally, the agent has two actions available at each time step n, $A_n \in \{\text{Measure}, \text{Skip}\}$. After choosing A_n the agent observes Y_n where

$$Y_n = \begin{cases} X_n & if\ A_n = \text{Measure} \\ \emptyset & if\ A_n = \text{Skip} \end{cases}$$

The agent's action A_{n+1} is a function of the history of observations up to time $H_n = (A_1, Y_1, ...A_n, Y_n)$. The agent also chooses a prediction rule $\psi = (\psi_1, \psi_2, ...)$ with $\psi_n : H_n \mapsto \{0, 1\}$.

Theorem 4. *Let $T_n = \sum_{\ell=1}^{n} \mathbf{1}_{\{A_n = \text{Measure}\}}$. If $\mathbb{P}_0(\psi_n \neq 0) \leq \alpha$ and $\mathbb{P}_1(\psi_n \neq 1) \leq \beta$, then*

$$\mathbb{E}_0[T_n] \geq \frac{d_{\mathrm{B}}(\alpha \,||\, 1-\beta)}{D(f_0||f_1)}.$$

and

$$\mathbb{E}_1[T_n] \geq \frac{d_{\mathrm{B}}(\beta \,||\, 1-\alpha)}{D(f_1||f_0)}.$$

Proof. As before, we only prove the first lower bound. We have,

$$\begin{aligned} d_{\mathrm{B}}(\alpha, 1-\beta) &\leq D(P_0(\psi_n)||P_1(\psi_n)) && \text{(error constraint)} \\ &\leq D(P_0(H_n)||P_1(H_n)) && \text{(data processing)} \\ &= \mathbb{E}_0[T_n]D(f_0||f_1). && \text{(chain rule)} \end{aligned}$$

The final equality requires additional justification. For $m < n$, let $H_{m:n}$ denote the sub-history $(A_m, Y_m, ...A_n, Y_n)$. Applying the chain rule, one has

$$\begin{aligned} D(P_0(H_n)||P_1(H_n)) &= D(P_0(A_1, Y_1)||P_1(A_1, Y_1)) + D(P_0(H_{2:n}|A_1, Y_1)||P_1(H_{2:n}|A_1, Y_1)) \\ &= \underbrace{D(P_0(A_1)||P_1(A_1))}_{=0} + D(P_0(Y_1|A_1)||P_1(Y_1|A_1)) \\ &\quad + D(P_0(H_{2:n}|A_1, Y_1)||P_1(H_{2:n}|A_1, Y_1)) \end{aligned}$$

where we use that the choice of A_1 is deterministic, and therefore does not depend on the distribution of $X_1, X_2, ...$ Now,

$$D(P_0(Y_1|A_1)||P_1(Y_1|A_1)) = \mathbb{P}_0(A_1 = \text{Measure})D(f_0||f_1)$$

since $\mathbb{P}_\theta(Y_1 = \emptyset) = 1$ under both $\theta = 0$ and $\theta = 1$. We can repeat this process, and find

$$D(P_0(H_{2:n}|A_1, Y_1)||P_1(H_{2:n}|A_1, Y_1)) = D(P_0(Y_2|A_2, Y_1, A_1)||P_1(Y_2|A_2, Y_1, A_1)) + D(P_0(H_{3:n}|H_{1:2})||P_1(H_{3:n}|H_{1:2}))$$

where

$$D(P_0(Y_2|A_2, Y_1, A_1)||P_1(Y_2|A_2, Y_1, A_1) = \mathbb{P}_0(A_2 = \text{Measure})D(f_0||f_1).$$

Repeating this inductively we find

$$D(P_0(H_n)||P_1(H_n)) = \sum_{\ell=1}^{n} \mathbb{P}(A_\ell = \text{Measure}) = \mathbb{E}[T_n].$$

□

2.4 Technical extension to unbounded stopping times

Stopping problems can be viewed as a special case of the optional sampling problem described above in which we place additional restrictions on the algorithm. Formally, in a stopping problem the agent has two actions available at each time step n, $A_n \in \{\text{Measure}, \text{Stop}\}$. After choosing A_n the agent observes Y_n where

$$Y_n = \begin{cases} X_n & if\ A_n = \text{Measure} \\ \emptyset & if\ A_n = \text{Stop} \end{cases}$$

However, now the agent's decision rule is restricted: once selecting stop, she may never select measure again. The agent still employs a a prediction rule $\psi = (\psi_1, \psi_2, ...)$ with $\psi_n : H_n \mapsto \{0, 1\}$, but now we require that if ψ_n does not change after the agent has chosen to stop. (Formally, if $A_n = \text{Stop}$, then $\psi_m(H_m)) = \psi_n(H_n))$ for all $m > n$.)

Let

$$T = \inf\{n \in \mathbb{N} : A_n = \text{Stop}\}$$

denote the time at which the agent stops, and let $T \wedge n = \min\{T, n\}$.

Proceeding just as in the previous subsection, we can conclude

$$D(P_0(\psi_{T\wedge n})||P_1(\psi_{T\wedge n})) \leq \mathbb{E}_0[T \wedge n]D(f_0||f_1) \qquad \forall n \in \mathbb{N}. \tag{1}$$

For any stopping time T with $\mathbb{E}_0[T] < \infty$ and $\mathbb{E}_1[T] < \infty$, one has that $\mathbb{P}_0(T > n) \to 0$ and $\mathbb{P}_1(T > n) \to 0$ as $n \to \infty$. As a result,

$$\lim_{n\to\infty} D(P_0(\psi_{T\wedge n})||P_1(\psi_{T\wedge n})) = D(P_0(\psi_T)||P_1(\psi_T))$$

and

$$\lim_{n\to\infty} \mathbb{E}_0[T \wedge n] = \mathbb{E}_0[T].$$

We conclude that for any sequential rule satisfying $\mathbb{P}_0(\Psi_T \neq 0) \leq \alpha < 1/2$ and $\mathbb{P}_1(\psi_T \neq 1) \leq \beta < 1/2$ must satisfy

$$d_{\mathrm{B}}(\alpha, 1-\beta) \leq D(P_0(\psi_T)||P_1(\psi_T)) \leq \mathbb{E}_0[T] D(f_0||f_1)$$

or

$$\mathbb{E}_0[T] \geq \frac{d_{\mathrm{B}}(\alpha, 1-\beta)}{D(f_0||f_1)}.$$

This matches our approximation to the expected sample size of Wald's sequential probability ratio test.

2.5 Chernoff's Sequential Design of Experiments

In Chernoff's sequential design of experiments problem, the agent is able to influence the observations they collect by choosing among a set of k possible experiments. Formally, at each time n the agent chooses an action $A_n \in \{\text{Stop}, e_1, e_2, ..., e_k\}$ and receives an observation Y_n. Once she chooses to stop, she must continue to play Stop in all subsequent periods. When, $A_n = \text{Stop}$, $Y_n = \emptyset$, reflecting that nothing is learned. Otherwise, upon playing $A_n = e_i$ she observes

$$Y_n | A_n = e_i \sim f_\theta(y|e_i)$$

and conditioned on A_n, this observation is independent of all past observations. Set

$$D(\theta, \theta', e_i) = \int \log\left(\frac{f_\theta(y|e_i)}{f_{\theta'}(y|e_i)}\right) f_\theta(y|e_i) dy$$

to be the KL divergence between θ and θ' under experiment e_i.

The agent's goal is to confidently conclude whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$ while collecting as few measurements as possible. Let $\psi^* : \Theta \mapsto \{0, 1\}$ denote the optimal prediction rule, which sets $\psi^*(\theta) = 1$ if and only if $\theta \in \Theta_1$. Let $H_n = (A_1, Y_1, A_2, Y_2, ...A_n, Y_n)$ denote the history of observations, and as before, let $\psi_n : H_n \mapsto \{0, 1\}$ denote the prediction rule employed at time n and

$$T = \inf\{n | A_n = \text{Stop}\}$$

denote the stopping time.

In the past lecture, we claimed that under some technical restrictions Chernoff's procedure guarantees

$$\mathbb{P}_\theta(\psi_n \neq \psi^*(\theta)) \leq \delta \qquad \forall \theta \in \Theta \tag{2}$$

while using only

$$\mathbb{E}_\theta[T] \leq \frac{\log(1/\delta)}{\Gamma^*(\theta)} + o(\log(1/\delta)) \qquad \text{as } \delta \to 0$$

samples. Here

$$\Gamma^*(\theta) := \max_{W \in \Delta_+^k} \min_{\theta' : \psi^*(\theta') \neq \psi^*(\theta)} \sum_{i=1}^k W_i D(\theta, \theta', e_i)$$

is Chernoff's complexity measure we derived in the previous class. We now establish that this is optimal in a very strong asymptotic sense.

Theorem 5. *Any sequential procedure satisfying* (2) *must satisfy*

$$\mathbb{E}_\theta[T] \geq \frac{d_{\mathrm{B}}(\delta, 1-\delta)}{\Gamma^*(\theta)}$$

for every $\theta \in \Theta$.

Remark 2. *We showed above that* $d_{\mathrm{B}}(\delta, 1-\delta) \sim \log(1/\delta)$ *as* $\delta \to 0$, *and so this matches the upper bound up to lower order terms.*

Proof Sketch. Fix some θ, and consider any θ' such that $\psi^*(\theta) \neq \psi^*(\theta')$. We proceed just as in subsections 2.3 and 2.4. First, for any $n \in \mathbb{N}$, proceeding as in Subsection 2.3 we find

$$\begin{aligned} D(P_\theta(\psi_n)||P_{\theta'}(\psi_n)) &\leq D(P_\theta(H_n)||P_{\theta'}(H_n)) \\ &= \sum_{\ell=1}^{n}\sum_{i=1}^{k} \mathbb{P}_\theta(A_\ell = e_i) D(\theta, \theta', e_i) \\ &= \sum_{i=1}^{k} \mathbb{E}_\theta[S_n(e_i)] D(\theta, \theta', e_i) \end{aligned}$$

where $S(n)(e_i) = \sum_{\ell=1}^{n} \mathbf{1}_{\{A_n = e_i\}}$ is the number of times e_i is played prior to time n. As in Subsection 2.4, taking $n \to \infty$, we find

$$\begin{aligned} D(P_\theta(\psi_T)||P_{\theta'}(\psi_T)) &\leq \sum_{i=1}^{k} \mathbb{E}_0[S_T(e_i)] D(\theta, \theta', e_i) \\ &= \mathbb{E}_\theta[T] \sum_{i=1}^{k} \frac{\mathbb{E}_\theta[S_T(e_i)]}{\mathbb{E}_\theta[T]} D(\theta, \theta', e_i) \end{aligned}$$

By our constraint on the probability of incorrect selection, we know $D(P_\theta(\psi_T)||P_{\theta'}(\psi_T)) \geq d_{\mathrm{B}}(\delta, 1-\delta)$, which implies

$$d_{\mathrm{B}}(\delta, 1-\delta) \leq \mathbb{E}_\theta[T] \sum_{i=1}^{k} \frac{\mathbb{E}_\theta[S_T(e_i)]}{\mathbb{E}_\theta[T]} D(\theta, \theta', e_i).$$

Now, to clean up notation, let $B(\theta) = \{\theta' : \psi^*(\theta') \neq \psi^*(\theta)\}$ denote the set of parameters under which the correct decision is different than under θ. Then, since the inequality above holds for all $\theta' \in B(\theta)$, we find

$$d_{\mathrm{B}}(\delta, 1-\delta) \leq \mathbb{E}_\theta[T] \min_{\theta' \in B(\theta)} \sum_{i=1}^{k} \left(\frac{\mathbb{E}_\theta[S_T(e_i)]}{\mathbb{E}_\theta[T]} \right) D(\theta, \theta', e_i).$$

Since $\mathbb{E}_\theta[T] = \sum_{i=1}^{k} \mathbb{E}_\theta[S_T(e_i)]$, we can see that $(\mathbb{E}_\theta[S_T(e_i)]/\mathbb{E}_\theta[T])_{i \in \{1,..,k\}}$ is a probability vector. Therefore, we have

$$d_{\mathrm{B}}(\delta, 1-\delta) \leq \mathbb{E}_\theta[T] \max_{W \in \Delta_k^+} \min_{\theta' \in B(\theta)} \sum_{i=1}^{k} W_i D(\theta, \theta', e_i) = \mathbb{E}_\theta[T]\Gamma^*(\theta).$$

□

References

T.M. Cover and J.A. Thomas. *Elements of information theory.* John Wiley & Sons, 2012.

R.M. Gray. *Entropy and information theory.* Springer, 2011.

Emilie Kaufmann. *Analyse de stratégies bayésiennes et fréquentistes pour l'allocation séquentielle de ressources.* PhD thesis, Paris, ENST, 2014.