**Information and Coding Theory** **Autumn 2017**

# Lecture 8: October 23, 2017

Lecturer: Madhur Tulsiani

In this lecture and the next one, we will use lower bounds on hypothesis testing developed before to understand how well we can "learn" properties of distributions using samples. Much of the presentation here (and the first half the next lecture) is based on the excellent set of lecture notes by John Duchi [Duc16] (also linked from the course webpage) which I highly recommend for a more in-depth treatment of the subject.

## 1 Minimax risk and reduction to hypothesis testing

Let $\Pi$ be a set of distributions on $U$ and let $\theta : \Pi \to \Theta$ be any map which we think as a "property" of $P$. We consider an estimator $\widehat{\theta} : U^n \to \Theta$, which takes $n$ independent samples from $P$ as input, and tries to estimate $\theta(P)$. The quality of the estimator is measured by a *loss function* $\ell : \Theta \times \Theta \to \mathbb{R}_+$. If we use an estimator $\widehat{\theta}$ and the data comes from a distribution $P$, the *expected* loss is $\mathbb{E}_{\overline{\mathbf{x}} \sim P^n}\left[\ell\left(\widehat{\theta}(\overline{\mathbf{x}}), \theta(P)\right)\right]$. The goal is to come up with an estimator, which minimizes the loss even for the worst-case distribution i.e., we want to understand

$$\mathcal{M}_n(\Pi, \ell) := \inf_{\widehat{\theta}} \sup_{P \in \Pi} \mathop{\mathbb{E}}_{\overline{\mathbf{x}} \sim P^n}\left[\ell\left(\widehat{\theta}(\overline{\mathbf{x}}), \theta(P)\right)\right] .$$

The quantity $\mathcal{M}_n(\Pi, \ell)$ is also called the *minimax risk*. As an example, consider the case $\Pi = \{P_v\}_{v \in \mathcal{V}}$, $\Theta = \mathcal{V}$ and $\theta(P_v) = v$. We take $\ell(\widehat{\theta}, \theta) = 1$ if $\widehat{\theta} \neq \theta$ and 0 otherwise. The goal is to find

$$\mathcal{M}_n(\Pi, \ell) = \inf_{\widehat{\theta}} \sup_{v \in \mathcal{V}} \mathop{\mathbb{P}}_{\overline{\mathbf{x}} \sim P_v^n}\left[\theta(\overline{\mathbf{x}}) \neq v\right] ,$$

which is very similar to the setting of multiple hypothesis testing introduced in the previous lecture. While the minimax risk requires bounding the probability of error for the *worst* distribution in $\Pi$, in the previous lecture we developed a lower bound on the probability that the estimator errs for a *randomly chosen* distribution from $\Pi$. Of course this is still a lower bound. If we have some additional information about $\mathcal{V}$, we can find a "hard set" $\Pi' \subseteq \Pi$ and apply the bound from the previous lecture for a randomly chosen distribution from $\Pi'$. This is still a lower bound on the minimax risk. All the lower bounds developed below are essentially of this form, where we identify a hard subset of distributions and apply the bounds for hypothesis testing. In general, the notion of a "hard subset" of distributions needs to be developed with respect to the loss function $\ell$.

We will restrict the discussion here to loss functions $\ell$ which only depend on some form of distance between $\widehat{\theta}$ and $\theta$. In particular, we consider

$$\ell\left(\widehat{\theta},\theta\right) \;=\; \Phi\left(d\left(\widehat{\theta},\theta\right)\right) \;=\; \Phi\circ d\left(\widehat{\theta},\theta\right),$$

where $d(\cdot,\cdot)$ is a metric (obeying triangle inequality) and $\Phi$ is a monotone function. In fact, $\ell(\widehat{\theta},\theta) = \left\|\widehat{\theta}-\theta\right\|_2^2$ will suffice for our purposes, but we state the reduction from lower bounds on minimax risk to hypothesis testing for any $\ell$ of the form above.

**Lemma 1.1** *Let $\{P_v\}_{v\in\mathcal{V}} \subseteq \Pi$ be a finite set of distributions such that $\forall v_1, v_2 \in \mathcal{V}$ with $v_1 \neq v_2$,m $d(\theta(P_{v_1}, P_{v_2})) \geq 2\delta$. Let $\ell$ be as above. Then,*

$$\mathcal{M}(\Pi,\ell) \;\geq\; \Phi(\delta)\cdot\inf_T\left\{\mathbb{P}\left[T(\overline{\mathbf{x}})\neq V\right]\right\}.$$

Note that the setting in the RHS above is exactly as considered in hypothesis testing. We think of $V$ as uniformly distributed over the set $\mathcal{V}$ and $\overline{\mathbf{x}}$ as drawn from $P_v^n$.

**Proof:** Let $\widehat{\theta} : U^n \to \mathcal{V}$ be any estimator. We define a classifier $T : U^n \to \mathcal{V}$ as follows

$$T(\overline{\mathbf{x}}) \;:=\; \arg\min_{v\in\mathcal{V}}\; d\left(\widehat{\theta}(\overline{\mathbf{x}}),\theta(P_v)\right).$$

Note that if $V = v$ and $T(\overline{\mathbf{x}}) = v' \neq v$, we must have $d\left(\widehat{\theta}(\overline{\mathbf{x}}),\theta(P_v)\right) \geq \delta$ (why?) This implies that if $T$ makes an error on input $\overline{\mathbf{x}}$, then we must have $\ell(\widehat{\theta},\theta) \geq \Phi(\delta)$. Thus, we get

$$\begin{aligned}
\mathcal{M}_n(\Pi,\ell) \;&\geq\; \sup_{P\in\Pi}\;\mathop{\mathbb{E}}_{\overline{\mathbf{x}}\sim P^n}\left[\Phi\circ d\left(\widehat{\theta}(\overline{\mathbf{x}}),\theta(P)\right)\right]\\
&\geq\; \mathop{\mathbb{E}}_{v\in\mathcal{V}}\mathop{\mathbb{E}}_{\overline{\mathbf{x}}\sim P^n}\left[\Phi\circ d\left(\widehat{\theta}(\overline{\mathbf{x}}),\theta(P_v)\right)\right]\\
&\geq\; \Phi(\delta)\cdot\mathbb{P}\left[T(\overline{\mathbf{x}})\neq V\right]\\
&\geq\; \Phi(\delta)\cdot\inf_T\left\{\mathbb{P}\left[T(\overline{\mathbf{x}})\neq V\right]\right\}.
\end{aligned}$$

The last inequality above used the fact that the error of the classifier $T$ here is lower bounded by the error of the *best* classifier. This completes the proof. ∎

## 2 Lower bounds via binary hyothesis testing (Le Cam's method)

We return to out favorite example of biased coins. Let $U = \{0,1\}$ and let $\Pi$ be the set of all distributions on $\{0,1\}$. For a distribution $P$ on $U$, let $\theta(P) := p(1) = \mathbb{E}_{x\sim P}[x]$ i.e., the

goal is to estimate the probability that the coin comes up heads (the mean of a Bernoulli random variable). We first consider a very simple estimator, which just takes the empirical mean of the given data i.e.,

$$\widehat{\theta}(\overline{\mathbf{x}}) \;=\; \widehat{\theta}(x_1, \ldots, x_n) \;:=\; \frac{1}{n} \cdot \sum_{i \in [n]} x_i \,.$$

Check that the expected error of this estimator, for the loss function $\ell(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$, is $O(1/n)$.

**Exercise 2.1** *Let $P : \{0,1\} \to [0,1]$ be any distribution with $\mathbb{E}_{x \sim P}[x] = p(1) = \mu$. Show that*

$$\mathop{\mathbb{E}}_{(x_1,\ldots,x_n) \sim P^n} \left[ \left| \frac{1}{n} \cdot \sum_{i \in [n]} x_i \;-\; \mu \right|^2 \right] \;=\; O\left(\frac{1}{n}\right) .$$

We will now show that the above bound is tight. Let $\mathcal{V} = \{0,1\}$, and let $P_0 = (1/2, 1/2)$ and $P_1 = (1/2 - 2\delta, 1/2 + 2\delta)$ be the corresponding two distributions (the value of $\delta$ will be chosen later). Note that

$$|\theta(P_0) - \theta(P_1)| \;=\; 2\delta \,.$$

Using the lemma from the previous section, we get that

$$\begin{aligned}
\mathcal{M}(\Pi, \ell) \;&\geq\; \delta^2 \cdot \inf_T \{\mathbb{P}\left[T(\overline{\mathbf{x}}) \neq V\right]\} \\
&\geq\; \delta^2 \cdot \inf_T \left\{ \frac{1}{2} \cdot \mathop{\mathbb{P}}_{\overline{\mathbf{x}} \sim P_0^n} \left[T(\overline{\mathbf{x}}) = 1\right] + \mathop{\mathbb{P}}_{\overline{\mathbf{x}} \sim P_1^n} \left[T(\overline{\mathbf{x}}) = 0\right] \right\} \\
&\geq\; \delta^2 \cdot \frac{1}{2} \cdot \inf_T \{\alpha(T) + \beta(T)\} \,,
\end{aligned}$$

where $\alpha(T)$ and $\beta(T)$ are the errors as defined in the setting of binary hypothesis testing. Using the bound in terms of total-variation distance, we get that

$$\begin{aligned}
\mathcal{M}(\Pi, \ell) \;&\geq\; \frac{\delta^2}{2} \cdot \left(1 - \frac{1}{2} \cdot \|P_0^n - P_1^n\|_1\right) \\
&\geq\; \frac{\delta^2}{2} \cdot \left(1 - \frac{1}{2} \cdot \sqrt{2 \ln 2 \cdot n \cdot D(P_0 \| P_1)}\right) .
\end{aligned}$$

We use the calculation from the previous lectures that $D(P_0 \| P_1) \leq c\delta^2$ for some constant $c$. Choosing $\delta = (c \cdot 2 \ln 2 \cdot n)^{-1/2}$ gives

$$\mathcal{M}(\Pi, \ell) \;\geq\; \frac{\delta^2}{2} \left(1 - \frac{1}{2}\right) \;=\; \Omega\left(\frac{1}{n}\right) .$$

## References

[Duc16] John Duchi. Lecture notes on Information Theory and Statistics, 2016. URL: https://stanford.edu/class/stats311/. 1