Lecture 10 — October 22

Lecturer: Lester Mackey *Scribe: Bryan He, Rahul Makhijani*

Warning: *These notes may contain factual and/or typographic errors.*

10.1 Minimaxity and least favorable prior sequences

In this lecture, we will extend our tools for deriving minimax estimators. Last time, we discovered that minimax estimators can arise from Bayes estimators under least favorable priors. However, it turns out that minimax estimators may not be Bayes estimators. Consider the following example, where our old approach fails.

Example 1 (Minimax for i.i.d. Normal random variables with unknown mean θ)**.** Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2)$, with σ^2 known. Our goal is to estimate θ under squared-error loss. For our first guess, pick the natural estimator $\overline{X}$. Note that it has constant risk $\frac{\sigma^2}{n}$, which suggests minimaxity because we know that Bayes estimators with constant risk are also minimax estimators. However, $\overline{X}$ is *not* Bayes for any prior, because under squared-error loss unbiased estimators are Bayes estimators only in the degenerate situations of zero risk (TPE Theorem 4.2.3), and $\overline{X}$ is unbiased. Thus, we cannot conclude by our previous results (e.g., TPE Corollary 5.1.5) that $\overline{X}$ is minimax.

We might try to consider the wider class of estimators $\delta_{a,\mu_0}(X) = a\overline{X} + (1-a)\mu_0$ for $a \in (0,1)$ and $\mu_0 \in \mathbb{R}$, because many of the Bayes estimators we've encountered are convex combinations of a prior and a data mean. Note however that the worst case risk for these estimators is infinite:

$$\begin{aligned}\sup_\theta \mathbb{E}_\theta \left[\theta - \delta(X)\right]^2 &= \sup_\theta \{a^2 \mathrm{Var}_\theta\left(\overline{X}\right) + (1-a)^2(\theta-\mu_0)^2\} \\ &= \frac{a^2\sigma^2}{n} + (1-a)^2 \sup_\theta (\theta-\mu_0)^2 \\ &= +\infty.\end{aligned}$$

Since these estimators have poorer worst case risk than $\overline{X}$, they certainly cannot be minimax. We could keep trying to find Bayes estimators with better worst-case performance than $\overline{X}$, but we would fail: it turns out that $\overline{X}$ is in fact minimax. To establish this, we will extend our minimax results to the *limits* of Bayes estimators, rather than restricting attention to Bayes estimators only.

Definition 1 (Least Favorable Sequence of Priors)**.** Let $\{\Lambda_m\}$ be a sequence of priors with minimal average risk $r_{\Lambda_m} = \inf_\delta \int R(\theta, \delta)\, d\Lambda_m(\theta)$. Then, $\{\Lambda_m\}$ is a least favorable sequence of priors if there is a real number r such that $r_{\Lambda_m} \to r < \infty$ and $r \geq r_{\Lambda'}$ for any prior Λ'.

The reason for studying the limit of priors is that it may help us establish minimaxity. Since there need not exist a prior Λ such that the associated Bayes estimator has average risk r, this definition is less restrictive than that of a least-favorable prior. We can prove an analogue of TPE Theorem 5.1.4 in this new setting.

Theorem 1 (TPE 5.1.12)**.** Suppose there is real number r such that $\{\Lambda_m\}$ is a sequence of priors with $r_{\Lambda_m} \to r < \infty$. Let δ be any estimator such that $\sup_\theta R(\theta, \delta) = r$. Then,

1. δ is minimax,
2. $\{\Lambda_m\}$ is least-favorable.

Proof. 1. Let δ' be any other estimator. Then, for any m,

$$\sup_\theta R(\theta, \delta') \geq \int R(\theta, \delta')\, d\Lambda_m(\theta) \geq r_{\Lambda_m},$$

so that sending $m \to \infty$ yields

$$\sup_\theta R(\theta, \delta') \geq r = \sup_\theta R(\theta, \delta),$$

which means that δ is minimax.

2. Let Λ' be any prior, then

$$r_{\Lambda'} = \int R(\theta, \delta_{\Lambda'})\, d\Lambda'(\theta) \leq \int R(\theta, \delta)\, d\Lambda'(\theta) \leq \sup_\theta R(\theta, \delta) = r,$$

which means that $\{\Lambda_m\}$ is least favorable.

□

Unlike Theorem 5.1.4, this result does not guarantee uniqueness, even if the Bayes estimators δ_{Λ_m} are unique. This is because the limiting step in the proof of (1) changes any strict inequality to nonstrict inequality. However, this result allows to check much wider class of estimators, since to check that the estimator is indeed a minimax estimator we need to find only the sequence of Bayes risks convergent to maximum risk of our candidate.

Example 2 (Minimax for i.i.d. Normal random variables, continued)**.** We now have the tools to confirm our suspicion that $\overline{X}$ is minimax. By Theorem 1 above, it suffices to find a sequence $\{\Lambda_m\}$ such that $r_{\Lambda_m} \to \frac{\sigma^2}{n} =: r$. Using the conjugate prior is a good starting point, so we let $\{\Lambda_m\}$ be the conjugate priors $\{N(0, m^2)\}$ with variance tending to ∞, so that Λ_m tends to the (improper with $\pi(\theta) = 1$, $\forall \theta \in \mathbb{R}$) uniform prior on $\mathbb{R}$. By TPE Example 4.2.2, the posterior for θ associated with each Λ_m is

$$\theta \mid X_1, \ldots, X_n \sim N\left(\frac{\frac{n\overline{X}}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{m^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{m^2}} \right).$$

In particular, the posterior variance does not depend on $X_1, \ldots, X_n$, so Lemma 1 below automatically yields the Bayes risk

$$r_{\Lambda_m} = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{m^2}} \xrightarrow{m \to \infty} \frac{\sigma^2}{n} = \sup_\theta R\left(\theta, \overline{X}\right).$$

It follows from Theorem 1 that $\overline{X}$ is minimax and $\{\Lambda_m\}$ is least favorable.

Lemma 1 (TPE 5.1.13)**.** If the posterior variance $\mathrm{Var}_{\Theta|X}\left(g\left(\Theta\right) \mid X = x\right)$ is constant in x, then under squared error loss, $r_\Lambda = \mathrm{Var}_{\Theta|X}\left(g\left(\Theta\right) \mid X = x\right)$.

We know that the posterior mean minimizes Bayes risk, so this result can be obtained by plugging in the posterior mean of $g(\theta)$ into the average risk.

10.2 Minimaxity via submodel restriction

The following example illustrates the technique of deriving a minimax estimator for a general family of models by restricting attention to a subset of that family. The idea comes from simple observation that if the estimator is minimax in submodel and its risk doesn't change when we go to a larger model then estimator is minimax in this larger class.

Example 3 (Minimax for i.i.d. Normal random variables, unknown mean and variance)**.** Reconsider Example 1 in the case that the variance is unknown. That is, let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\theta, \sigma^2\right)$, with both θ and σ^2 unknown. Note that

$$\sup_{\theta,\sigma^2} R\left((\theta, \sigma^2), \overline{X}\right) = \sup_{\sigma^2} \frac{\sigma^2}{n} = \infty,$$

and in fact, the maximum risk of *any* estimator in this setting is infinite, so the question of minimaxity is uninteresting. Therefore, we restrict attention to the family parameterized by $\Omega = \{(\theta, \sigma^2) : \theta \in \mathbb{R}, \sigma^2 \leq B\}$, where B is a known constant. Assume δ is any other estimator. Calculating the risk of $\overline{X}$ within this family, we find

$$\begin{aligned}
\sup_{\theta\in\mathbb{R},\sigma^2\leq B} R\left((\theta, \sigma^2), \overline{X}\right) &= \frac{B}{n} \\
&= \sup_{\theta\in\mathbb{R},\sigma^2=B} R\left((\theta, \sigma^2), \overline{X}\right) \\
&\leq \sup_{\theta\in\mathbb{R},\sigma^2=B} R\left((\theta, \sigma^2), \delta\right) \text{ [submodel minimax]} \\
&\leq \sup_{\theta\in\mathbb{R},\sigma^2\leq B} R\left((\theta, \sigma^2), \delta\right),
\end{aligned}$$

where the first inequality follows from the fact that $\overline{X}$ is minimax for i.i.d. normals with known σ^2, and the second inequality follows from the fact that we are taking the supremum over a larger set. Hence, we are able to show that $\overline{X}$ is minimax over Ω by focusing on the case where σ^2 is known. Notice further that the form of the estimator does not depend on the upper bound B, though the bound is necessary for minimaxity to be worth investigating.

10.3 Dependence on the Loss Function

In general, minimax estimators can vary depending on the loss being considered. Below, we provide an example of minimax estimation under weighted squared error loss.

Example 4 (Minimax for binomial random variables, weighted squared error loss)**.** Let $X \sim \text{Bin}(n, \theta)$ with the loss function $L(\theta, d) = \frac{(d-\theta)^2}{\theta(1-\theta)}$. This is a simple weighted squared-error loss with the weights $w(\theta) = \frac{1}{\theta(1-\theta)}$ but it is arguably more realistic than the usual squared error in this situation because it penalizes errors near 0 and 1 more strongly than errors near $\frac{1}{2}$.

Note that for any θ, $R\left(\theta, \frac{X}{n}\right) = \frac{1}{n}$; that is, the risk is constant in θ, suggesting $\frac{X}{n}$ is minimax. We will show that this is indeed the case. We should be careful since TPE Theorem 4.2.3 is only valid under the squared-error loss. Since our loss function is different, an unbiased estimator can be Bayes. In this example, this is indeed the case.

Recall from TPE Corollary 4.1.2 that the Bayes estimator associated with the loss $L(d, \theta) = w(\theta)(d-\theta)^2$ is given by $\frac{\mathbb{E}_{\Theta|X}[\Theta w(\Theta)|X]}{\mathbb{E}_{\Theta|X}[w(\Theta)|X]}$. Invoking this result, we find that the Bayes estimator has the form

$$\delta_\Lambda(X) = \frac{\mathbb{E}_{\Theta|X}\left[\frac{1}{1-\Theta} \mid X\right]}{\mathbb{E}_{\Theta|X}\left[\frac{1}{\Theta(1-\Theta)} \mid X\right]}. \tag{10.1}$$

This is true for arbitrary priors Λ, but to calculate a closed form Bayes estimator, we use a prior conjugate to the binomial likelihood: $\Theta \sim \Lambda_{a,b} = \text{Beta}(a, b)$, for some $a, b > 0$. Suppose we observe $X = x$. If $a + x > 1$ and $b + n + x > 1$, then substituting the result of Remark 1 below into equation 10.1 proves that the estimator

$$\delta_{a,b}(x) = \frac{a + x - 1}{a + b + n - 2},$$

minimizes the posterior risk.

In particular, the estimator $\delta_{1,1}(x) = \frac{x}{n}$ minimizes the posterior risk with respect to the uniform prior after observing $0 < x < n$. If we can verify that this form remains unchanged when $x \in \{0, n\}$, then the estimator $\delta_{1,1}(X) = \frac{X}{n}$ is Bayes with constant risk, and hence minimax.

To see that this is the case, note that the posterior risk under the prior $\Lambda_{1,1}$ after observing $X = x$ and deciding $\delta(x) = d$ is

$$\int_0^1 \frac{(d-\theta)^2}{\theta(1-\theta)} \cdot \frac{\Gamma(x+1+n-x+1)}{\Gamma(x+1)\Gamma(n-x+1)} \cdot \theta^x (1-\theta)^{n-x} \, d\theta,$$

which, in the case $X = 0$, simplifies to

$$\int_0^1 \frac{(d-\theta)^2}{\theta} (1-\theta)^{n-1} \, d\theta.$$

This integral converges for $d = 0$ and diverges otherwise, so the posterior risk is minimized by choosing $\delta(0) = 0$. Similarly, in the case $X = n$, the posterior risk is minimized by choosing $\delta(n) = 1 = \frac{n}{n}$. This confirms that $\delta_{1,1}(X) = \frac{X}{n}$ minimizes the posterior risk for any outcome X, and is indeed Bayes. Since as we mentioned before this estimator has constant risk we can conclude that $\frac{X}{n}$ is minimax.

Notice that the form of the minimax estimator here depends on the type of loss being used: $\frac{X}{n}$ has constant risk for the type of weighted squared error loss considered here.

Remark 1. Recall that the Beta function can be evaluated as

$$\int_0^1 x^{k_1-1}(1-x)^{k_2-1}\,dx = \frac{\Gamma(k_1)\,\Gamma(k_2)}{\Gamma(k_1+k_2)}, \tag{10.2}$$

whenever $k_1, k_2 > 0$.

Therefore, if $Y \sim \text{Beta}(a,b)$, where $a, b > 0$, we can explicitly evaluate the expectation

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{1-Y}\right] &= \int_0^1 \frac{1}{1-y}\left[\frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} y^{a-1}(1-y)^{b-1}\right] dy \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} \int_0^1 \left[y^{a-1}(1-y)^{b-2}\right] dy \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} \cdot \frac{\Gamma(a)\,\Gamma(b-1)}{\Gamma(a+b-1)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a+b-1)} \cdot \frac{\Gamma(b-1)}{\Gamma(b)} \cdot \frac{\Gamma(a)}{\Gamma(a)} \\
&= \frac{a+b-1}{b-1},
\end{aligned}$$

where in the second step we require $b > 1$ in order to apply the relation 10.2. A similar argument yields

$$\mathbb{E}\left[\frac{1}{Y(1-Y)}\right] = \frac{(a+b-2)(a+b-1)}{(a-1)(b-1)},$$

whenever $a > 1$. Combining these identities, we have that, whenever $a, b > 1$,

$$\frac{\mathbb{E}\left[\frac{1}{1-Y}\right]}{\mathbb{E}\left[\frac{1}{Y(1-Y)}\right]} = \frac{a-1}{a+b-2}.$$

10.4 Randomized Minimax Estimators

So far, we have had little occasion to consider randomized estimators, that is, functions $\delta(X,U)$ of both the data and an independent source of randomness $U \sim \text{Unif}(0,1)$. Randomized estimators played little role in our exploration of average risk optimality, since non-randomized estimators of equal or better average risk are always available. However, they turn out to play a role when we consider the minimax criterion.

Notice that when working with convex losses, we can dispense with randomized estimators, because we can find a deterministic estimator with the same or better performance. Indeed, the data X is always sufficient, so by the Rao-Blackwell theorem, the non-random estimator $\tilde{\delta}(X) = \mathbb{E}[\delta(X,U) \mid X]$ is no worse than $\delta(X,U)$.

However, there are non-convex losses for which no deterministic minimax estimator exists, as the following example demonstrates.

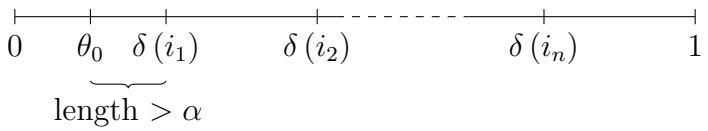

Figure 10.1. By choosing α small enough, we can ensure that any choice of $n+1$ values for the non-random estimator δ will leave some θ_0 a distance at least α away from any of the δs.

Example 5 (Randomized minimax estimator)**.** Let $X \sim \text{Bin}(n, \theta)$, where $\theta \in [0, 1]$, and consider estimation of θ under the 0-1 loss,

$$L(\theta, d) = \begin{cases} 0 & \text{if} \quad |d - \theta| < \alpha \\ 1 & \text{otherwise} \end{cases}.$$

First consider an arbitrary non-random estimator δ. Since X can take on only the $n+1$ values $\{0, 1, \ldots, n\}$, the estimator $\delta(X)$ can take on only $n+1$ values, $\{\delta(0), \delta(1), \ldots, \delta(n)\}$. If $\alpha < \frac{1}{2(n+1)}$, then we can always find θ_0 such that $|\delta(x) - \theta_0| \geq \alpha$ for every $x \in \{0, \ldots, n\}$; see Figure 10.1. Hence, $R(\theta_0, \delta(X)) = 1$ is the maximum risk of any non-random δ.

Consider instead the estimator $\delta'(X, U) = U$, which is completely random and independent of the data X. Then, for any $\theta \in [0, 1]$,

$$\begin{aligned} R(\theta, \delta') &= \mathbb{E}\left[L(\theta, \delta'(X, U))\right] \\ &= \mathbb{P}(|U - \theta| \geq \alpha) \\ &= 1 - \mathbb{P}(\theta - \alpha < U < \alpha + \theta) \\ &\leq 1 - \alpha < 1, \end{aligned}$$

and since $\alpha > 0$, the maximum risk of δ' is smaller than the maximum risk of any non-random δ. Hence, in this setting, there can be no deterministic minimax estimator.