STATS 300A: Theory of Statistics **Fall 2015**

Lecture 9 — October 20

Lecturer: Lester Mackey *Scribe: Poorna Kumar, Shengjie (Jessica) Zhang*

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

9.1 Recap

Let's take a moment to recall what we have covered so far and then give a brief preview of what we will cover in the future. We have been studying optimal point estimation. First, we showed that uniform optimality is typically not an attainable goal. Thus, we need to look for other meaningful optimality criteria. One approach is to find optimal estimators from classes of estimators possessing certain desirable properties, such as unbiasedness and equivariance. Another approach is to collapse the risk function. Examples of this approach include Bayesian estimation (minimizing the average risk) and minimax estimation (minimizing the worst-case risk). After the midterm, we will study these same principles in the context of hypothesis testing. Today, we will complete our discussion of average risk optimality and introduce the notion of minimax optimality.

Example 1. Let $X_1, ..., X_n \sim \mathcal{N}(\theta, \sigma^2)$, with $\sigma^2 > 0$ known.

From last lecture, we know that $\bar{X}$ is not a Bayes estimator for θ under squared error in the normal location model. However, $\bar{X}$ does minimize a form of average risk: it minimizes the average risk with respect to the Lebesgue measure, that is with respect to the density $\pi(\theta) = 1$ for all θ. We call this choice of π an **improper prior**, since the integral $\int \pi(\theta)d\theta = \infty$ and hence π does not define a proper probability distribution. Nevertheless, we may define a **formal posterior** for this improper prior with analogy to the definition of a proper posterior distribution:

$$\text{Formal Posterior} \propto \text{Likelihood} \times \text{Improper Prior}.$$

Quite often a formal posterior is a valid probability distribution even if the prior is improper. In the case of this normal example,

$$\text{Formal Posterior} \propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(X_i - \theta)^2\right) \times 1 \tag{9.1}$$

$$\propto \exp\left(\frac{n\bar{X}}{\sigma^2}\theta - \frac{n}{2\sigma^2}\theta^2\right) \tag{9.2}$$

which is a valid normal distribution with mean $\bar{X}$ and variance σ^2/n. Thus, $\bar{X}$ is the mean of the formal posterior which means that it minimizes the improper average risk with respect to the (improper) prior π, and hence $\bar{X}$ is a **generalized Bayes estimator**. We emphasize however that generalized Bayes estimators do not inherit many of the desirable properties of Bayes rules.

Our next example demonstrates an intimate connection between Bayes estimators in location models and the Pitman estimator.

Example 2 (Location model)**.** Suppose $X_1, ..., X_n$ are drawn i.i.d. according to a density $f(x-\theta)$ with f known, θ unknown, and prior $\pi(\theta)$. The posterior is proportional to

$$\prod_i f(x_i - \theta)\pi(\theta).$$

Therefore, the posterior mean, which is the Bayes estimator with respect to squared loss, is,

$$\frac{\int \theta \prod_i f(x_i-\theta)\pi(\theta)d\theta}{\int \prod_i f(x_i-\theta)\pi(\theta)d\theta}.$$

Note that if we choose the improper prior $\pi(\theta) = 1$, we get the Pitman estimator of location, the MRE estimator under squared error! Hence, the Pitman estimator is generalized Bayes with respect to the improper prior $\pi(\theta) = 1$, under squared error.

9.2 Admissibility of Unique Bayes Estimators

Earlier in the course we discovered that best unbiased estimators are quite often inadmissible. The story is rather different for Bayes estimators. In particular, any unique Bayes estimator is guaranteed to be admissible:

Theorem 1 (TPE 5.2.4)**.** A unique Bayes estimator (a.s. for all P_θ) is admissible.

Recall that an estimator is admissible if it is not uniformly dominated by some other estimator. That is δ is inadmissible if and only if there exists δ' such that

$$R(\theta, \delta') \le R(\theta, \delta) \text{ for all } \theta \in \Omega, \text{ and}$$
$$R(\theta, \delta') < R(\theta, \delta) \text{ for some } \theta \in \Omega.$$

Proof. Suppose δ_Λ is Bayes for Λ, and for some $\delta', R(\theta, \delta') \le R(\theta, \delta_\Lambda)$ for all $\theta \in \Omega$. If we take expectations with respect to θ, the inequality is preserved, and we get

$$\int_{\theta\in\Omega} R(\theta, \delta')d\Lambda(\theta) \le \int_{\theta\in\Omega} R(\theta, \delta_\Lambda)d\Lambda(\theta)$$

This implies that δ' is also Bayes since δ' has risk less than or equal to δ_Λ, which minimizes the average risk, and thus $\delta' = \delta_\Lambda$ with probability 1 for all P_θ. □

This theorem naturally raises the question: when is a Bayes estimator δ_Λ unique? The next result provides a set of conditions under which uniqueness is guaranteed.

Theorem 2 (TPE 4.1.4)**.** Let Q be the marginal distribution of X, that is

$$Q(E) = \int \mathbb{P}_\theta(X \in E)\, d\Lambda(\theta).$$

Then, under a strictly convex loss function, δ_Λ is unique (a.s. for all $\mathbb{P}_\theta$) if

1. $r(\Lambda, \delta_\Lambda)$ is finite, and

2. Whenever a property holds a.e. on Q, that property also holds a.e. on $\mathbb{P}_\theta$ for all $\theta \in \Omega$. In other words, $\mathbb{P}_\theta$ is **absolutely continuous** with respect to Q for all θ, sometimes written as $\mathbb{P}_\theta \ll Q$.

In this theorem, it is clear that we need the Bayes risk to be finite. Otherwise, any estimator is Bayes. Given that the risk is finite, any Bayes estimator with respect to Λ is a.s. unique under Q by the first property. Thus, the second property implies that the Bayes estimator is also unique with respect to $\mathbb{P}_\theta$.

The second property holds for most of the models we will consider and holds necessarily if Ω is open and equal to the support of Λ and $\mathbb{P}_\theta(X \in A)$ is a continuous function of θ for all measurable sets A.

9.3 Why consider Bayes estimators?

By definition, Bayes estimators are optimal with respect to the objective of minimizing average risk. In this section, we will describe several additional reasons why Bayes estimators are worthy of consideration.

9.3.1 All admissible estimators are limits of Bayes estimators

Under very weak conditions on a decision problem (see, e.g., Wald's 1949 work "Statistical Decision Functions"), *every* admissible estimator is either a Bayes estimator or a **limit of Bayes estimators**. That is, there exists a sequence of prior distributions (Λ_m) such that $\delta_{\Lambda_m}(x) \to \delta(x)$ a.e. $\mathbb{P}_\theta$ as $m \to \infty$.

A result of this flavor can be found in TPE Theorem 5.7.15. Hence, if our goal is to find an admissible estimator (and it typically is), we can safely restrict our attention to Bayes estimators and their limits.

9.3.2 Prior information

Another good reason to use Bayes estimators is that they allow us to incorporate relevant prior information and experience into our estimators.

Example 3. Suppose we are given a freshly minted coin, and we want to determine the probability that the coin will come up heads when it is flipped. Let θ denote this probability. If I have measurements of the probability that a coin flip will come up heads for 1000 coins from the same mint, I can incorporate this information into the prior distribution over θ.

9.3.3 Evaluating Bayes estimators under other criteria

A third reason to use Bayes estimators is that they offer a general method for generating reasonable estimators under various optimality criteria. By their nature, Bayes estimators are average risk optimal, but we can evaluate the quality of our Bayes estimators (just like

any other estimator) using alternative quality criteria as well. For instance, we will see soon that a search for minimax estimators often begins with Bayes estimators. In addition, Bayes estimators are often admissible. In particular, a Bayes estimator is admissible whenever any of the following conditions holds:

1. The Bayes estimator is unique.
2. The support of the prior distribution π is Ω, and either
   (a) Ω is discrete, or
   (b) π is a density with respect to the Lebesgue measure, and $R(\theta, \delta)$ is a continuous function of θ for all δ.

Note that some of our other techniques for finding estimators (such as the UMVUE framework) frequently produced inadmissible estimators.

9.4 How to choose a prior?

A primary difficulty in Bayesian decision theory, however, lies in the choice of prior. We summarize several popular strategies for prior choice below:

- *Subjective.* If prior knowledge about or experience with a model parameter is available, we can incorporate this information into the prior choice.
- *Objective.* When no prior knowledge is available, we can choose a maximally non-informative or *reference* prior (see the pioneering work of Jeffreys and the more modern work of Bernardo and Berger).
- *Family of priors.* Often, rather than choosing a single prior, we will select a family of priors based on experience, flexibility and convenience of the family. Various strategies are then available for deriving an estimator from this prior family.
  - *Hierarchical Bayes.* We can impose a *hyper prior* on the parameters of the prior distribution, which averages across the prior class. Often, the choice of hyper prior has a quantifiably smaller impact on the final Bayes estimator than the choice of a single prior from the prior family.
  - *Empirical Bayes.* We estimate the prior distribution based on the data.
  - *Robust Bayes.* We look for an estimator that performs well with respect to all priors in the prior family.

9.5 Minimax Estimators and Worst-Case Optimality

In minimax estimation, we collapse our risk function by looking at the worse-case risk. Given $X \sim \mathbb{P}_\theta$, where $\theta \in \Omega$, and a loss function $L(\theta, d)$, we want to find an estimator δ that minimizes the maximum risk:

$$\sup_{\theta \in \Omega} R(\theta, \delta).$$

Any such δ is called a **minimax** estimator.

In Bayes estimation, we essentially had a single, general-purpose way of deriving a Bayes estimator: minimize the posterior risk. The derivation of minimax estimators is often less prescriptive and more problem-specific. However, we will develop a few tools which, when they apply, will identify minimax estimators.

Perhaps surprisingly, one of the most effective ways of finding minimax estimators is to restrict our attention to Bayes estimators. To understand the connection, we will need to introduce some additional notation. First we recall the definition of the Bayes risk (or minimum average risk) under any prior distribution Λ,

$$r_\Lambda = \inf_\delta r(\Lambda, \delta) = \inf_\delta \int_{\theta\in\Omega} R(\theta, \delta) d\Lambda(\theta).$$

Definition 1. We say that a prior Λ is a **least favorable prior** if $r_\Lambda \geq r_{\Lambda'}$ for any other prior distribution Λ'. (Note that we always use the unmodified word "prior" to mean a proper prior.)

Theorem 3 (TPE 5.1.4)**.** Suppose δ_Λ is Bayes for Λ with

$$r_{\Lambda\in\Omega} = \sup_\theta R(\theta, \delta_\Lambda)$$

That is, the Bayes risk of δ_Λ is the maximum risk of δ_Λ. Then,

1. δ_Λ is minimax
2. Λ is a least favorable prior
3. If δ_Λ is the unique Bayes estimator for Λ (a.s. for all P_θ), then it is the unique minimax estimator.

Proof. If δ is any other estimator, then we have that

$$\sup_{\theta\in\Omega} R(\theta, \delta) \geq \int R(\theta, \delta) d\Lambda(\theta) \geq \int R(\theta, \delta_\Lambda) d\Lambda(\theta) = \sup_{\theta\in\Omega} R(\theta, \delta_\Lambda)$$

where the first step holds because the worst-case risk of δ is greater than (or equal to) the average risk of δ, the second step holds because δ_Λ is Bayes (and hence has an average risk no higher than that of δ), and the third step holds because of our assumption that the Bayes risk of δ_Λ is equal to the worst-case risk. This implies that δ_Λ is minimax.

If δ_Λ is the unique Bayes estimator, then the second inequality above is strict for $\delta \neq \delta_\Lambda$, which implies that δ_Λ is the unique minimax.

Let Λ' be any other prior distribution. Then, we have that

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

The first step first and second steps are by the definition of Bayes risk, and the third step holds because the worst-case risk of δ_Λ is no less than its average risk over the distribution Λ'. Since the worst-case risk of δ_Λ is its Bayes risk over Λ (by our assumption), we can infer that that Λ is a least favorable prior distribution. □

Thus, we can find a minimax estimator by finding a Bayes estimator with Bayes risk equal to its maximum risk. The following corollary highlights an important special case of this strategy.

Corollary 1 (TPE 5.1.5)**.** If a Bayes estimator δ_Λ has constant risk (that is, $R(\theta, \delta_\Lambda) = R(\theta', \delta_\Lambda)$ for all θ and θ'), then δ_Λ is minimax. Note that this is a sufficient but not necessary condition.

It is often relatively easy to check whether an estimator has constant risk, and this is typically our first line of attack for determining whether an estimator is minimax. More generally we could find a prior support set ω such that $\Lambda(\omega) = 1$ and for which $R(\theta, \delta_\Lambda)$ is maximum for all $\theta \in \omega$:

Corollary 2 (TPE 5.1.6)**.** Define

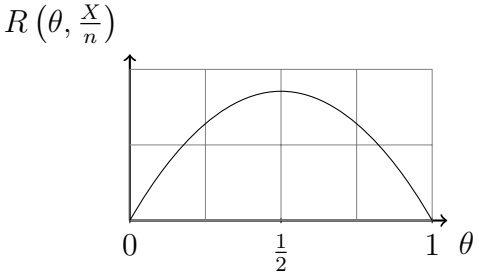
$$\omega_\Lambda = \{\theta : R(\theta, \delta_\Lambda) = \sup_{\theta'} R(\theta', \delta_\Lambda)\}$$

Then, a Bayes estimator δ_Λ is minimax if $\Lambda(\omega_\Lambda) = 1$. (TPE misstates this result as if and only if, but the only if component is false. In other words, this condition is sufficient, but not necessary.)

Example 4. Suppose $X \sim \text{Binom}(n, \theta)$ for some $\theta \in (0, 1)$ and that we use the squared error loss function. Is the sample proportion $\frac{X}{n}$ minimax? The risk of this estimator is

$$R\left(\theta, \frac{X}{n}\right) = \frac{\theta(1-\theta)}{n}.$$

The graph of $R\left(\theta, \frac{X}{n}\right)$ versus θ looks like the following:



The risk has a unique maximum at $\theta = \frac{1}{2}$, so the worst-case risk is

$$\sup_{\theta \in \Omega} R\left(\theta, \frac{X}{n}\right) = R\left(\frac{1}{2}, \frac{X}{n}\right) = \frac{1}{4n}.$$

Unfortunately, we cannot apply Corollary 5.1.6 directly because if $\Lambda(\{\frac{1}{2}\}) = 1$, then $\delta_\Lambda(X) = \frac{1}{2} \neq \frac{X}{n}$.

However, we can use the Corollary 5.1.5 to find a minimax estimator and then compare the risk of the minimax estimator with that of $\frac{X}{n}$. To find a minimax estimator, we will search for a prior such that the Bayes estimator has constant risk.

Recall the following useful fact. Under the prior distribution Beta(a, b), the Bayes estimator under the squared error loss is

$$\delta_{a,b}(X) = \frac{X + a}{n + a + b}.$$

For any a and b,

$$\begin{aligned} R(\theta, \delta_{a,b}) &= \mathbb{E}_\theta \left[\left(\frac{X + a}{n + a + b} - \theta \right)^2 \right] \\ &= \frac{1}{(n + a + b)^2} \mathbb{E}_\theta \left[(X + a - (n + a + b)\theta)^2 \right] \\ &= \frac{1}{(n + a + b)^2} \mathbb{E}_\theta \left[(X - n\theta - a(\theta - 1) - \theta b)^2 \right] \\ &= \frac{1}{(n + a + b)^2} \left(n\theta(1 - \theta) + (a(\theta - 1) + \theta b)^2 \right). \end{aligned}$$

This is a quadratic function of θ. To eliminate the θ dependence in $R(\theta, \delta_{a,b})$, we need the coefficients of the linear and quadratic terms to equal zero. The coefficient of θ^2 is

$$-n + (a + b)^2,$$

so we need $a + b = \sqrt{n}$ (since $a, b > 0$). The coefficient of θ is

$$n - 2a(a + b) = n - 2a\sqrt{n},$$

so we need $a = b = \frac{\sqrt{n}}{2}$. With these choices of a and b, the risk of $R(\theta, \delta_{a,b})$ is constant, which implies that Beta $\left(\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2} \right)$ is a least favorable prior with constant risk. Then our Bayes estimator

$$\delta_{\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}}(X) = \frac{X + \frac{\sqrt{n}}{2}}{n + \sqrt{n}},$$

is minimax with constant risk of

$$\frac{1}{4(\sqrt{n} + 1)^2}.$$

Since the worst-case risk of $\frac{X}{n}$ is $\frac{1}{4n} > \frac{1}{4(\sqrt{n}+1)^2}$, we can conclude that $\frac{X}{n}$ is not minimax.