Lecture 7 — October 13

Lecturer: Lester Mackey *Scribe: Jing Miao and Xiuyuan Lu*

7.1 Recap

So far, we have investigated various criteria for optimal inference. We first considered uniformly best risk estimation but quickly realized this can't really be done in an effective manner. We then moved on to exploring estimators under certain constraints. The first constraint we considered was unbiasedness, and we followed that with equivariance, a constraint that exploits symmetries inherent to the estimation problem being considered.

An alternative approach to optimal inference is to "collapse" the risk function rather than impose constraints. We will ultimately consider two approaches to this, the Bayesian approach (which we begin to cover in this lecture), and the minimax approach, which will be addressed later.

In the previous lectures, we have been mostly focusing on estimation. Later, we will discuss optimal inference under the context of hypothesis testing.

7.2 Risk unbiased Estimator

Recall from the last lecture the following definition.

Definition 1. An estimator δ of $g(\theta)$ is risk unbiased for a loss function $L(\theta, d)$ if for all θ and θ', $\mathbb{E}_\theta [L(\theta, \delta(x))] \leq \mathbb{E}_\theta [L(\theta', \delta(x))]$.

Intuitively, this means that the true parameter penalizes less than any false parameter. Now, we relate risk unbiasedness to MRE.

Theorem 7.1. (TPE.3.1.27) If δ is MRE for a location invariant decision problem, then δ is also risk unbiased.

Proof. We want to show that $\mathbb{E}_\theta [\rho(\delta(X) - \theta')] \geq \mathbb{E}_\theta [\rho(\delta(X) - \theta)]$ for all θ and θ'. By location equivariance, we have

$$\delta(X) - \theta = \delta(X - \theta) = \delta(U) \text{ for } U \sim f_0.$$

In other words, we must show $\mathbb{E}_0 [\rho(\delta(U))] \leq \mathbb{E}_0 [\rho(\delta(U) - (\theta' - \theta))]$ for all θ and θ', which is equivalent to showing that $\mathbb{E}_0 [\rho(\delta(U))] \leq \mathbb{E}_0 [\rho(\delta(U) - a)]$ for all $a \in \mathbb{R}$. But $\delta(U) - a$ is location equivariant if $\delta(U)$ is, which implies that δ has risk no larger than $\delta - a$ since δ is MRE. Therefore, the inequality holds for all $a \in \mathbb{R}$. □

7.3 Location-Scale Model

(TPE 3.3) Now we can extend our considerations of location invariance and equivariance to a larger class with new types of symmetries. Consider the Location-Scale Model, where $X = (X_1, ..., X_n)$ follows a joint density of the form

$$f_{\theta,\tau}(x) = \frac{1}{\tau^n} f\left(\frac{x_1 - \theta}{\tau}, \ldots, \frac{x_n - \theta}{\tau}\right),$$

where f is known but $\tau > 0$ and $\theta \in \mathbb{R}$ are unknown. Here τ is called the scale parameter and θ is called the location parameter. Let us first look at an example.

Example 1. Let $X = (X_1, \ldots, X_n)$ where $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \tau^2)$ model. The joint density is then given by

$$f_{\theta,\tau}(x) = \frac{1}{\tau^n \left(\sqrt{2\pi}\right)^n} \exp\left(-\frac{1}{2}\sum_i \left(\frac{x_i - \theta}{\tau}\right)^2\right).$$

We will also write $X \sim \text{LocScale}(\theta, \tau)$ to mean that X is distributed according to some location-scale model, with true parameters θ and τ. Note that if $X \sim \text{LocScale}(\theta, \tau)$ and $X_i' = aX_i + b$ for all i for some $a > 0, b \in \mathbb{R}$. Then

$$(X_1', ..., X_n') \sim \text{LocScare}(a\theta + b, a\tau) \sim \text{LocScale}(\theta', \tau'),$$

where $\theta' = a\theta + b, \tau' = a\tau$.

Definition 2. A model $\mathcal{P} = \{f_{\theta,\tau} : (\theta, \tau) \in \Omega\}$ is called *location-scale invariant* if

$$f_{a\theta+b,a\tau}(ax + b) = f_{\theta,\tau}(x)$$

for all $a > 0, b \in \mathbb{R}$, and $(\theta, \tau) \in \Omega$.

We now consider estimation in the context of a location-scale invariant model. Our goal for today is to estimate only the location parameter θ. We treat τ as a *nuisance* parameter, that is, a parameter which we do not know and are uninterested in estimating. We begin our considerations with a number of definitions.

Definition 3. A loss L is *location-scale invariant* for estimating θ if

$$L((\theta, \tau), d) = L((a\theta + b, a\tau), ad + b)$$

for all $a > 0$ and $b \in \mathbb{R}$. Note that any such function must be of the form

$$L((\theta, \tau), d) = \rho\left(\frac{d - \theta}{\tau}\right)$$

for some function ρ.

Definition 4. A decision problem is *location-scale invariant* for estimating θ if both the model and the loss are location-scale invariant.

Definition 5. An estimator δ of θ is *location-scale equivariant* if

$$\delta(ax + b) = a\delta(x) + b$$

for all $a > 0$ and $b \in \mathbb{R}$.

Theorem 7.2. Consider a location-scale invariant decision problem, and let δ^*_τ be a minimum risk location equivariant estimator of θ under the location invariant submodel with τ fixed. If $\delta^*_\tau = \delta^*$ is independent of τ and location-scale equivariant, then δ^* is a minimum risk location-scale equivariant estimator for the location-scale model. That is, for any location-scale equivariant estimator δ', the risk function R satisfies

$$R((\theta, \tau), \delta) \leq R((\theta, \tau), \delta')$$

for all (θ, τ).

Before proving the theorem, we provide a few examples that show when and how it may be applied.

Example 2. Consider the location-scale invariant decision problem where $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \tau^2)$ and both θ and τ are unknown. Under the loss $\frac{(\theta-d)^2}{\tau^2}$ (which we can check is location-invariant), $\overline{X}$ is the minimum risk location equivariant estimator for the location submodel for any fixed τ, and $\overline{X}$ is location-scale equivariant. Thus, by Theorem 2, $\overline{X}$ is a minimum risk location-scale equivariant estimator of θ.

Example 3. Let $X_i \overset{\text{iid}}{\sim} \text{Unif}(\theta - \tau, \theta + \tau)$. For any fixed τ, the estimator

$$\frac{X_{(1)} + X_{(n)}}{2}$$

is the minimum risk location equivariant estimator of θ under the loss $\frac{(\theta-d)^2}{\tau^2}$, and the estimator is also location-scale equivariant, so again by Theorem 2 this estimator is a minimum risk location-scale equivariant estimator as well.

Example 4. Let $X_i \overset{\text{iid}}{\sim} \text{Exp}(\theta, b)$, where both θ and b are unknown. Under the loss $\frac{(\theta-d)^2}{b^2}$, for fixed b, $X_{(1)} - \frac{b}{n}$ is the minimum risk equivariant estimator for θ. Since this depends on b, Theorem 2 cannot be applied.

Proof of Theorem 2. Assume that some location-scale equivariant δ' has strictly better risk than δ^* at some (θ_0, τ_0). Then δ' has strictly better risk at θ_0 in the location submodel with $\tau = \tau_0$ fixed. This is a contradiction since $\delta^*_{\tau_0}$ is the minimum risk equivariant estimator by hypothesis. □

For a very modern usage of equivariance, see the 2013 paper "Optimal Estimation of a Large-Dimensional Covariance Matrix Under Stein's Loss" by Ledoit and Wolf. This paper enforces a multivariate notion of equivariance, **rotation equivariance** in the setting of covariance matrix estimation:

$$\delta(\mathbf{OX}) = \mathbf{O}\delta(\mathbf{X})\mathbf{O}^\top$$

where $\mathbf{X} \in \mathbb{R}^{p\times n}$ has i.i.d. columns $\mathbf{X}_1, \ldots, \mathbf{X}_n$, $\mathbf{O} \in \mathbb{R}^{p\times p}$ is an arbitrary orthogonal matrix, and δ is an estimator for the covariance matrix of $\mathbf{X}_1$.

7.4 Bayes Estimators and Average Risk Optimality

So far we have explored optimality achievable by constraining the set of candidate decision procedures. We will next explore an alternative path to optimality: finding decision procedures which minimize a collapsed (scalar) summary of the risk function. Recall that every decision problem has the following components:

- The data $\mathbf{X}$.
- The model $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$, which is a collection of probablity distributions on the sample space.
- The Loss function L, where $L(\theta, d)$ measures the loss incurred by the decision d when the true value of the parameter is θ.
- The risk function R of a decision rule δ, where $R(\theta, \delta) = E_\theta[L(\theta, \delta)]$.

To define our first notion of collapsed optimality, **average risk optimality**, we will need to introduce a new component, measure Λ over the parameter space Ω (we assume that the parameter space has a measurable structure). Intuitively, the measure Λ should be viewed as an assignment of different importance weights to each parameter value $\theta \in \Omega$ *a priori* (that is, before the data has been observed).

Our optimality goal, given a measure Λ, is to find an estimator δ_Λ which minimizes the **average risk**,

$$r(\Lambda, \delta) = \int R(\theta, \delta) d\Lambda(\theta).$$

If Λ is a probability distribution on Ω, we call Λ the **prior** distribution. The estimator δ_Λ, if it exists, is called the **Bayes estimator** with respect to Λ, and the minimized average risk $r(\Lambda, \delta_\Lambda)$ is called the **Bayes risk**. In this Bayesian setup, we may interpret $\Theta \in \Omega$ as a random variable with distribution Λ and P_θ as the conditional distribution of X given Θ. Then the average risk $r(\Lambda)$ may be expressed as $\mathbb{E}\left[L(\Theta, \delta(X))\right]$, where the expectation is now taken jointly over (X, Θ). Using the tower property of conditional expectation, we may rewrite the average risk as

$$\begin{aligned} r(\Lambda, \delta) &= \mathbb{E}\left[L(\Theta, \delta(X))\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[L(\Theta, \delta(X))|\Theta\right]\right] \\ &= \mathbb{E}\left[R(\Theta, \delta)\right]. \end{aligned}$$

To find Bayes estimators, the big idea is that for average risk optimality, it suffices to consider the conditional risk

$$\mathbb{E}\left[L(\Theta, \delta(X)) \mid X = x\right]$$

at (almost) every value of X, where the expectation is taken with respect to the posterior distribution of $\Theta \mid X = x$.

Theorem 7.3. Suppose $\Theta \sim \Lambda$, and $X|\Theta = \theta \sim P_\theta$. If

1. there exists δ_0 an estimator of $g(\theta)$ with finite risk for all θ, and
2. there exists a value $\delta_\Lambda(x)$ that minimizes

$$\mathbb{E}\left[L(\Theta, \delta_\Lambda(X))|X = x\right] \text{ for almost every } x,$$

then δ_Λ is a Bayes estimator with respect to Λ.

The almost sure statement is with respect to the marginal (unconditional) distribution of X, where the marginal distribution is given by

$$P(X \in A) = \int P_\theta(X \in A) d\Lambda(\theta).$$

Proof. Under the assumptions of the theorem, for any other estimator δ', and for almost every x,

$$\mathbb{E}\left[L(\Theta, \delta_\Lambda(X))|X = x\right] \leq \mathbb{E}\left[L(\Theta, \delta'(X))|X = x\right].$$

Taking expectations over X, we have

$$\mathbb{E}\left[L(\Theta, \delta_\Lambda(X))\right] \leq \mathbb{E}\left[L(\Theta, \delta'(X))\right]$$

for all δ'. □

Example 5. If $L(\theta, d) = (\theta - d)^2$, we need to minimize $\mathbb{E}\left[(g(\Theta) - \delta(X))^2|X = x\right]$, and in this case, the Bayes estimator turns out to be $\delta_\Lambda(X) = \mathbb{E}\left[g(\Theta)|X\right]$, where the expectation is taken with respect to the posterior distribution of Θ given X. Here, $\mathbb{E}\left[(g(\Theta) - \delta(X))^2|X\right]$ is called the posterior risk and $\mathbb{E}\left[g(\Theta)|X\right]$ is called the posterior mean.

Example 6. Suppose that $X \sim \text{Bin}(n, \theta)$ given $\Theta = \theta$ and that Θ has prior distribution $\text{Beta}(a, b)$. The prior density is given by

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} \mathbb{I}\left(0 < \theta < 1\right)$$

The **likelihood** (model density) is given by

$$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{(n-x)}$$

The marginal density is given by

$$f(x) = \int f(x|\theta)\pi(\theta) d\theta.$$

The posterior density may be calculated using Bayes rule which states that

$$\text{posterior} = \frac{\text{joint}}{\text{marginal}} = \frac{\text{prior} \cdot \text{likelihood}}{\text{marginal}}.$$

In our notation, the posterior density is given by the formula

$$\begin{aligned}\pi(\theta|x) &= \frac{\pi(\theta)f(x|\theta)}{f(x)} \\ &= \frac{\pi(\theta)f(x|\theta)}{\int \pi(\theta')f(x|\theta')d\theta'}\end{aligned}$$

Note that the marginal component is simply a normalizing constant, a function of the likelihood and prior that does not depend on θ. Often we can avoid computing the normalizing constant and determine the posterior directly from the form of the product of likelihood and prior. Hence the following is a useful mnemonic:

$$\text{posterior} \propto \text{prior} \cdot \text{likelihood}.$$

Returning to our example,

$$\begin{aligned}\pi(\theta|x) &\propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\ &\propto \theta^{x+a-1}(1-\theta)^{n-x+b-1} \sim \text{Beta}(x+a, n-x+b).\end{aligned}$$

Hence, the Bayes estimator of θ under the squared error loss is given by

$$\mathbb{E}\left[\Theta|X=x\right] = \frac{x+a}{n+a+b}.$$

This posterior mean may be expressed as

$$\frac{X+a}{n+a+b} = \frac{n}{n+a+b}\left(\frac{X}{n}\right) + \frac{a+b}{n+a+b}\left(\frac{a}{a+b}\right).$$

Hence, the Bayes estimate is a convex combination of the sample proportion X/n (which is the UMVUE) and the prior mean $a/(a+b)$. Thus, the Bayes estimate modifies the sample estimate in light of prior information by "shrinking" the sample estimate towards the prior mean. (This is a commonly recurring property of Bayes estimators.) In addition, as the sample size n tends to infinity, the weight of the prior mean tends to zero, the empirical evidence increasingly outweighs the prior information, and the posterior mean becomes less distinguishable from the sample proportion.