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

Lecture 4 — October 1

Lecturer: Lester Mackey *Scribe: Rina Friedberg ; Weizhuang Zhou*

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

4.1 Completeness and Ancillarity

Last time we defined our ideal notion of optimal data reduction:

Definition 1. A statistic T is **complete** if no non-constant function of T is first order ancillary.

Example 1. If $X \sim p(x;\theta) \propto h(x)e^{\theta x}$, then the statistic $T(X) = X$ is complete.

Proof. Outline of steps:

(1) Suppose $\int f(x)h(x)e^{\theta x}dx = 0$ for all θ.

(2) Then we can decompose f as $f(x) = f_+(x) - f_-(x)$, with $f_+ \geq 0, f_- \geq 0$

(3) f_+ and f_- can be viewed as unnormalized densities $p_+(x)$, $p_-(x)$

(4) (1) implies that the moment generating functions of $p_+(x)$ and $p_-(x)$ are equal: $\int p_+(x)e^{\theta x}dx = \int p_-(x)e^{\theta x}dx, \forall\theta$. Hence $p_+(x)$ and $p_-(x)$ represent the same distribution, so $p_+(x) = p_-(x)$, $f_+ = f_-$, and $f \equiv 0$.

□

This ideal reduction is realized, for example, by the sufficient statistics of any full-rank exponential family.

Theorem 1 (TSH 4.3.1)**.** $(T_1, ..., T_s)$ is complete for any s-dimensional full rank exponential family.

In addition, a complete sufficient statistic is guaranteed to be independent of any ancillary statistic.

Theorem 2 (Basu's Theorem)**.** If T is complete and sufficient for $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Omega\}$, and A is ancillary then $T(X) \perp\!\!\!\perp A(X)$.

Let's look at an application of Basu's Theorem regarding the independence of sample mean and sample variance for normal model.

Example 2. $X_1, ..., X_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$ (μ, σ^2 both unknown),
Claim: $\bar{X} \perp\!\!\!\perp \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$**.**

Proof. Fix any $\sigma > 0$, and consider the submodel $\mathcal{P}_\sigma = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}\}$. In each submodel, $\bar{X}$ is complete and sufficient, and $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ is ancillary. By Basu's Theorem, $\bar{X} \perp\!\!\!\perp \sum_{i=1}^n (X_i - \bar{X})^2$ under $\mathcal{N}(\mu, \sigma^2)$ for any μ. Since σ is arbitrary, we have $\bar{X} \perp\!\!\!\perp \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ for the full model. □

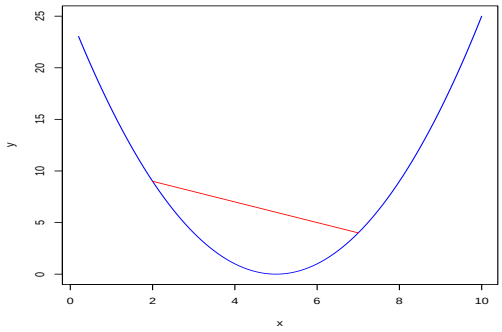

Figure 4.1. Strictly Convex

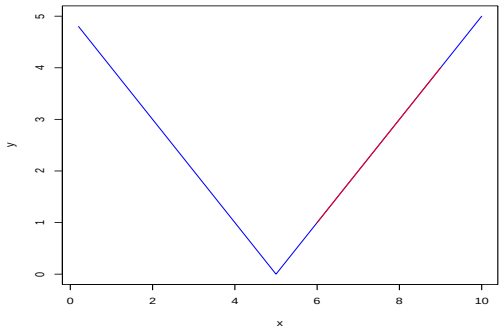

Figure 4.2. Convex but Not Strictly Convex

4.2 From Data Reduction to Risk Reduction

How does all of our work on optimal data compression translate into better decision procedures? When the loss function is convex (see Definition 2), we can develop methods to reduce risk using sufficient statistics. We focus on point estimators $g(\theta)$ for known g.

Definition 2. A function $f : C \to \mathbb{R}$ with C convex is a **convex function** if $x \neq y \in C$ and $\gamma \in (0, 1)$,

$$f(\gamma x + (1-\gamma)y) \leq \gamma f(x) + (1-\gamma)f(y).$$

f is said to be **strictly convex** if inequality above holds strictly.

Example 3. For any θ, the function $f(d) = (d-\theta)^2$ is strictly convex on $\mathbb{R}$. Figure 4.1 shows the function when $\theta = 5$. A sign of strict convexity is that the line segment joining any two points on the curve lies strictly above the curve (at all points save the end points).

Example 4. For any $\theta \in \Omega$, $f(d) = |d-\theta|$ is convex but not strictly convex. Figure 4.2 shows the function when $\theta = 5$. The line segment connecting any two points on the curve still lies on or above the curve (a sign of convexity) but not necessarily strictly above.

The following property of convex functions will have important implications for reducing risks of decision procedures.

Theorem 3 (Jensen's Inequality, K 3.25)**.** If $f : C \to \mathbb{R}$ is convex on an open set C, $\mathbb{P}(X \in C) = 1$, and $\mathbb{E}(X)$ exists, then $f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$. If f is strictly convex, the inequality holds strictly unless $X = \mathbb{E}(X)$, w.p. 1.

We will now apply this result in our decision theoretic setting. Recall that a loss function $L(\theta, d)$ is the penalty incurred by estimating d when θ is the true parameter. When $L(\theta, \cdot)$ is convex (as a function of $d \in \mathcal{D}$), we have the following result relating risk and sufficiency as a direct consequence of Jensen's inequality.

Theorem 4 (Rao-Blackwell Theorem, K 3.28)**.** Suppose that T is sufficient for $\mathbb{P} = \{\mathbb{P}_\theta, \theta \in \Omega\}$, that $\delta(X)$ is an estimator for $g(\theta)$ for which $\mathbb{E}(\delta(X))$ exists, and that $R(\theta, \delta) = \mathbb{E}_\theta L(\theta, \delta(X)) < \infty$. If $L(\theta, \cdot)$ is convex (as a function of $d \in \mathcal{D}$), then

$$R(\theta, \eta) \leq R(\theta, \delta) \quad \text{for} \quad \eta(T(X)) = \mathbb{E}(\delta(X)|T(X)).$$

If $L(\theta, \cdot)$ is strictly convex, then $R(\theta, \eta) < R(\theta, \delta)$ for any θ unless $\eta(T'(x)) = \delta$ w.p. 1.

The Rao-Blackwell Theorem is stronger than Theorem 1 of Lecture 1 because it states that when loss function is convex, we can find a *deterministic* estimator that is no worse than the original estimator using only a compression of the data, $T(X)$.

Example 5. Let $X_1, ..., X_n \overset{i.i.d}{\sim} \text{Ber}(\theta)$ for $\theta \in (0, 1)$, and consider the loss function $L(\theta, d) = (\theta - d)^2$. Suppose that we begin with the somewhat unreasonable estimator $\delta(X) = X_1$ that only makes use of the first data point. We know that $T(X) = \bar{X}$ is sufficient, so we will use the Rao-Blackwell theorem to improve our estimator δ:

$$\begin{aligned}\eta(T(X)) &= \mathbb{E}(\delta(X)|T(X)) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_1|\bar{X}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i|\bar{X}) \\ &= \mathbb{E}[\bar{X}|\bar{X}] = \bar{X}\end{aligned}$$

The second and third equality is based on the fact that our variables are identically distributed. Now we saw in the first lecture that $R(\theta, \eta) = \frac{\theta(1-\theta)}{n} < \theta(1-\theta) = R(\theta, \delta)$, indicating a strict improvement as suggested by the Rao-Blackwell Theorem.

It it important to note that while Rao-Blackwell allows us to improve upon a given estimator, it need not lead to a uniformly optimal estimator. For example, if we begin with the estimator $\delta_{\text{goofy}}(X) = \frac{1}{2}$, then $\eta(T(X)) = \mathbb{E}(\delta_{\text{goofy}}|\bar{X}) = \frac{1}{2}$ as well. Since $R(\theta, \eta) = (\frac{1}{2} - \theta)^2$, neither Rao-Blackwellized outcome (of 1/2 and $\bar{X}$) is uniformly better across all parameter values θ. In fact, we have shown in Lecture 1 that a uniformly best estimator does not exist.

So in order to obtain a meaningful notion of optimality, we need to either constrain the class of possible estimators or collapse the risk function into a scalar. We will begin by imposing a classical statistical constraint, that of unbiasedness.

4.3 Unbiased Estimation

Recall that an estimator is **unbiased** if $\mathbb{E}_\theta[\delta(X)] = g(\theta)$. Although uniformly best estimator does not exist, quite often, we can find a unbiased estimator with uniformly minimum risk, that is, an unbiased δ satisfying $R(\theta, \delta) \leq R(\theta, \delta')$, for $\forall\theta$ and any other unbiased estimators δ'. Such an estimator is called a **uniformly minimum risk unbiased estimator (UMRUE)**.

When $L(\theta, d) = (\theta - d)^2$, an UMRUE becomes a **uniformly minimum variance unbiased estimator (UMVUE).** The name follows from the fact that the mean squared error decomposes into two terms, a variance term and a squared bias term:

$$\begin{aligned}\mathbb{E}_\theta\left[(\theta - \delta(X))^2\right] &= \Big(\mathbb{E}_\theta[\delta(X)] - \theta\Big)^2 + \mathbb{E}_\theta\Big\{\Big(\delta(X) - \mathbb{E}_\theta[\delta(X)]\Big)^2\Big\} \\ &= \text{Bias}^2 + \text{Variance}.\end{aligned} \tag{4.1}$$

When the bias is constrained to be zero, minimizing risk is equivalent to minimizing variance. The following theorem demonstrates an important and fruitful connection between UMRUEs and complete sufficient statistics, when the loss is convex.

Theorem 5 (Lehmann-Scheffe Theorem)**.** If T is a complete and sufficient statistic, and $\mathbb{E}_\theta[h(T(X))] = g(\theta)$(i.e., $h(T(x))$ is unbiased for $g(\theta)$), then $h(T(X))$ is

(1) the only function of $T(X)$ that is unbiased for $g(\theta)$

(2) an UMRUE under any convex loss function,

(3) the unique UMRUE (up to a $\mathcal{P} - null$ set) under any strictly convex loss function,

(4) the unique UMVUE (up to a $\mathcal{P} - null$ set).

Proof. (1) Suppose $\mathbb{E}_\theta\left[\tilde{h}(T(X))\right] = g(\theta)$. Then $\mathbb{E}_\theta\left[\tilde{h}(T(X)) - h(T(X))\right] = 0$, $\forall\theta$. Thus $\tilde{h}(T(X)) = h(T(X))$ almost surely for each θ, by completeness.

(2) Consider any unbiased $\delta(X)$, and let $\tilde{h}(T(X)) = \mathbb{E}_\theta[\delta(X) \mid T(X)]$. Then $\mathbb{E}_\theta\left[\tilde{h}(T(X))\right] = \mathbb{E}_\theta[\delta(X)] = g(\theta)$ by the tower property of conditional expectation. By (1), $\tilde{h}(T(X)) = h(T(X))$, and by the Rao-Blackwell Theorem $R(\theta, h(T(\cdot))) = R(\theta, \tilde{h}(T(\cdot))) \leq R(\theta, \delta)$, for all θ, if the loss function is convex. Therefore, h(T(X)) is an UMRUE under any convex loss function.

(3) If the loss function is strictly convex, $R(\theta, h(T(\cdot))) < R(\theta, \delta)$ unless $\delta(X) = h(T(X))$. Thus, $h(T(X))$ is the unique UMRUE.

(4) By (3), $h(T(X))$ is the unique UMVUE.

□

Indeed, the Lehmann-Scheffe theorem gives rise to several useful strategies for finding UMRUEs under convex loss functions.

We now provide three different strategies for finding UMRUEs under convex loss.

(A) **Rao-Blackwellize / Condition**

(1) We first find a sufficient and complete statistic $T(X)$.

(2) We then find any unbiased estimator $\delta_0(X)$.

(3) We then compute $\mathbb{E}[\delta_0(X) \mid T(X)]$, which is an UMVUE by the Lehmann-Scheffe Theorem.

(B) **Solve** for the (unique) δ satisfying $\mathbb{E}_\theta[\delta(T(X))] = g(\theta)$, for all θ.

(C) **Stumble** upon some unbiased function of $T(X)$.

We now show how to use the strategies above in the following examples.

Example 6. Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \text{Ber}(\theta)$.

We know that $T(X) = \sum_{i=1}^n X_i$ is a complete and sufficient statistic from previous lectures, and furthermore the statistic $\frac{T(X)}{n}$ is an unbiased estimator for θ since

$$\mathbb{E}\left[\frac{T(X)}{n}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \theta.$$

Therefore, $\frac{T(X)}{n}$ is an UMRUE for θ under any convex loss function.

Suppose that we want to estimate θ^2 instead. If we choose $\delta(X) = \mathbb{I}(X_1 = X_2 = 1) = X_1 \cdot X_2$, then $\mathbb{E}_\theta[\delta(X)] = \theta^2$ is unbiased. We again apply Strategy A to find an UMRUE. We do this by conditioning on $T(X)$:

$$\begin{aligned}
\mathbb{E}[\delta(X) \mid T(X) = t] &= \mathbb{P}(X_1 = X_2 = 1 \mid T(X) = t) \\
&= \frac{\mathbb{P}(X_1 = X_2 = 1, \sum_{i=3}^n X_i = t-2)}{\mathbb{P}(T(X) = t)} \\
&= \frac{\theta^2 \binom{n-2}{t-2} \theta^{t-2}(1-\theta)^{n-t}\mathbb{I}(t \geq 2)}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
&= \frac{t(t-1)\mathbb{I}(t \geq 2)}{n(n-1)}
\end{aligned} \tag{4.2}$$

Notice that in this case, $\mathbb{I}(t \geq 2) = 0$ is an extraneous term, since for $t = 0$ or 1, the term $t(t-1) = 0$ already. Therefore

$$UMVUE = \frac{T(X)(T(X)-1)}{n(n-1)}\mathbb{I}(T(X) \geq 2) = \frac{T(X)(T(X)-1)}{n(n-1)}.$$

Example 7. Now suppose $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0, \theta)$. In this case, $T(X) = X_{(n)}$ is a complete and sufficient statistic, and $\delta(X) = 2X_1$ is an unbiased estimator of θ. Given knowledge of $X_{(n)}$, X_1 is equal to $X_{(n)}$ with probability $1/n$ and distributed according to $\text{Unif}(0, X_{(n)})$ with probability $1 - (1/n)$, so

$$\mathbb{P}(X_1 = x_1 | T(X)) = \frac{1}{n}\mathbb{I}(T(X) = x_1) + \frac{(1-\frac{1}{n})\mathbb{I}(0 < x_1 < T(X))}{T(X)}$$

Hence, our UMVUE is,

$$\begin{aligned}
\mathbb{E}[\delta(X) \mid T(X)] &= 2\mathbb{E}[X_1 \mid T(X)] \\
&= 2\left(\frac{1}{n}T(X) + (1-\frac{1}{n})\int_0^{T(x)} \frac{x_1}{T(X)}\mathrm{d}x_1\right) \\
&= 2(\frac{T(X)}{n} + (1-\frac{1}{n})\frac{T(X)}{2}) \\
&= \left(\frac{n+1}{n}\right)T(X).
\end{aligned} \tag{4.3}$$