Rao-Blackwell Theorem: Intuition, Lemmas and Start of Proof

Lecture #4: Tuesday, 13 January 2005
Lecturer: Prof. Charles Elkan
Scribe: Hyun Min Kang
Reviewer: Jan W. Young

1 Introduction to the Rao-Blackwell theorem

Last time we introduced the concept of sufficiency, discussing how to achieve minimal sufficiency with the Bernoulli trials example. What we were interested in at the start of the last lecture was how to find unbiased estimators with minimum variance. In this lecture, we introduce the Rao-Blackwell theorem, which is the central result that lets us find MVUEs.

Theorem 1.1 (Rao-Blackwell Theorem). Let $\{P_\theta\}$ be a family of distributions on a sample space X. Suppose $\tilde{g} : X \to \mathbb{R}$ is any unbiased estimator of $g(\theta)$. Let $t : X \to Y$ be a sufficient statistic. Define $\hat{g}(x) = E_\theta [\, \tilde{g}(x') \mid t(x') = t(x)]$. Then, the following two properties hold: $\hat{g} : X \to \mathbb{R}$ is unbiased, and $\mathrm{Var}[\hat{g}] \leq \mathrm{Var}[\tilde{g}]$.

In order to understand what the theorem says, let us describe the definition of conditional expectation. As we know, the definition of expectation is $E_\theta[f(x)] = \sum_{x \in X} P_\theta(x) f(x)$ for the case where the sample space is discrete. An *event* is any subset of the sample space X. For example, $B = \{x \in X \mid f(x) = b\}$ is an event. For any event B, the conditional expectation is defined as follows:

$$E_\theta [f(x)|B] = \sum_{x \in B} P_\theta(x|x \in B) f(x)$$

where the conditional probability $P_\theta(x|x \in B)$ is defined as follows:

$$P_\theta(x|x \in B) = \frac{P_\theta(x \text{ and } x \in B)}{P_\theta(x \in B)} = \begin{cases} 0 & \text{if } x \notin B \\ \frac{P_\theta(x)}{P_\theta(x \in B)} & \text{if } x \in B \end{cases}$$

Given the above, $\hat{g}(x)$ in the Rao-Blackwell theorem can be rewritten as follows:

$$\begin{aligned} \hat{g}(x) &= E_\theta [\tilde{g}(x') \mid t(x') = t(x)] \\ &= E_\theta [\tilde{g}(x') \mid t(x') = b] \qquad \text{for } b = t(x). \end{aligned}$$

2 Intuition behind the Rao-Blackwell theorem

The following reasoning gives an intuitive explanation why this theorem is true. $\hat{g}(x)$ is an average over all x' that have the same value for t as the observed sample x. Averaging over many x' reduces random variability, i.e. it reduces sensitivity to a particular x. But we still use all useful information in x because $t(x)$ is a sufficient statistic, and the value of $t(x)$ is fixed throughout the averaging procedure, since $t(x') = t(x)$.

The above reasoning is plausible, but it is not a proof. To prove the Rao-Blackwell theorem formally, we need two lemmas, the nested expectations lemma and Jensen's inequality lemma. These appear in the upcoming sections.

3 Nested expectations lemma

Lemma 3.1 (Nested expectations). Let $\{A\}$ be any partition of the sample space X, and let $f : X \to \mathbb{R}$ be any function. Consider $E_{\{A\}}[E[f(x)|A]]$ where the outer expectation $E_{\{A\}}[\cdot]$ averages over $\{A\}$. Then the following equation holds:

$$E\left[E\left[f(x)|A\right]\right] = E\left[f(x)\right].$$

Proof. (Discrete case only.) Let us define a function $\mathcal{A} : X \to \{A\}$ so that $\mathcal{A}(x)$ is the subset containing x in the partition $\{A\}$.

$$\begin{aligned}
E[E[f(x)|A]] &= \sum_{A\in\{A\}} P(A) \left[\sum_{x\in X} P(x|x\in A)f(x)\right] \\
&= \sum_{A\in\{A\}} P(A) \left[\sum_{x\in A} P(x|x\in A)f(x) + \sum_{x\in X-A} P(x|x\in A)f(x)\right] \\
&= \sum_{A\in\{A\}} P(A) \left[\sum_{x\in A} P(x|x\in A)f(x) + 0\right] \\
&= \sum_{A\in\{A\}} \sum_{x\in A} P(A)P(x|x\in A)f(x) \\
&= \sum_{x\in X} P(\mathcal{A}(x))P(x|x\in \mathcal{A}(x))f(x) \\
&= \sum_{x\in X} P(\mathcal{A}(x))\frac{P(x)}{P(x\in\mathcal{A}(x))}f(x) \\
&= \sum_{x\in X} P(x)f(x) \\
&= E[f(x)].
\end{aligned}$$

Note that the nested sum $\sum_{A\in\{A\}}\sum_{x\in A}$ above can be rewritten to $\sum_{x\in X}$ because $\{A\}$ is a partition of X.

4 Jensen's inequality

Definition 1 (Convex function). A function $c : \mathbb{R} \to \mathbb{R}$ is *convex* if

$$\lambda c(x) + (1-\lambda)c(y) \geq c(\lambda x + (1-\lambda)y)$$

for all $x, y \in \mathbb{R}$ and all $0 \leq \lambda \leq 1$.

Lemma 4.1 (Jensen's inequality). Let $u : X \to \mathbb{R}$ be an estimator, and let $c : \mathbb{R} \to \mathbb{R}$ be a convex function. Then the following property always holds:

$$E\left[c(u)\right] \geq c\left(E[u]\right).$$

Informally, the average of $c(u)$ is always equal or greater than c of the average of u.

Proof. For the discrete case where u has only n different values that have nonzero probability, we prove this lemma by induction.

(base case) If $n = 1$, it is obvious that $E[c(u)] = c(E[u])$ since there is only a single value of $u(x)$. Thus, the lemma holds for the base case.

(inductive case) Assume that the lemma holds for $n = 1, \cdots, k-1$ regardless of u. We want to prove the lemma holds for $n = k$.

Let $U = \{u_1, \cdots, u_{k-1}\}$ be the $k-1$ values of $u(x)$ with nonzero probability. The inductive hypothesis says that $E[c(u)] \geq c(E[u])$ for any probability distribution such that $\sum_{u \in U} P(u(x) = u) = 1$.

Even if $u(x)$ has more than $k-1$ values, $\sum_{u \in U} P(u(x) = u | u \in U) = 1$ is always true. Consequently, the inductive hypothesis also implies that $E[c(u)|u \in U] \geq c(E[u|u \in U])$ is true where $u(x)$ may take k different values, i.e. $u \in U \cup \{u_k\}$. We want to prove $E[c(u)] \geq c(E[u])$ for such u.

Let $p = P(u = u_k)$ and let $z = E[u|u \in U]$. Then $E[u] = pu_k + (1-p)z$ holds, and $E[c(u)]$ can be rewritten as follows:

$$\begin{aligned}
E[c(u)] &= pc(u_k) + (1-p)E[c(u)|u \in U] \\
&\geq pc(u_k) + (1-p)c(z) \qquad \text{(by inductive hypothesis)} \\
&\geq c(pu_k + (1-p)z) \qquad \text{(by definition of convexity)} \\
&= c\left(E[u]\right).
\end{aligned}$$

According to the base case and inductive case, $E[c(u)] \geq c(E[u])$ holds for any u that has discrete values with nonzero probability.

We can extend Jensen's inequality for conditional expectations also as the following proposition.

Lemma 4.2 (Jensen's inequality for conditional expectations). Let $u : X \to \mathbb{R}$ be any estimator. If $c : \mathbb{R} \to \mathbb{R}$ is convex, then for any $A \subseteq X$,

$$E[c(u(x))|x \in A] \geq c(E[u(x)|x \in A]).$$

5 Proof of Rao-Blackwell theorem (first part)

Recall that, in the statement of the Rao-Blackwell theorem, $t : X \to Y$ is a sufficient statistic for a family of probability distributions $\{P_\theta\}$. Define $\hat{g}(x) = E_\theta\left[\tilde{g}(x')|t(x') = t(x)\right]$. We must prove the following three claims.

1. $\hat{g}(x)$ really is an estimator, i.e. it really is a function of x only, regardless of θ.
2. $E_\theta\left[\hat{g}(x)\right] = g(\theta)$ for every θ, i.e. $\hat{g}(x)$ is an unbiased estimator.
3. $\mathrm{Var}[\hat{g}] \leq \mathrm{Var}[\tilde{g}]$.

In this section, we only prove the first claim.

We know that $t : X \to Y$ is sufficient for $\{P_\theta\}$. This means that for any a such that $a = t(x)$ for a particular x, $P(x'|t(x') = a)$ does not depend on θ. Consider the conditional expectation

$$E[f(x')|t(x') = a] = \sum_{x' \text{ s.t. } t(x')=a} f(x')P(x'|t(x') = a).$$

If $f(x')$ does not depend on θ, then the conditional expectation does not either. Let $f(x') = \tilde{g}(x')$ which is not a function of θ. It follows that

$$\hat{g}(x) = E[\tilde{g}(x')|t(x') = t(x)]$$

is a function of x only, not of θ or anything else.