Statistics 210A Lecture 7 Notes

Daniel Raban

September 16, 2021

1 Computing UMVU Estimators and Lower Bounds for Unbiased Estimation

1.1 Computing UMVU estimators

Last time, we proved **Jensen's inequality** for convex f:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

The **Rao-Blackwell theorem** told us that if $L(\theta; d)$ is convex in d, $\delta(X)$ is an estimator, and $T(X)$ is sufficient, then $\mathbb{E}[\delta \mid T]$ is better than δ. We also saw that if $T(X)$ is complete sufficient and $g(\theta)$ is U-estimable, there is a unique unbiased estimator of the form $\delta(T)$. It is UMVU (dominates all other unbiased estimators for any convex L). We saw that there were 2 ways to find UMVU estimators:

1. Directly find an unbiased $\delta(T)$.
2. Rao-Blackwellize any unbiased $\delta(X)$.

Example 1.1. If $X_1, \dots, X_n \overset{\text{iid}}{\sim} U[0, \theta]$, then $X_{(n)}$ is complete sufficient for estimating θ. We saw that $\frac{n+1}{n} X_{(n)}$ is UMVU. However, Keener shows that among estimators of the form $cX_{(n)}$, $\frac{n+2}{n+1} X_{(n)}$ actually has the best MSE.

Example 1.2. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Pois}(\theta)$ with $\theta > 0$ and pmf

$$p_\theta^{(1)}(x) = \frac{\theta^x e^{-\theta}}{x!}, \quad x = 0, 1, \dots.$$

Then $T(X) = \sum_i X_i \sim \text{Pois}(n\theta)$ is complete sufficient with pmf

$$p_\theta^T(t) = \frac{(n\theta)^t e^{-n\theta}}{t!}.$$

Let's estimate θ^2 with an unbiased estimator. First, we'll use Method 1: $\overline{X}^2$ is not unbiased because $\mathbb{E}[\overline{X}] = \theta$, so $\mathbb{E}[\overline{X}^2] > \theta^2$ by Jensen's inequality. Observe that

$$\begin{aligned}\delta(T) \text{ is unbiased} &\iff \sum_{t=0}^{\infty} \delta(t) p_\theta^T(t) = \theta^2 \quad \forall \theta > 0 \\ &\iff \sum_{t=0}^{\infty} \delta(t) \frac{n^t \theta^t}{t!} = \theta^2 e^{n\theta} \quad \forall \theta > 0.\end{aligned}$$

Write $\theta^2 e^{n\theta} = \sum_{k=0}^{\infty} \frac{n^k \theta^{k+2}}{k!} = \sum_{j=2}^{\infty} \frac{n^{j-2}}{(j-2)!} \theta^j$. So we get $\delta(0) = \delta(1) = 0$, and for $t \geq 2$, $\delta(t) = \frac{n^{t-2}}{(t-2)!} \cdot \frac{t!}{n^t} = \frac{t(t-1)}{n^2}$. We can write this more compactly as

$$\delta(t) = \frac{t(t-1)}{n^2}, \qquad t = 0, 1, \ldots$$

Now we use Method 2, Rao-Blackwellization: We know that $\mathbb{E}_\theta[X_1 X_2] = (\mathbb{E}_\theta[X_1])^2 = \theta^2$, so we want to condition $X_1 X_2$ on $T = \sum_i X_i$. Since $X \mid T = t \sim \text{Multinomial}(t, 1/n 1_n)$, we can check that $X_1 \mid T = t \sim \text{Binom}(t, 1/n)$ and $X_2 \mid X_1 = x_1, T = t \sim \text{Binom}(t - x_1, 1/(n-1))$. So we can compute

$$\mathbb{E}\left[X_1 X_2 \mid \sum_i X_i\right] = \delta(T),$$

as before.

1.2 Differential identities for the score function

Assume that $\mathcal{P}$ has densities p_θ with respect to μ with $\Theta \subseteq \mathbb{R}^d$. Suppose there is a **common support** $\{x : p_\theta(x) > 0\}$ which is the same for all θ. We have the log-likelihood $\ell(\theta; x) = \log p_\theta(x)$.

Definition 1.1. Define the **score function** to be $\nabla \ell(\theta; x)$.

We have

$$p_{\theta+\eta}(x) = e^{\ell(\theta+\eta;x)} \approx p_\theta(x) e^{\eta^\top \nabla \ell(\theta,x)}$$

for small η. So we can think of this as locally looking like an exponential family withe the score function looking like a complete sufficient statistic.

We have differential identities, similar to in an exponential family. Start with

$$1 = \int_{\mathcal{X}} e^{\ell(\theta,x)} \, d\mu(x)$$

Taking $\frac{\partial}{\partial \theta_j}$ on both sides, we get

$$0 = \int_{\mathcal{X}} \frac{\partial}{\partial_j} \ell(\theta; x) e^{\ell(\theta;x)} \, d\mu(x).$$

This gives the identity

$$\mathbb{E}_\theta[\nabla \ell(\theta; X)] = 0.$$

It is important that we are integrating using the same θ that we plug into the score function.

If we differentiate again with respect to θ_k, we get

$$0 = \int_{\mathcal{X}} \left(\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k} + \frac{\partial \ell}{\partial \theta_j} \frac{\partial \ell}{\partial \theta_k} \right) = \mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta_j \partial \theta_k} \ell(\theta; X) \right] + \mathbb{E} \left[\frac{\partial \ell}{\partial \theta_j}(\theta; X) \frac{\partial \ell}{\partial \theta_k}(\theta; X), \right]$$

which gives the identity

$$J(\theta) := \mathbb{E}_\theta[-\nabla^2 \ell(\theta; X)] = \mathrm{Var}_\theta(\nabla \ell(\theta; X)).$$

The quantity $J(\theta)$ is called the **Fisher information**.

1.3 The Cramér-Rao lower bound

Let's relate this back to a statistic $\delta(X)$. Suppose

$$g(\theta) = \mathbb{E}_\theta[\delta(X)] = \int_{\mathcal{X}} \delta(x) e^{\ell(\theta;x)} \, d\mu(x).$$

Then

$$\begin{aligned} \nabla g(\theta) &= \int \delta \nabla \ell(\theta) e^{\ell} \, d\mu \\ &= \mathbb{E}_\theta[\delta(X) \nabla \ell(\theta; X) \\ &= \mathrm{Cov}_\theta(\delta(X); \nabla \ell(\theta; X).] \end{aligned}$$

If we have only one parameter, so $\theta \in \mathbb{R}$, then Cauchy-Schwarz gives

$$\mathrm{Var}_\theta(\delta) \, \mathrm{Var}(\dot{\ell}(\theta; X)) \geq \mathrm{Cov}_\theta(\delta, \dot{\ell}(\theta))^2.$$

So we get

Theorem 1.1 (Cramér-Rao)**.** *Let $\delta(X)$ be an unbiased estimator for $g(\theta)$. If $\theta \in \mathbb{R}$,*

$$\mathrm{Var}_\theta(\delta(X)) \geq \frac{g'(\theta)^2}{J(\theta)}.$$

More generally, if $\theta \in \mathbb{R}^d$ and $g(\theta) \in \mathbb{R}$,

$$\mathrm{Var}_\theta(\delta) \geq \nabla g(\theta)^\top J(\theta)^{-1} \nabla g(\theta).$$

Remark 1.1. This technically holds for any estimator δ with $\mathbb{E}_\theta[\delta(X)] = g(\theta)$. We are just interpreting it as $g(\theta)$ coming first and δ being unbiased for $g(\theta)$.

Example 1.3 (iid sample)**.** Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} p_\theta^{(1)}(x)$ with $\theta \in \Theta$, so $X \sim p_\theta(x) = \prod_i p_\theta^{(1)}(x_i)$. Writing $\ell_1(\theta; x_i) = \log p_\theta^{(1)}(x_i)$, we have

$$\ell(\theta; x) = \sum_i \ell_1(\theta, x_i).$$

Then

$$\begin{aligned} J(\theta) &= \operatorname{Var}_\theta(\nabla \ell(\theta; X)) \\ &= n \operatorname{Var}_\theta(\nabla \ell_1(\theta; X_i)) \\ &= n J_1(\theta), \end{aligned}$$

where $J_1(\theta)$ is the Fisher information in a single observation. So Fisher information scales linearly. This means that the Cramér-Rao lower bound scales like $1/n$.

1.4 The Hammersley-Chapman-Robbins inequality

The Cramér-Rao lower bound requires differentiation under the integral. The Hammersley-Chapman-Robbins inequality gives a more general bound using finite differences. The idea is that

$$\frac{p_{\theta+\varepsilon}(x)}{p_\theta(x)} - 1 = e^{\ell(\theta+\varepsilon;x) - \ell(\theta;x)} - 1 \approx \varepsilon^\top \nabla \ell(\theta; x)$$

for small ε. So in the limit, we will get a similar bound to Cramér-Rao.

Theorem 1.2 (Hammersley-Chapman-Robbins)**.** *Let δ be unbiased for $g(\theta)$, and assume that for some collection of ε, $p_\varepsilon \ll p$. Then*

$$\operatorname{Var}_\theta(\delta) \geq \sup_\varepsilon \frac{g(\theta+\varepsilon) - g(\theta)}{\mathbb{E}_\theta\left[\left(\frac{p_{\theta+\varepsilon}(X)}{p_\theta(X)} - 1\right)^2\right]}.$$

Proof. Observe that

$$\begin{aligned} \mathbb{E}_\theta\left[\frac{p_{\theta+\varepsilon}(x)}{p_\theta(x)} - 1\right] &= \int \left(\frac{p_{\theta+\varepsilon}}{p_\theta} - 1\right) p_\theta \, d\mu \\ &= \int (p_{\theta+\varepsilon} - p_\theta) \, d\mu = 0, \end{aligned}$$

as long as $p_{\theta+\varepsilon} \ll p_\theta$. Furthermore,

$$\operatorname{Cov}\left(\delta(X), \frac{p_{\theta+\varepsilon}(X)}{p_\theta(X)} - 1\right) = \int \delta \left(\frac{p_{\theta+\varepsilon}}{p_\theta} - 1\right) p_\theta \, d\mu$$

$$
\begin{aligned}
&= \int \delta p_{\theta+\varepsilon}\, d\mu - \int \delta p_\theta \, d\mu \\
&= \mathbb{E}_{\theta+\varepsilon}[\delta(X)] - \mathbb{E}_\theta[\delta(X)] \\
&= g(\theta+\varepsilon) - g(\theta).
\end{aligned}
$$

Using Cauchy-Schwarz, we get

$$
\mathrm{Var}_\theta(\delta) \cdot \mathbb{E}_\theta\left[\left(\frac{p_{\theta+\varepsilon}(X)}{p_\theta(X)} - 1\right)^2\right] \geq g(\theta+\varepsilon) - g(\theta).
$$

So we get

$$
\mathrm{Var}_\theta(\delta) \geq \frac{g(\theta+\varepsilon) - g(\theta)}{\mathbb{E}_\theta\left[\left(\frac{p_{\theta+\varepsilon}(X)}{p_\theta(X)} - 1\right)^2\right]}.
$$

This lower bound holds for every ε, so we can take the sup over ε on the right hand side. □

Remark 1.2. If we let $\varepsilon \to 0$, we get the Cramér-Rao lower bound, but taking the sup over ε gives a better bound.

1.5 Efficiency

The Cramér-Rao lower bound is not always achievable.

Definition 1.2. The **efficiency** is

$$
\mathrm{eff}_\theta(\delta) = \frac{\mathrm{CRLB}}{\mathrm{Var}_\theta(\delta)} \leq 1.
$$

We say that $\delta(X)$ is **efficient** if $\mathrm{eff}_\theta(\delta) = 1$ for all θ.

Note that

$$
\mathrm{eff}_\theta(\delta) = \mathrm{Corr}_\theta(\delta(X), \ell'(\theta; X))^2
$$

Example 1.4. For exponential families,

$$
p_\eta(x) = e^{\eta^\top T(x) - A(\eta)} h(x), \qquad \ell(\eta; x) = \eta^\top T(x) - A(\eta) + \log h(x).
$$

So the score is

$$
\nabla \ell(\eta; x) = T(x) - \mathbb{E}_\eta[T(X)].
$$

This tells us that the Fisher information is

$$
\begin{aligned}
\mathrm{Var}_\eta(\nabla \ell(\eta; X)) &= \mathrm{Var}_\eta(T(X)) \\
&= \nabla^2 A(\eta) \\
&= \mathbb{E}_\eta[-\nabla^2 A(\eta)]
\end{aligned}
$$

Example 1.5. Consider a curved exponential family with $\theta \in \mathbb{R}$:

$$p_\theta(x) = e^{\eta(\theta)^\top T(x) - B(\theta)} h(x).$$

Then the log-likelihood is

$$\ell(\theta; x) = \eta(\theta)^\top T(x) - B(\theta) - \log h(x),$$

so the chain rule gives the score as

$$\frac{d}{d\theta}\ell(\theta; x) = \dot{\eta}(\theta)^\top T(x) - \dot{B}(\theta)$$

Note that $\frac{d}{d\theta}B(\theta) = \frac{d}{d\theta}A((\eta(\theta)) = \sum_{j=1}^n \dot{\eta}(\theta)\frac{\partial}{\partial \eta_j}A(\eta) = \dot{\eta}(\theta)^\top(\nabla A(\eta))$.

$$= \dot{\eta}(\theta)^\top(T(x) - \mathbb{E}_\eta[T(X)])$$