Presentation Supplement: Proofs of the Selected Theorems

Volkan Cevher
Georgia Institute of Technology
Atlanta, GA
cevher@ieee.org

I. The Factorization Theorem

The factorization theorem is introduced at Slide 15. The proof of this theorem is done for the case in which Γ is discrete and is due to [1]. A general proof can be found in [2].

Let $p_\theta(\mathbf{y}|t)$ denote the density of $\mathbf{y}$ given $t = T(\mathbf{y})$. By the Bayes formula one have

$$\begin{aligned} p_\theta(\mathbf{y}|t) &\triangleq P_\theta(\mathbf{Y}=\mathbf{y}|T(\mathbf{Y})=t) \\ &= \frac{P_\theta(T(\mathbf{Y})=t|\mathbf{Y}=\mathbf{y})P_\theta(\mathbf{Y}=\mathbf{y})}{P_\theta(T(\mathbf{Y})=t)} \end{aligned} \tag{I.1}$$

Since $P_\theta(T(\mathbf{Y}) = t|\mathbf{Y} = \mathbf{y}) = 1$ if $T(\mathbf{Y}) = t$ and 0 if $T(\mathbf{Y}) \neq t$, and $P_\theta(\mathbf{Y} = \mathbf{y}) = p_\theta(\mathbf{y})$, Eq.I.1 becomes

$$p_\theta(\mathbf{y}|t) = \begin{cases} p_\theta(\mathbf{y})/P_\theta(T(\mathbf{Y})=t) & \text{if } T(\mathbf{y})=t, \\ 0 & \text{otherwise.} \end{cases} \tag{I.2}$$

Now $P_\theta(T(\mathbf{Y}) = t) = \sum_{\mathbf{y}|T(\mathbf{Y})=t} p_\theta(\mathbf{y})$. To prove the if part of the theorem observe the following

$$\begin{aligned} P_\theta(T(\mathbf{Y})=t) &= \sum_{\mathbf{y}|T(\mathbf{Y})=t} g_\theta[T(\mathbf{y})]h(\mathbf{y}) \\ &= g_\theta(t)\sum_{\mathbf{y}|T(\mathbf{Y})=t} h(\mathbf{y}) \end{aligned} \tag{I.3}$$

in addition one also have $p_\theta(\mathbf{y}) = g_\theta[T(\mathbf{y})]h(\mathbf{y}) = g_\theta(t)h(\mathbf{y})$. From Eq. I.2 one then have

$$p_\theta(\mathbf{y}|t) = \begin{cases} h(\mathbf{y})/\sum_{\mathbf{y}|T(\mathbf{Y})=t} h(\mathbf{y}) & \text{if } T(\mathbf{y})=t, \\ 0 & \text{otherwise.} \end{cases} \tag{I.4}$$

Since the right hand side of Eq. I.4 does not depend on θ, T is a sufficient statistic for the parameter set $\theta \in \mathbf{\Lambda}$.

To prove the only if statement in the theorem, let T be any sufficient statistic for θ. From Eq. I.2 one can write

$$p_\theta(\mathbf{y}) = p_\theta(\mathbf{y}|T(\mathbf{y}))P_\theta[T(\mathbf{Y})=T(\mathbf{y})] \tag{I.5}$$

Since T is sufficient for θ, $p_\theta(\mathbf{y}|T(\mathbf{y}))$ depends only on $\mathbf{y}$ and not on θ. On defining $h(\mathbf{y}) \triangleq p_\theta(\mathbf{y}|T(\mathbf{y}))$ and $g_\theta[T(\mathbf{y})] \triangleq P_\theta[T(\mathbf{Y}) = T(\mathbf{y})]$, one can see that Eq. I.5 implies the factorization theorem. Hence, the proof is complete.

II. The Rao-Blackwell Theorem

Slide 17 presents the Rao-Blackwell theorem, which is very useful for minimum variance unbiased estimators. The theorem and its proof can also be found in [1].

To prove that $\tilde{g}[T(\mathbf{Y})]$ is unbiased, take the expectation

$$\begin{aligned} E_\theta\{\tilde{g}[T(\mathbf{Y})]\} &= E_\theta\{E_\theta\{\hat{g}(\mathbf{Y})|T(\mathbf{Y})\}\} \\ \Rightarrow \tilde{g}[T(\mathbf{Y})] &= E_\theta\{\hat{g}(\mathbf{Y})\} = g(\theta) \end{aligned} \tag{II.1}$$

First note that the expextation defining $\tilde{g}$ does not depend on θ due to the sufficiency of T. Secondly, the second equality can be obtained by using the fact that $E\{E\{X|Z\}\} = E\{X\}$ and the unbiasedness of $\hat{g}$.

In order to see that $Var_\theta(\tilde{g}[T(\mathbf{Y})]) \leq Var_\theta(\hat{g}(\mathbf{Y}))$, note the following

$$\begin{aligned} Var_\theta(\tilde{g}[T(\mathbf{Y})]) &= E_\theta\{[\tilde{g}[T(\mathbf{Y})]]^2\} - g^2(\theta) \\ Var_\theta(\hat{g}(\mathbf{Y})) &= E_\theta\{[\hat{g}(\mathbf{Y})]^2\} - g^2(\theta) \end{aligned} \tag{II.2}$$

Hence, if it can be shown that $E_\theta\{[\tilde{g}[T(\mathbf{Y})]]^2\} \leq E_\theta\{[\hat{g}(\mathbf{Y})]^2\}$, the proof is complete.

$$\begin{aligned} E_\theta\{[\tilde{g}[T(\mathbf{Y})]]^2\} &= E_\theta\{[E_\theta\{\hat{g}(\mathbf{Y})|T(\mathbf{Y})\}]^2\} \\ &\leq E_\theta\{E_\theta\{[\hat{g}(\mathbf{Y})]^2|T(\mathbf{Y})\}\} \\ &= E_\theta\{[\hat{g}(\mathbf{Y})]^2\}, \end{aligned} \tag{II.3}$$

The second equality follows from Jensen's inequality [1] and the final equality follows from iterated expectation operations. The equality in Jensen's inequality is satisfied if and only if $P_\theta[\hat{g}(\mathbf{Y}) = E_\theta\{\hat{g}(\mathbf{Y})|T(\mathbf{Y})\}|T(\mathbf{Y})] = 1$, and using the definition of $\tilde{g}$ it is easy to see that this condition is equivalent to $P_\theta[\hat{g}(\mathbf{Y}) = \tilde{g}[T(\mathbf{Y})]] = 1$. This completes the proof of the Rao-Blackwell theorem.

III. Cramer-Rao Bound

The Cramer-Rao bound establishes a lower bound on the error covariance matrix for any unbiased estimator $\hat{\theta}$ for a parameter θ and was introduced in Slide 39. To set up the Cramer-Rao bound, we need to define a function called the score function, interpret it, and establish its statistical properties. The proof here follows the one in chapter 6 of [3].

The score function is defined to be the gradient of the log-likelihood function:

[1] *Jensen's Inequality*: For any random variable X and convex function C, $E\{C(X)\} \geq C(E\{X\})$ with equality if and only if $P(X = E\{X\}) = 1$ when C is strictly convex.

$$s(\theta,\mathbf{y}) = \frac{\partial}{\partial\theta}L(\theta,\mathbf{y}) = \frac{\partial}{\partial\theta}\log p_\theta(\mathbf{y}) \tag{III.1}$$

When the realization $\mathbf{y}$ is replaced by the random variable $\mathbf{Y}$, then the log-likelihood and score functions become random variables:

$$s(\theta,\mathbf{Y}) = \frac{\partial}{\partial\theta}L(\theta,\mathbf{Y}) = \frac{\partial}{\partial\theta}\log p_\theta(\mathbf{Y}) \tag{III.2}$$

The score function scores values of θ as the random vector $\mathbf{Y}$ assumes values from the distribution $p_\theta(\mathbf{y})$. Scores are good scores and scores different from zero are bad scores. The score function has zero mean:

$$\begin{aligned} E\{s(\theta,\mathbf{y})\} &= E\{\frac{\partial}{\partial\theta}\log p_\theta(\mathbf{Y})\} \\ &= \int d\mathbf{y} p_\theta(\mathbf{y})\frac{\partial}{\partial\theta}\log p_\theta(\mathbf{y}) \\ &= \int d\mathbf{y}\frac{\partial}{\partial\theta}\log p_\theta(\mathbf{y}) = \frac{\partial}{\partial\theta}\int d\mathbf{y} p_\theta(\mathbf{y}) = \mathbf{0} \end{aligned} \tag{III.3}$$

The covariance matrix of the score function $s(\theta,\mathbf{Y})$ is called the *Fisher information matrix* and is denoted by $\mathbf{J}(\theta)$:

$$\mathbf{J}(\theta) = E\{s(\theta,\mathbf{Y})s^T(\theta,\mathbf{Y})\} = E\{\frac{\partial}{\partial\theta}\log p_\theta(\mathbf{Y})(\frac{\partial}{\partial\theta}\log p_\theta(\mathbf{Y}))^T\}. \tag{III.4}$$

This result for the Fisher information matrix can be cast in a different, but equivalent, form by noting that the function $\frac{\partial}{\partial\theta}\log p_\theta(\mathbf{y})$ may be rewritten as

$$\frac{\partial}{\partial\theta}\log p_\theta(\mathbf{y}) = \frac{1}{p_\theta(\mathbf{y})}\frac{\partial}{\partial\theta}p_\theta(\mathbf{y}). \tag{III.5}$$

The second gradient of $\log p_\theta(\mathbf{y})$ may then be rewritten as

$$\frac{\partial}{\partial\theta}(\frac{\partial}{\partial\theta}\log p_\theta(\mathbf{y}))^T = \frac{\frac{\partial}{\partial\theta}(\frac{\partial}{\partial\theta}p_\theta(\mathbf{y}))^T}{p_\theta(\mathbf{y})} - \frac{\partial}{\partial\theta}\log p_\theta(\mathbf{y})(\frac{\partial}{\partial\theta}p_\theta(\mathbf{y}))^T \tag{III.6}$$

The expectation of the first term on the right-hand side is zero, so

$$E\{\frac{\partial}{\partial\theta}(\frac{\partial}{\partial\theta}\log p_\theta(\mathbf{Y}))^T\} = -E\{\frac{\partial}{\partial\theta}\log p_\theta(\mathbf{Y})(\frac{\partial}{\partial\theta}p_\theta(\mathbf{Y}))^T\}. \tag{III.7}$$

This identity produces formula for the Fisher information matrix:

$$\mathbf{J}(\theta) = -E\{\frac{\partial}{\partial\theta}(\frac{\partial}{\partial\theta}\log p_\theta(\mathbf{Y}))^T\}. \tag{III.8}$$

These results are summarized by recording the i,j element of the Fisher information matrix:

$$\begin{aligned} [\mathbf{J}(\theta)]_{i,j} &= E\{\frac{\partial}{\partial\theta_i}\log p_\theta(\mathbf{Y})(\frac{\partial}{\partial\theta_j}\log p_\theta(\mathbf{Y}))^T\} \\ &= E\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta(\mathbf{Y})\} \end{aligned} \tag{III.9}$$

There is one more property we will need. The cross-covariance between the score function and the error of any unbiased estimator $\hat\theta$ is identity:

$$E\{s(\theta,\mathbf{Y})[\hat\theta-\theta]^T\} = \mathbf{I} \tag{III.10}$$

To establish this remarkable property, we note that the unbiasedness of $\hat\theta$ implies $E\{[\hat\theta-\theta]^T\} = \mathbf{0}^T$. This may be written as $\int d\mathbf{y}p_\theta(\mathbf{y})[\hat\theta-\theta]^T = \mathbf{0}^T$. Taking the gradient with respect to θ, one can obtain:

$$\begin{aligned} \int d\mathbf{y}\frac{\partial}{\partial\theta}p_\theta(\mathbf{y})[\hat\theta-\theta]^T - \int d\mathbf{y}p_\theta(\mathbf{y})\mathbf{I} = \mathbf{0} \\ \int d\mathbf{y}p_\theta(\mathbf{y})\frac{\partial}{\partial\theta}\log p_\theta(\mathbf{y})[\hat\theta-\theta]^T = \mathbf{I} \\ E\{s(\theta,\mathbf{Y})[\hat\theta-\theta]^T\} = \mathbf{I}. \end{aligned} \tag{III.11}$$

Then the error covariance matrix for $\hat\theta$ is bounded as follows:

$$\mathbf{C} = E\{[\hat\theta-\theta][\hat\theta-\theta]^T\} \geq \mathbf{J}^{-1}, \tag{III.12}$$

provided that $\mathbf{J}$ is positive definite. That is, the matrix $\mathbf{C}-\mathbf{J}^{-1}$ is nonnegative definite, as is the matrix $\mathbf{J}-\mathbf{C}^{-1}$.

Form the following $2m\times 1$ vector:

$$\begin{bmatrix} \hat\theta-\theta \\ s(\theta,\mathbf{Y}) \end{bmatrix} \tag{III.13}$$

This vector has zero mean. Its covariance matrix is given by

$$\begin{aligned} \mathbf{Q} &= E\{\begin{bmatrix} \hat\theta-\theta \\ s(\theta,\mathbf{Y}) \end{bmatrix}[(\hat\theta-\theta)^T s^T(\theta,\mathbf{Y})]\} \\ &= \begin{bmatrix} \mathbf{C} & \mathbf{I} \\ \mathbf{I} & \mathbf{J} \end{bmatrix} \end{aligned} \tag{III.14}$$

The nonnegative definite covariance matrix $\mathbf{Q}$ may be diagonalized as follows:

$$\begin{bmatrix} \mathbf{I} & -\mathbf{J}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}\begin{bmatrix} \mathbf{C} & \mathbf{I} \\ \mathbf{I} & \mathbf{J} \end{bmatrix}\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{J}^{-1} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \mathbf{C}-\mathbf{J}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{J} \end{bmatrix} \tag{III.15}$$

Thus, the covariance matrix $\mathbf{Q}$ is similar to the matrix on the right-hand side. Therefore, $\mathbf{C}-\mathbf{J}^{-1}$ is nonnegative definite, meaning $\mathbf{C}\geq\mathbf{J}^{-1}$ or $\mathbf{J}\geq\mathbf{C}^{-1}$. The i,i element of $\mathbf{C}$ is the mean-squared error of the estimator of θ_i:

$$C_{i,i} = E\{(\hat\theta_i-\theta_i)^2\} \geq (\mathbf{J}^{-1})_{i,i}. \tag{III.16}$$

So, the i,i element of the inverse of the Fisher information matrix lower bounds the mean-squared error of any unbiased estimator of θ_i.

References

[1] Poor, H. V. (1994), *An Introduction to Signal Detection and Estimation* (Dowen& Culver, Inc.)

[2] Lehmann, E. L. (1986), *Testing Statistical Hypotheses* (Wiley: New York)

[3] Scharf, L. S. (1991), *Statistical Signal Processing* (Addison-Wesley Publishing Company, Inc.)