

## Section 6: Sufficiency, Rao-Blackwell, NEF

Ricky Truong (rickytruong@college.harvard.edu),  
Emily Xing (exing@college.harvard.edu)

### 1 Introduction

#### 1.1 Logistics

Welcome! The goal of our weekly section is to review content from lecture, with emphasis on big-picture intuition, mathematical proof, and hands-on practice.

Specifically, we aim for our sections to be interactive, well-paced, inclusive, and fun! Please do not hesitate to reach out if you have any questions/feedback!

#### 1.2 Office Hours

- Mondays, 7:30 - 9:30 PM in Adams D-Hall (Ricky).
- Fridays, 11 AM - 12 PM in Maxwell-Dworkin 2nd Floor (Emily).
- Saturdays, 10:30 - 11:30 AM in Cabot D-Hall (Emily).

### 2 Big Picture

This week, we introduce three deeply connected ideas: *sufficient statistics*, *Rao-Blackwellization*, and the *Natural Exponential Family (NEF)*.

The unifying theme? *Data compression*: given  $n$  observations  $Y_1, \dots, Y_n$ , can we incorporate all of the relevant information about  $\theta$  into a lower-dimensional summary  $T(\vec{Y})$  without any loss? *Sufficient statistics* answer this question. Informally,  $T$  is sufficient for  $\theta$  if, once you know  $T$ , looking at the full data  $\vec{Y}$  tells you nothing new about  $\theta$ !

*Rao-Blackwell theorem* shows why this matters for estimation: any estimator that *ignores* the sufficient statistic can always be improved (in MSE, mean squared error) by conditioning on it. This gives us a principled recipe for upgrading a “naive” estimator into a better one.

Finally, the *Natural Exponential Family* is a very important class of well-known distributions (including Bernoulli, Poisson, Normal, Exponential, and more) that share a common formula structure. This structure makes it easy to (i) identify a sufficient statistic, (ii) find the MLE, and (iii) compute moments—all at once!

### 3 Sufficient Statistics

**Idea.** A statistic  $T(\vec{Y})$  “*suffices*” for an estimator  $\theta$  if it captures everything the data can tell us about  $\theta$ . Once we know  $T$ , the remaining randomness in  $\vec{Y}$  is irrelevant to finding  $\theta$ . Formally:

**Definition 1** (Sufficient statistic). For  $Y_1, \dots, Y_n$  from a parametric model  $F_{Y;\theta}$ , a statistic  $T(\vec{Y})$  is **sufficient** for  $\theta$  if the conditional distribution of  $(Y_1, \dots, Y_n) \mid T$  does not depend on  $\theta$ .

**Example 1** (Bernoulli trials). Let  $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(p)$  and  $T = \sum_{i=1}^n Y_i$ . Is  $T$  sufficient for  $p$ ?

For any  $\vec{y}$  with  $\sum y_i = t$ ,

$$P(\vec{Y} = \vec{y} \mid T = t) = \frac{P(\vec{Y} = \vec{y}, T = t)}{P(T = t)} = \frac{p^t(1-p)^{n-t}}{\binom{n}{t} p^t(1-p)^{n-t}} = \frac{1}{\binom{n}{t}},$$

which does not depend on  $p$ . Thus  $T = \sum_{i=1}^n Y_i$  is sufficient. Intuitively, all that matters is the *total* number of successes, not which specific trials succeeded.

- Intuitively: Knowing  $T$  renders the ordering and identity of individual observations irrelevant for learning about  $\theta$ .
- The full data  $\vec{Y}$  is always sufficient, but this is trivial. We seek the most efficient sufficient statistic possible.
- Sufficient statistics are not unique: If  $T$  is sufficient, then so is any one-to-one function  $g(T)$ .
- The *minimal* sufficient statistic achieves the greatest simplification.

### 3.1 The Factorization Criterion

Checking sufficiency from the definition—by computing conditional distributions—can be really tedious. The *Factorization Criterion* gives a much faster alternative!

**Theorem 1** (Factorization Criterion).  $T(\vec{Y})$  is sufficient for  $\theta$  if and only if the joint density (or PMF) factors as

$$f_{\vec{Y}}(\vec{y}; \theta) = g_{\theta}(T(\vec{y})) \cdot h(\vec{y}),$$

where  $g_{\theta}$  may depend on  $\theta$  (but only through  $T(\vec{y})$ ), and  $h$  does not depend on  $\theta$  at all.

- **Strategy:** Write down the joint density/PMF. Then, try to factor it into a piece that only depends on  $(\theta, T(\vec{y}))$  and a leftover piece free of  $\theta$ .
- $\otimes$ : The factorization only needs to hold up to multiplicative constants that are free of  $\theta$ . Constant factors can always be absorbed into  $h$ .

**Example 2** (Poisson). Let  $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Pois}(\lambda)$ . The joint PMF is

$$f_{\vec{Y}}(\vec{y}; \lambda) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} = e^{-n\lambda} \lambda^{\sum_{i=1}^n y_i} \cdot \prod_{i=1}^n \frac{1}{y_i!}.$$

Set  $g_{\lambda}(t) = e^{-n\lambda} \lambda^t$  (with  $t = \sum y_i$ ) and  $h(\vec{y}) = \prod_{i=1}^n \frac{1}{y_i!}$ . Since  $h$  is free of  $\lambda$ ,  $T = \sum_{i=1}^n Y_i$  (equivalently,  $\vec{Y}$ ) is sufficient for  $\lambda$ .

**Concept Checker 1.** For each model below, use the Factorization Criterion to identify a sufficient statistic for the named parameter.

1.  $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Expo}(\lambda)$ , sufficient statistic for  $\lambda$ .

2.  $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$  with  $\sigma^2$  **known**, sufficient statistic for  $\mu$ .
3.  $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(0, \theta)$ , sufficient statistic for  $\theta$ .

### Solution

1. The joint density is

$$f_{\vec{Y}}(\vec{y}; \lambda) = \prod_{i=1}^n \lambda e^{-\lambda y_i} = \lambda^n e^{-\lambda \sum_{i=1}^n y_i}.$$

Set  $g_\lambda(t) = \lambda^n e^{-\lambda t}$  (with  $t = \sum y_i$ ) and  $h(\vec{y}) = 1$ . Thus  $T = \sum_{i=1}^n Y_i$  (equivalently,  $\bar{Y}$ ) is sufficient for  $\lambda$ .

2. The joint density is

$$f_{\vec{Y}}(\vec{y}; \mu) = \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp\left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right).$$

Expanding the exponent:  $\sum (y_i - \mu)^2 = \sum y_i^2 - 2\mu \sum y_i + n\mu^2$ . The factor  $e^{-\frac{1}{2\sigma^2}(\sum y_i^2)}$  does not involve  $\mu$ , so set

$$g_\mu(t) = \exp\left( \frac{\mu t}{\sigma^2} - \frac{n\mu^2}{2\sigma^2} \right), \quad h(\vec{y}) \propto \exp\left( -\frac{\sum y_i^2}{2\sigma^2} \right),$$

with  $t = \sum y_i$ . Thus  $T = \sum_{i=1}^n Y_i$  (equivalently,  $\bar{Y}$ ) is sufficient for  $\mu$ .

3. The joint density is

$$f_{\vec{Y}}(\vec{y}; \theta) = \frac{1}{\theta^n} \mathbb{1}\{y_{(n)} \leq \theta\},$$

since all  $y_i \in [0, \theta]$  iff  $y_{(n)} \leq \theta$ . Set  $g_\theta(t) = \frac{1}{\theta^n} \mathbb{1}\{t \leq \theta\}$  (with  $t = y_{(n)}$ ) and  $h(\vec{y}) = 1$ . Thus  $T = Y_{(n)}$  is sufficient for  $\theta$ .

## 4 Rao-Blackwell

**Idea.** Got an unbiased estimator, but it doesn't use all the data we have? Rao-Blackwell says: *Condition on the sufficient statistic*. At worst, you'll end up with the same estimator, At best, in fact, you'll almost always make things strictly better (lower MSE)!

**Theorem 2** (Rao-Blackwell). *Let  $\hat{\theta}$  be any estimator of  $\theta$ , and let  $T$  be a sufficient statistic for  $\theta$ . Define the Rao-Blackwellized estimator*

$$\hat{\theta}_{RB} = \mathbb{E}[\hat{\theta} \mid T].$$

Then:

- (i)  $\hat{\theta}_{RB}$  has the **same bias** as  $\hat{\theta}$  (in particular, unbiasedness is preserved).
- (ii)  $\hat{\theta}_{RB}$  has **lower or equal variance** than  $\hat{\theta}$ .

(iii) Therefore,  $MSE(\hat{\theta}_{RB}) \leq MSE(\hat{\theta})$ , with equality iff  $\hat{\theta}$  is already a function of  $T$ .

*Proof.* (i) **Bias.** By Adam's Law:  $\mathbb{E}[\hat{\theta}_{RB}] = \mathbb{E}[\mathbb{E}[\hat{\theta} | T]] = \mathbb{E}[\hat{\theta}]$ . So  $\text{Bias}(\hat{\theta}_{RB}) = \text{Bias}(\hat{\theta})$ .

(ii) **Variance.** By Eve's Law:

$$\text{Var}(\hat{\theta}) = \underbrace{\mathbb{E}[\text{Var}(\hat{\theta} | T)]}_{\geq 0} + \text{Var}(\mathbb{E}[\hat{\theta} | T]) = \mathbb{E}[\text{Var}(\hat{\theta} | T)] + \text{Var}(\hat{\theta}_{RB}) \geq \text{Var}(\hat{\theta}_{RB}).$$

Equality holds iff  $\text{Var}(\hat{\theta} | T) = 0$  a.s., meaning  $\hat{\theta}$  is already a deterministic function of  $T$ .  $\square$

• **Why does this work?** The sufficient statistic captures all the information about  $\theta$  in the data. If your estimator doesn't fully exploit  $T$ , it contains residual randomness that is uninformative about  $\theta$ —and that noise that increases variance. Conditioning on  $T$  averages out this uninformative variation.

• **Intuition:** If  $\hat{\theta}$  is a function of  $T$ , then we can take out what's known. E.g., let  $\hat{\theta} = \bar{Y}$  and  $T = \sum_{i=1}^n Y_i$ . Thus,  $\hat{\theta}_{RB} = \mathbb{E}[\bar{Y} | \sum_{i=1}^n Y_i] = \bar{Y} \mathbb{E}[1 | \sum_{i=1}^n Y_i] = \bar{Y}$ , the original estimator!

•  $\otimes$ :  $\hat{\theta}_{RB} = \mathbb{E}[\hat{\theta} | T]$  is a function of  $T$  only (since  $T$  is sufficient, the conditional expectation cannot depend on  $\theta$ ). This is what makes it a valid statistic.

**Example 3** (Rao-Blackwell for Poisson). Each page of a book has a  $\text{Pois}(\lambda)$  number of typos, independently. We observe  $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Pois}(\lambda)$  and wish to estimate  $\theta = P(Y_1 = 0) = e^{-\lambda}$ . A naive (but unbiased) estimator uses only the first observation:

$$\hat{\theta} = \mathbf{1}\{Y_1 = 0\}.$$

This is unbiased ( $\mathbb{E}[\hat{\theta}] = e^{-\lambda}$ ), but it throws away  $Y_2, \dots, Y_n$ . We know  $T = \sum_{i=1}^n Y_i$  is sufficient. By the Chicken-Egg story (or direct Bayes' rule calculation),  $Y_1 | T = t \sim \text{Bin}(t, \frac{1}{n})$ , so

$$\hat{\theta}_{RB} = \mathbb{E}[\hat{\theta} | T] = P(Y_1 = 0 | T) = \left(1 - \frac{1}{n}\right)^T = \left(1 - \frac{1}{n}\right)^{n\bar{Y}}.$$

For large  $n$ ,  $(1 - \frac{1}{n})^{n\bar{Y}} \approx e^{-\bar{Y}} = \hat{\theta}_{\text{MLE}}$ , so the Rao-Blackwellized estimator is essentially the MLE. Both are better than the naive estimator.

**Concept Checker 2.** Let  $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(p)$ . Consider the estimator  $\hat{p} = Y_1$  (using only the first observation).

1. Show  $\hat{p} = Y_1$  is unbiased for  $p$ .
2. What is the sufficient statistic  $T$ ?
3. Compute  $\hat{p}_{RB} = \mathbb{E}[Y_1 | T]$ .
4. Does this result make intuitive sense?

**Solution**

1.  $\mathbb{E}[Y_1] = p$ , so  $\hat{p} = Y_1$  is unbiased.
2. From the factorization criterion (or our earlier example),  $T = \sum_{i=1}^n Y_i \sim \text{Bin}(n, p)$  is sufficient for  $p$ .
3. We need  $\mathbb{E}[Y_1 | T = t]$ . By symmetry,  $\mathbb{E}[Y_i | T = t] = \mathbb{E}[Y_j | T = t]$  for all  $i, j$ . Since  $\sum_{i=1}^n Y_i = t$ , summing over  $i$ :

$$\sum_{i=1}^n \mathbb{E}[Y_i | T = t] = t \implies n \cdot \mathbb{E}[Y_1 | T = t] = t \implies \mathbb{E}[Y_1 | T = t] = \frac{t}{n}.$$

Therefore  $\hat{p}_{\text{RB}} = \mathbb{E}[Y_1 | T] = \frac{T}{n} = \bar{Y}$ .

4. Yes! The Rao-Blackwellized estimator is simply the sample mean  $\bar{Y}$ , which is the MLE (and intuitively the most natural estimator). Rao-Blackwell automatically recovers the “obvious” estimator from a wasteful one.

**Concept Checker 3.** Can the MLE be improved by Rao-Blackwellization?

**Solution**

Since the MLE is always a function of a sufficient statistic, it is not improved by Rao-Blackwellization.

More formally, let  $\hat{\theta}_{\text{MLE}}$  be the MLE for some model  $f_{\vec{y}}(\vec{y}; \theta)$ , and suppose  $T(\vec{Y})$  is a sufficient statistic for  $\theta$ . By the Factorization Criterion,  $f_{\vec{y}}(\vec{y}; \theta) = g(T(\vec{y}), \theta)h(\vec{y})$ , but since  $h(\vec{y})$  is a multiplicative constant,  $\mathcal{L}(\theta, \vec{y}) = g(T(\vec{y}), \theta)$ . Thus,  $\hat{\theta}_{\text{MLE}} = \arg \max_{\theta \in \Theta} (\mathcal{L}(\theta, \vec{y})) = \arg \max_{\theta \in \Theta} (g(T(\vec{y}), \theta))$ , so  $\hat{\theta}_{\text{MLE}}$  is always a function of a sufficient statistic.

## 5 Natural Exponential Family (NEF)

**Idea.** Many of the most important distributions in statistics share a common algebraic structure in PMF/PDF. This structure—the *Natural Exponential Family*—makes it straightforward to identify sufficient statistics, derive MLEs, and compute means and variances from the formula.

**Definition 2** (Natural Exponential Family). A distribution belongs to the **Natural Exponential Family (NEF)** if its density (or PMF) can be written as

$$f_Y(y; \theta) = e^{\theta y - \psi(\theta)} h(y),$$

where:

- $\theta$  is the **natural parameter** (a reparameterization of the original parameter),
  - $\psi(\theta)$  is the **log-partition function** (a normalizing term that ensures the density integrates to 1),
  - $h(y) \geq 0$  is a **base measure** that does not depend on  $\theta$ .
- **Strategy:** Rewrite the density  $f_Y(y)$  as  $\exp(\log(f_Y(y)))$ .

## 5.1 Properties of NEF

**Theorem 3** (Mean and Variance via  $\psi$ ). *If  $Y \sim NEF(\theta)$  with log-partition function  $\psi$ , then*

$$\mathbb{E}[Y] = \psi'(\theta), \quad \text{Var}(Y) = \psi''(\theta).$$

• **Intuition:**  $\psi(\theta)$  encodes all of our distributional information. Its first derivative gives the mean, its second derivative gives the variance. This is a remarkable fact—we always have our moments once you identify  $\psi$ .

**Theorem 4** (Sufficient statistic and MLE in NEF). *Let  $Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} NEF(\theta)$ . Then:*

(i) *The sample mean  $\bar{Y}$  is a sufficient statistic for  $\theta$ .*

(ii) *The MLE of  $\mu = \mathbb{E}[Y] = \psi'(\theta)$  satisfies  $\hat{\mu}_{MLE} = \bar{Y}$ .*

*Proof.* (i) The joint density is

$$f_{\bar{Y}}(\bar{y}; \theta) = \prod_{i=1}^n e^{\theta y_i - \psi(\theta)} h(y_i) = e^{\theta \sum_{i=1}^n y_i - n\psi(\theta)} \prod_{i=1}^n h(y_i) = e^{n(\theta \bar{y} - \psi(\theta))} \cdot \prod_{i=1}^n h(y_i).$$

The first factor depends on the data only through  $\bar{y}$ , and the second factor ( $h^n$ ) does not involve  $\theta$ . By the Factorization Criterion,  $\bar{Y}$  is sufficient.

(ii) The log-likelihood is  $\ell(\theta; \bar{y}) = n(\theta \bar{y} - \psi(\theta))$ . Setting  $\ell'(\theta) = n(\bar{y} - \psi'(\theta)) = 0$  gives  $\psi'(\hat{\theta}_{MLE}) = \bar{y}$ . Since  $\mu = \psi'(\theta)$ , this means  $\hat{\mu}_{MLE} = \bar{y}$ .  $\square$

## 5.2 Common NEF Distributions

**Example 4** (Bernoulli as NEF). Let  $Y \sim \text{Bern}(p)$ . We have

$$f_Y(y; p) = p^y(1-p)^{1-y} = \exp\left(y \log \frac{p}{1-p} - \log \frac{1}{1-p}\right) \cdot 1.$$

Setting  $\theta = \log \frac{p}{1-p}$  (the log-odds, or logit of  $p$ ) and  $\psi(\theta) = \log(1 + e^\theta)$ , we see  $\text{Bern}(p)$  is NEF. We can verify:

$$\psi'(\theta) = \frac{e^\theta}{1 + e^\theta} = p = \mathbb{E}[Y], \quad \psi''(\theta) = p(1-p) = \text{Var}(Y). \checkmark$$

**Example 5** (Poisson as NEF). Let  $Y \sim \text{Pois}(\lambda)$ . We have

$$f_Y(y; \lambda) = \frac{e^{-\lambda} \lambda^y}{y!} = \exp(y \log \lambda - \lambda) \cdot \frac{1}{y!}.$$

Setting  $\theta = \log \lambda$  and  $\psi(\theta) = e^\theta = \lambda$ , we see  $\text{Pois}(\lambda)$  is NEF. Check:

$$\psi'(\theta) = e^\theta = \lambda = \mathbb{E}[Y], \quad \psi''(\theta) = e^\theta = \lambda = \text{Var}(Y). \checkmark$$

**Concept Checker 4.** Let  $Y \sim \mathcal{N}(\mu, \sigma^2)$  with  $\sigma^2$  known.

1. Show that the Normal distribution is in the NEF by identifying  $\theta$ ,  $\psi(\theta)$ , and  $h(y)$ .
2. Use the NEF properties to verify  $\mathbb{E}[Y] = \mu$  and  $\text{Var}(Y) = \sigma^2$ .

## Solution

1. The density of  $Y \sim \mathcal{N}(\mu, \sigma^2)$  is

$$f_Y(y; \mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) = \exp\left(\frac{\mu y}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) \cdot \underbrace{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}}}_{h(y)}.$$

This is NEF with natural parameter  $\theta = \frac{\mu}{\sigma^2}$  and log-partition function  $\psi(\theta) = \frac{\mu^2}{2\sigma^2} = \frac{\theta^2 \sigma^2}{2}$ .

2. With  $\psi(\theta) = \frac{\sigma^2 \theta^2}{2}$ :

$$\psi'(\theta) = \sigma^2 \theta = \sigma^2 \cdot \frac{\mu}{\sigma^2} = \mu = \mathbb{E}[Y]. \checkmark$$

$$\psi''(\theta) = \sigma^2 = \text{Var}(Y). \checkmark$$

Distribution	$\theta$	$\psi(\theta)$	$\mathbb{E}[Y] = \psi'(\theta)$	$\text{Var}(Y) = \psi''(\theta)$	Suff. stat.
Bern( $p$ )	$\log \frac{p}{1-p}$	$\log(1 + e^\theta)$	$p$	$p(1-p)$	$\sum Y_i$
Bin( $m, p$ ), $m$ known	$\log \frac{p}{1-p}$	$m \log(1 + e^\theta)$	$mp$	$mp(1-p)$	$\sum Y_i$
Pois( $\lambda$ )	$\log \lambda$	$e^\theta$	$\lambda$	$\lambda$	$\sum Y_i$
$\mathcal{N}(\mu, \sigma^2)$ , $\sigma^2$ known	$\frac{\mu}{\sigma^2}$	$\frac{\sigma^2 \theta^2}{2}$	$\mu$	$\sigma^2$	$\bar{Y}$
Expo( $\lambda$ )	$-\lambda$	$-\log(-\theta)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\sum Y_i$
Gamma( $\alpha, \lambda$ ), $\alpha$ known	$-\lambda$	$-\alpha \log(-\theta)$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$	$\sum Y_i$

Table 1: Common NEF distributions, where  $f_Y(y; \theta) = e^{\theta y - \psi(\theta)} h(y)$ . The  $\mathbb{E}[Y]$  and  $\text{Var}(Y)$  columns are expressed in terms of the original parameter for readability. Sufficient statistics assume  $Y_1, \dots, Y_n$  i.i.d.

## 6 Practice Problems

**Problem 1.** Let  $Y$  be a random variable from a distribution belonging to the NEF (i.e.,  $f_Y(y; \theta) = e^{\theta y - \psi(\theta)} h(y)$ , with  $\theta$  unknown). Assume regularity conditions hold. Show  $\mathbb{E}[Y] = \psi'(\theta)$  and  $\text{Var}[Y] = \psi''(\theta)$ .

*Hint:* For  $\theta = \theta^*$  under regularity conditions,  $\mathbb{E}[s(\theta; \vec{Y})] = 0$  and  $\text{Var}[s(\theta; \vec{Y})] = -\mathbb{E}[s'(\theta; \vec{Y})]$ .

## Solution

First, let's show  $\mathbb{E}[Y] = \psi'(\theta)$ .

$$\begin{aligned}
\mathcal{L}(\theta; Y) &= f_Y(Y; \theta) && \text{by definition of likelihood} \\
&= e^{\theta Y - \psi(\theta)} h(Y) && \text{since } Y \text{ is NEF} \\
&= e^{\theta Y - \psi(\theta)} && \text{since } h(Y) \text{ is a mult. constant} \\
\ell(\theta; Y) &= \theta Y - \psi(\theta) && \text{by definition of log-likelihood} \\
s(\theta; Y) &= Y - \psi'(\theta) && \text{by definition of score} \\
\mathbb{E}[s(\theta; Y)] &= \mathbb{E}[Y] - \psi'(\theta) && \text{by taking expectation} \\
0 &= \mathbb{E}[Y] - \psi'(\theta) && \text{since } \mathbb{E}[s(\theta; \vec{Y})] = 0 \\
\mathbb{E}[Y] &= \psi'(\theta) && \text{by rearranging}
\end{aligned}$$

Next, let's show  $\text{Var}[Y] = \psi''(\theta)$ . Notice  $s(\theta; Y) = Y - \psi'(\theta) \implies \text{Var}[s(\theta; Y)] = \text{Var}[Y]$  since  $\psi'(\theta)$  is not random.

$$\begin{aligned}
s(\theta; Y) &= Y - \psi'(\theta) && \text{from before} \\
s'(\theta; Y) &= -\psi''(\theta) && \text{by taking partial derivative} \\
\mathbb{E}[s'(\theta; Y)] &= -\psi''(\theta) && \text{by taking expectation} \\
-\mathbb{E}[s'(\theta; Y)] &= \psi''(\theta) && \text{by multiplying by } -1 \\
\text{Var}[s(\theta; \vec{Y})] &= \psi''(\theta) && \text{since } \text{Var}[s(\theta; \vec{Y})] = -\mathbb{E}[s'(\theta; \vec{Y})] \\
\text{Var}[Y] &= \psi''(\theta) && \text{since } \text{Var}[s(\theta; Y)] = \text{Var}[Y]
\end{aligned}$$

**Problem 2.** Let  $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Gamma}(\alpha, \lambda)$  with density  $f_Y(y; \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y}$  for  $y > 0$ , where  $\alpha > 0$  is **known** and  $\lambda > 0$  is unknown.

- Show the Gamma distribution is in the NEF (with  $\alpha$  known), and identify  $\theta$ ,  $\psi(\theta)$ , and  $h(y)$ .
- Use the NEF properties to find  $\mathbb{E}[Y]$  and  $\text{Var}(Y)$ .
- Identify a sufficient statistic for  $\lambda$  and find  $\hat{\lambda}_{\text{MLE}}$ .

### Solution

(a) We can write

$$f_Y(y; \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y} = \exp((-\lambda)y - (-\alpha \log \lambda + \log \Gamma(\alpha))) \cdot y^{\alpha-1}.$$

Setting  $\theta = -\lambda$  (so  $\lambda = -\theta$  and  $\theta < 0$ ),  $\psi(\theta) = -\alpha \log(-\theta) + \log \Gamma(\alpha)$ , and  $h(y) = y^{\alpha-1}$ , we see this is NEF.

(b) Differentiating  $\psi(\theta) = -\alpha \log(-\theta) + \log \Gamma(\alpha)$ :

$$\psi'(\theta) = \frac{-\alpha}{-\theta} \cdot (-1) = \frac{\alpha}{-\theta} = \frac{\alpha}{\lambda} = \mathbb{E}[Y].$$

$$\psi''(\theta) = \frac{\alpha}{\theta^2} = \frac{\alpha}{\lambda^2} = \text{Var}(Y).$$

(c) By the NEF theorem,  $\bar{Y}$  is sufficient for  $\lambda$ . The MLE satisfies  $\hat{\mu}_{\text{MLE}} = \bar{Y}$ , and since  $\mu = \mathbb{E}[Y] = \frac{\alpha}{\lambda}$ , we have  $\frac{\alpha}{\hat{\lambda}_{\text{MLE}}} = \bar{Y}$ , so  $\hat{\lambda}_{\text{MLE}} = \frac{\alpha}{\bar{Y}}$ .

**Problem 3.** Let  $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(p)$  with  $p$  unknown.

- (a) Consider the estimator  $\hat{p} = \frac{Y_1 + Y_2}{2}$  (using only two observations). Is this unbiased? Find its variance.
- (b) Apply Rao-Blackwell to find  $\hat{p}_{\text{RB}} = \mathbb{E}\left[\frac{Y_1 + Y_2}{2} \mid T\right]$  where  $T = \sum_{i=1}^n Y_i$ .
- (c) Compute  $\text{Var}(\hat{p}_{\text{RB}})$  and verify  $\text{Var}(\hat{p}_{\text{RB}}) \leq \text{Var}(\hat{p})$ .
- (d) What is  $\hat{p}_{\text{RB}}$  when  $n = 2$ ? Does this make sense?

### Solution

(a)  $\mathbb{E}\left[\frac{Y_1 + Y_2}{2}\right] = \frac{p+p}{2} = p$ , so it is unbiased. Since  $Y_1, Y_2 \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(p)$ ,

$$\text{Var}\left(\frac{Y_1 + Y_2}{2}\right) = \frac{1}{4} (\text{Var}(Y_1) + \text{Var}(Y_2)) = \frac{p(1-p)}{2}.$$

(b) By linearity of conditional expectation,

$$\hat{p}_{\text{RB}} = \mathbb{E}\left[\frac{Y_1 + Y_2}{2} \mid T\right] = \frac{1}{2} (\mathbb{E}[Y_1 \mid T] + \mathbb{E}[Y_2 \mid T]).$$

By the same symmetry argument as in the Concept Checker,  $\mathbb{E}[Y_i \mid T] = \frac{T}{n}$  for each  $i$ . So

$$\hat{p}_{\text{RB}} = \frac{1}{2} \cdot \frac{T}{n} + \frac{1}{2} \cdot \frac{T}{n} = \frac{T}{n} = \bar{Y}.$$

(c)  $\text{Var}(\bar{Y}) = \frac{p(1-p)}{n}$ . Since  $n \geq 2$ , we have  $\frac{p(1-p)}{n} \leq \frac{p(1-p)}{2}$ , confirming  $\text{Var}(\hat{p}_{\text{RB}}) \leq \text{Var}(\hat{p})$ .

(d) When  $n = 2$ ,  $\hat{p}_{\text{RB}} = \bar{Y} = \frac{Y_1 + Y_2}{2} = \hat{p}$ , so the two estimators coincide. This makes sense:  $\hat{p}$  already uses all the data, so there is nothing to improve by Rao-Blackwell.

**Problem 4.** Suppose  $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} f_Y(y; \theta) = \theta y^{\theta-1}$  for  $0 < y < 1$  and  $\theta > 0$ .

- (a) Find a sufficient statistic for  $\theta$  using the Factorization Criterion.
- (b) Find  $\hat{\theta}_{\text{MLE}}$  using calculus (this distribution is not NEF).
- (c) Consider the estimator  $\hat{\theta} = \frac{1}{-\log Y_1 - \log Y_2}$  (using only the first two observations; assume  $n \geq 3$ ). Show  $\hat{\theta}$  is unbiased for  $\theta$ .

- (d) Apply Rao-Blackwell to improve  $\hat{\theta}$  using your answer to (a). *Hint:* Use the fact that  $Z_i = -\log Y_i \stackrel{\text{i.i.d.}}{\sim} \text{Expo}(\theta)$ , that  $S = \sum_{i=1}^n Z_i \sim \text{Gamma}(n, \theta)$ , and that  $\frac{Z_1+Z_2}{S} \mid S \sim \text{Beta}(2, n-2)$ .

### Solution

(a) The joint density is

$$f_{\vec{Y}}(\vec{y}; \theta) = \prod_{i=1}^n \theta y_i^{\theta-1} = \theta^n \left( \prod_{i=1}^n y_i \right)^{\theta-1} = \theta^n \underbrace{\left( \prod_{i=1}^n y_i \right)^{\theta}}_{g_{\theta}(T(\vec{y}))} \cdot \underbrace{\left( \prod_{i=1}^n y_i \right)^{-1}}_{h(\vec{y})}.$$

By the Factorization Criterion,  $T = \prod_{i=1}^n Y_i$  (equivalently,  $T = \sum_{i=1}^n \log Y_i$ ) is sufficient for  $\theta$ .

(b) The log-likelihood is

$$\ell(\theta; \vec{y}) = n \log \theta + (\theta - 1) \sum_{i=1}^n \log y_i.$$

Setting  $\ell'(\theta) = \frac{n}{\theta} + \sum_{i=1}^n \log y_i = 0$  gives

$$\hat{\theta}_{\text{MLE}} = -\frac{n}{\sum_{i=1}^n \log Y_i}.$$

Since  $\ell''(\theta) = -\frac{n}{\theta^2} < 0$ , this is indeed a maximum.

(c) Let  $Z_i = -\log Y_i \stackrel{\text{i.i.d.}}{\sim} \text{Expo}(\theta)$ , so  $\hat{\theta} = 1/(Z_1 + Z_2)$  and  $Z_1 + Z_2 \sim \text{Gamma}(2, \theta)$  with density  $f(z) = \theta^2 z e^{-\theta z}$ . Then

$$\mathbb{E}\left[\frac{1}{Z_1 + Z_2}\right] = \int_0^{\infty} \frac{1}{z} \cdot \theta^2 z e^{-\theta z} dz = \theta^2 \int_0^{\infty} e^{-\theta z} dz = \theta^2 \cdot \frac{1}{\theta} = \theta.$$

So  $\hat{\theta}$  is unbiased. (The key: the  $1/z$  and  $z$  from the Gamma density cancel, leaving a clean exponential integral.)

(d) Let  $S = \sum_{i=1}^n Z_i \sim \text{Gamma}(n, \theta)$  (our sufficient statistic, since  $S = -\sum \log Y_i$  is a one-to-one function of  $T$ ). By the hint,  $W = (Z_1 + Z_2)/S \mid S \sim \text{Beta}(2, n-2)$ , so

$$\hat{\theta}_{\text{RB}} = \mathbb{E}\left[\frac{1}{Z_1 + Z_2} \mid S\right] = \mathbb{E}\left[\frac{1}{SW} \mid S\right] = \frac{1}{S} \mathbb{E}[1/W].$$

For  $W \sim \text{Beta}(2, n-2)$ , using the Beta integral:

$$\mathbb{E}[1/W] = \frac{\Gamma(n)}{\Gamma(2)\Gamma(n-2)} \int_0^1 w^0 (1-w)^{n-3} dw = \frac{\Gamma(n)}{\Gamma(2)\Gamma(n-2)} \cdot \frac{\Gamma(1)\Gamma(n-2)}{\Gamma(n-1)} = n-1.$$

Therefore  $\hat{\theta}_{\text{RB}} = \frac{n-1}{S} = -\frac{n-1}{\sum_{i=1}^n \log Y_i}$ . Note this is the unbiased correction of  $\hat{\theta}_{\text{MLE}} =$

$-n/\sum \log Y_i$ , differing only in  $n \rightarrow n-1$ .