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

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 θ if it captures everything the data can tell us about θ. Once we know T, the remaining randomness in $\vec{Y}$ is irrelevant to finding θ. Formally:

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

Example 1 (Bernoulli trials)**.** Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$ 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 θ.

• 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 θ 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_θ may depend on θ (but only through $T(\vec{y})$), and h does not depend on θ 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 θ.*

• ☠*: The factorization only needs to hold up to multiplicative constants that are free of θ. Constant factors can always be absorbed into h.*

Example 2 (Poisson)**.** Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Pois}(\lambda)$. 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 λ, $T = \sum_{i=1}^n Y_i$ (equivalently, $\bar{Y}$) is sufficient for λ.

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

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

2. $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with σ^2 **known**, sufficient statistic for μ.

3. $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, \theta)$, sufficient statistic for θ.

Solution

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 θ, and let T be a sufficient statistic for θ. 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: $\mathbb{E}[\hat{\theta}_{\text{RB}}] = \mathbb{E}[\mathbb{E}[\hat{\theta} \mid T]] = \mathbb{E}[\hat{\theta}]$. So $\text{Bias}(\hat{\theta}_{\text{RB}}) = \text{Bias}(\hat{\theta})$.
(ii) Variance. By Eve's Law:

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

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

• **Why does this work?** The sufficient statistic captures all the information about θ in the data. If your estimator doesn't fully exploit T, it contains residual randomness that is uninformative about θ—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, $\hat{\theta}_{\text{RB}} = \mathbb{E}[\bar{Y} \mid \sum_{i=1}^n Y_i] = \bar{Y}\mathbb{E}[1 \mid \sum_{i=1}^n Y_i] = \bar{Y}$, the original estimator!

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

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

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

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

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

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

Concept Checker 2. Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$. 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 $\hat{p}_{\text{RB}} = \mathbb{E}[Y_1 \mid T]$.
4. Does this result make intuitive sense?

Solution

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

Solution

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:

- θ 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 θ.

• **Strategy**: Rewrite the density $f_Y(y)$ as $\exp(\log(f_Y(y)))$.

5.1 Properties of NEF

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

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

• **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 ψ.

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

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

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

Proof. **(i)** The joint density is

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

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

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

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 Bern($p$) is NEF. We can verify:

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

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 Pois($\lambda$) is NEF. Check:

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

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

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

Solution

Distribution	θ	$\psi(\theta)$	$\mathbb{E}[Y] = \psi'(\theta)$	$\mathrm{Var}(Y) = \psi''(\theta)$	**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(λ)	$\log \lambda$	e^θ	λ	λ	$\sum Y_i$
$\mathcal{N}(\mu, \sigma^2)$, σ^2 known	$\frac{\mu}{\sigma^2}$	$\frac{\sigma^2\theta^2}{2}$	μ	σ^2	$\bar{Y}$
Expo(λ)	$-\lambda$	$-\log(-\theta)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\sum Y_i$
Gamma(α, λ), α 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 $\mathrm{Var}(Y)$ columns are expressed in terms of the original parameter for readability. Sufficient statistics assume $Y_1, \ldots, Y_n$ 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 θ unknown). Assume regularity conditions hold. Show $\mathbb{E}[Y] = \psi'(\theta)$ and $\mathrm{Var}[Y] = \psi''(\theta)$.

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

Solution

Problem 2. Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\alpha, \lambda)$ 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.

(a) Show the Gamma distribution is in the NEF (with α known), and identify θ, $\psi(\theta)$, and $h(y)$.

(b) Use the NEF properties to find $\mathbb{E}[Y]$ and $\text{Var}(Y)$.

(c) Identify a sufficient statistic for λ and find $\hat{\lambda}_{\text{MLE}}$.

Solution

Problem 3. Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$ 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 $\hat{p}_{\text{RB}} = \mathbb{E}\left[\frac{Y_1+Y_2}{2} \,\middle|\, T\right]$ 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

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

(a) Find a sufficient statistic for θ 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 θ.

(d) Apply Rao-Blackwell to improve $\hat{\theta}$ using your answer to (a). *Hint*: Use the fact that $Z_i = -\log Y_i \overset{\text{i.i.d.}}{\sim} \text{Expo}(\theta)$, 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