

Outline:

- Classical estimator performance.
- Bias and mean-square error (MSE).
- Definition of a minimum-variance unbiased (MVU) estimator and (lack of) importance of (strict) unbiasedness.
- Exponential family of distributions.

Reading:

- Chapter 2 in Kay-I,
- sufficient statistics: Sections 5.3–5.4.

Estimator Performance

We now continue with the DC-level-in-Gaussian noise estimation example from handout # 0.

Consider the following two estimators:

$$\begin{aligned}\hat{a}_1 &= \hat{a}_1(\mathbf{X}) = \frac{1}{N} \sum_{n=0}^{N-1} X[n] \\ \hat{a}_2 &= \hat{a}_2(X[0]) = X[0]\end{aligned}$$

where $\mathbf{X} = [X[0], X[1], \dots, X[N-1]]^T$. Here, $\hat{a}_1(\mathbf{X})$ is the *ML estimate* of the DC level a , i.e. it maximizes the likelihood function of a . (Interestingly, the ML estimate of a is the same regardless of whether σ^2 is known or not.) It is also intuitively appealing: a is the *average level* of $X[n]$, since $W[n]$ has zero mean.

Which estimator is better?

For a given realization $\mathbf{X} = \mathbf{x}$ of the measurements, it is possible that either $\hat{a}_1 = \hat{a}_1$ or $\hat{a}_2 = \hat{a}_2$ is closer to a . Hence, we need statistical analysis to answer this question.

Estimator Performance (cont.)

Substitute the measurement model to perform statistical analysis. We have

$$\begin{aligned}\hat{a}_1(\mathbf{X}) &= \frac{1}{N} \sum_{n=0}^{N-1} \underbrace{\{a + W[n]\}}_{X[n]} \\ \hat{a}_2(X[0]) &= \underbrace{a + W[0]}_{X[0]}.\end{aligned}$$

Take expectation:

$$\begin{aligned}\mathbb{E}_{\mathbf{X} | a}[\hat{a}_1(\mathbf{X}) | a] &= \frac{1}{N} \sum_{n=0}^{N-1} (a + \overbrace{\mathbb{E}_W\{W[n]\}}^0) = a \\ \mathbb{E}_{X | a}[\hat{a}_2(X[0]) | a] &= a + \overbrace{\mathbb{E}_W\{W[0]\}}^0 = a.\end{aligned}$$

On average, both estimators are around the correct value, i.e. they are *unbiased*.

Simplified Notation. In the *classical scenario* that we describe here, the DC level a is a constant. This is why $\mathbb{E}_{\mathbf{X} | a}[\cdot | a]$ is equivalent to $\mathbb{E}_{\mathbf{X}}[\cdot]$. Now,

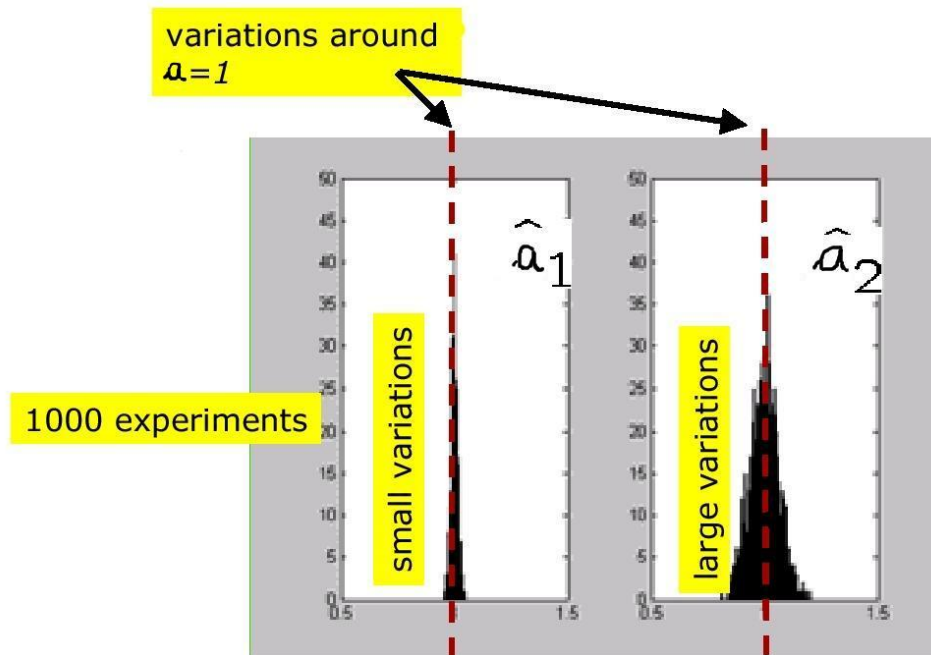
$$\mathbb{E}_{\mathbf{X} | a}[\cdot | a]$$

is obviously more precise than

$$E_{\mathbf{x}}[\cdot] \tag{1}$$

but also more cumbersome. When clear from context that we are dealing with the classical scenario and that a is the value of the parameter, we will simplify the notation and use $E_{\mathbf{x}}[\cdot]$.

Estimator Performance: An Example



Histograms of $\hat{a}_1(\mathbf{X})$ and $\hat{a}_2(\mathbf{X})$.

But, $\hat{a}_1(\mathbf{X})$ is *better* than $\hat{a}_2(\mathbf{X})$ because its pdf is more concentrated around the true value. On average, $\hat{a}_1(\mathbf{X})$ is closer to the true a , equal to $a = 1$ in the above example.

(More) Simplified Notation:

$$\hat{a}_i = \hat{a}_i(\mathbf{X}). \quad (2)$$

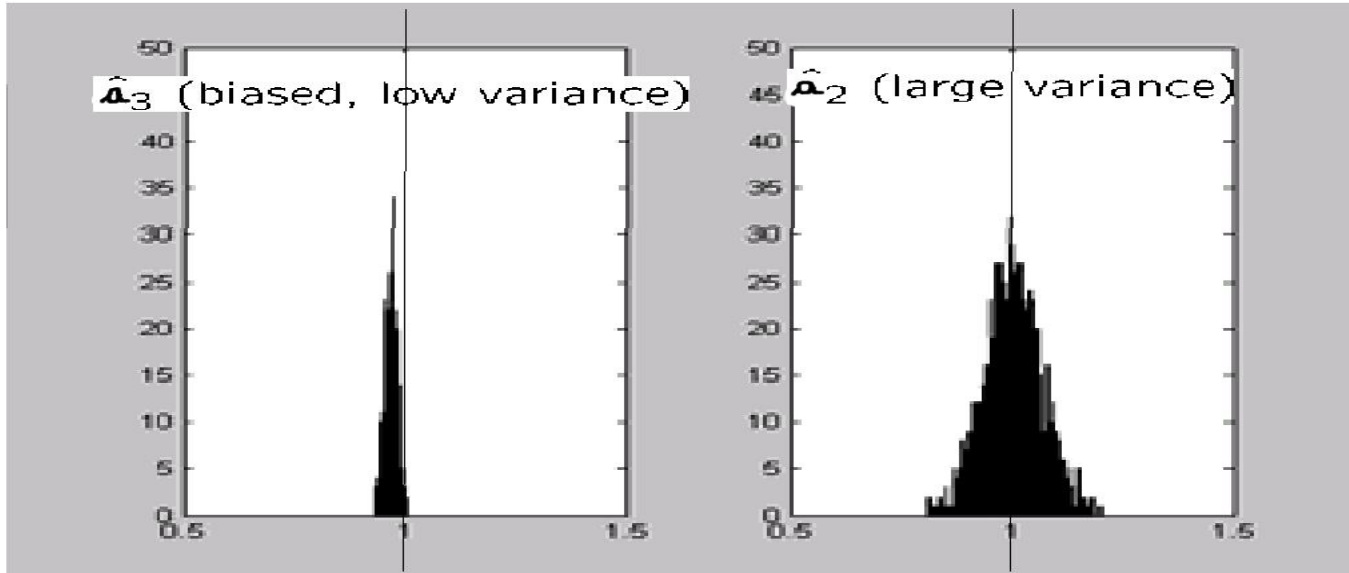
Proof. We simplify the notation using (1) and (2). Then,

$$\begin{aligned}
 E_{\mathbf{X}}(\hat{a}_1) &= E_{\mathbf{X}}(\hat{a}_2) = a \\
 \text{var}_{\mathbf{X}}(\hat{a}_1) &= E_{\mathbf{X}}\left\{[\hat{a}_1 - \overbrace{E_{\mathbf{X}}(\hat{a}_1)}^a]^2\right\} \\
 &= E_{\mathbf{W}}\left\{\left(\frac{1}{N} \sum_{n=0}^{N-1} W[n]\right)^2\right\} \\
 &\stackrel{W[n] \text{ i.i.d.}}{=} \frac{1}{N^2} \sum_{n=0}^{N-1} \text{var}_W(W[n]) = \frac{\sigma^2}{N} \\
 \text{var}_X(\hat{a}_2) &= \text{var}_X(X[0]) = \text{var}_W(W[0]) = \sigma^2.
 \end{aligned}$$

□

What is the justification for taking these conditional expectations? We implicitly assume that we could repeat this experiment many times and plot the histogram, say, of the resulting estimates, as was done on the previous page. This is what Bayesians criticize: in the classical approach, “data that have never been observed are used for inference.” Bayesians do not need this virtual-data argument. Suppose that we are fine with the classical argument and let’s continue.

Which Estimator is the Best?



Definition. Bias and mean-square error (MSE) of an estimator $\hat{\theta} = \hat{\theta}(\mathbf{X})$ under the classical setting:

$$\begin{aligned}\text{bias}\{\hat{\theta}\} &= b(\theta) = E_{\mathbf{X}}(\hat{\theta}) - \theta \\ \text{MSE}\{\hat{\theta}\} &= E_{\mathbf{X}}[(\hat{\theta} - \theta)^2] = \text{var}_{\mathbf{X}}(\hat{\theta}) + b^2(\theta).\end{aligned}$$

The proper (but cumbersome) notation:

$$\begin{aligned}\text{bias}\{\hat{\theta}(\mathbf{X})\} &= E_{\mathbf{X}|\theta}[\hat{\theta}(\mathbf{X}) | \theta] - \theta \\ \text{MSE}\{\hat{\theta}(\mathbf{X})\} &= E_{\mathbf{X}|\theta}\{[\hat{\theta}(\mathbf{X}) - \theta]^2 | \theta\}.\end{aligned}$$

We wish to minimize the above MSE, which may lead to a minimum MSE (MMSE) estimator. The MSE expression can

be written as (using the greatly simplified notation, be careful)

$$\begin{aligned}
 \text{MSE}\{\hat{\theta}\} &= \mathbb{E}_{\mathbf{X}}[(\hat{\theta} - \theta)^2] \\
 &= \mathbb{E}_{\mathbf{X}}(\{\hat{\theta} - \mathbb{E}_{\mathbf{X}}[\hat{\theta}] + \mathbb{E}_{\mathbf{X}}[\hat{\theta}] - \theta\}^2) \\
 &= \mathbb{E}_{\mathbf{X}}\{[\hat{\theta} - \mu(\theta) + \mu(\theta) - \theta]^2\} \\
 &= \underbrace{\mathbb{E}_{\mathbf{X}}\{[\hat{\theta} - \mu(\theta)]^2\}}_{\text{var}_{\mathbf{X}}(\hat{\theta})} + \underbrace{[\mu(\theta) - \theta]^2}_{b^2(\theta)} \\
 &\quad + 2 \underbrace{\mathbb{E}_{\mathbf{X}}\{[\hat{\theta} - \mu(\theta)] \cdot [\underbrace{\mathbb{E}_{\mathbf{X}}(\hat{\theta}) - \theta}_{\text{const}}]\}}_0
 \end{aligned}$$

where we have defined

$$\mu(\theta) = \mathbb{E}_{\mathbf{X}}(\hat{\theta}).$$

This $\text{MSE}\{\hat{\theta}\}$ is different from Bayesian MSE. Since, in Bayesian inference, we assign a prior distribution to θ , Bayesian MSE is obtained by taking the expectation of $\text{MSE}\{\hat{\theta}\}$ with respect to θ , see (2) in handout # 0.

In the classical scenario, minimizing the MSE would be a reasonable estimator design criterion. However, $\text{MSE}\{\hat{\theta}(\mathbf{X})\} = \text{var}_{\mathbf{X}}(\hat{\theta}) + b^2(\theta)$ is usually a function of θ , and minimizing it over some family of estimators will usually produce an optimal “estimator” $\hat{\theta}(\mathbf{X})$ that depends on θ .

Example 1. DC level in **A**dditive **W**hite **G**aussian **N**oise (AWGN), see also Section 2.4 in Kay-I:

$$X[n] = a + \underbrace{W[n]}_{\text{AWGN}}$$

$$W[n] \sim \mathcal{N}(0, \sigma^2) \quad n = 0, 1, \dots, N - 1.$$

Consider the following family of estimators of a :

$$\check{a} = c \bar{X} \quad (3)$$

where

$$\bar{X} = \frac{1}{N} \sum_{n=0}^{N-1} X[n] \quad (\text{sample mean}).$$

Here

$$\mathbb{E}_{\mathbf{X}}(\check{a}) = c a, \quad \text{var}_{\mathbf{X}}(\check{a}) = c^2 \sigma^2 / N.$$

Find the best c that minimizes the MSE for the family (3). In other words, can we improve upon the sample mean?

$$\text{MSE}\{\check{a}\} = \text{MSE}(c) = c^2 \sigma^2 / N + \underbrace{(c a - a)^2}_{\text{depends on } a}$$

$$\frac{\text{MSE}(c)}{dc} = 2 c \sigma^2 / N + 2 (c a - a) a = 0$$

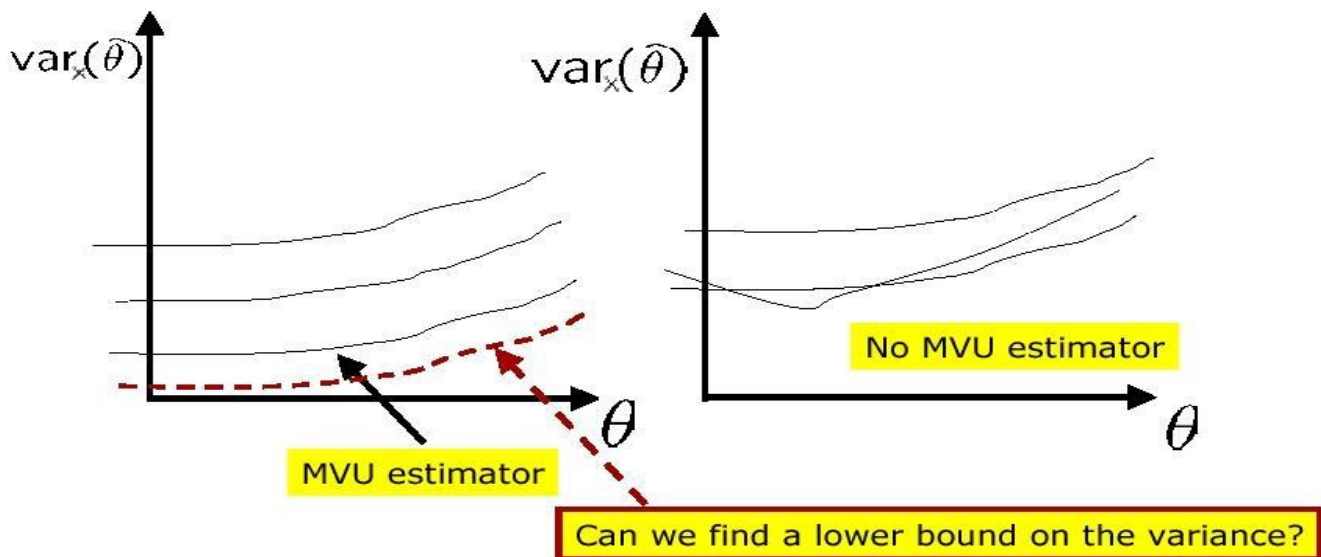
$$c_{\text{opt}} = \frac{a^2}{a^2 + \sigma^2 / N}$$

depends on the unknown parameter a . Hence not useful, at least not directly. Observe the shrinkage form of the above “estimator.”

Minimum-variance Unbiased (MVU) Estimation

How do we construct a realizable (proper) estimator?

An idea: Constrain the bias to be zero and then minimize the estimator variance (which is equal to MSE in this scenario since the bias is zero) for all values of θ : MVU estimator $\hat{\theta}_{\text{MVU}} = \hat{\theta}_{\text{MVU}}(\mathbf{x})$.



MVU estimator $\hat{\theta}_{\text{MVU}}$ does not always exist, since $\hat{\theta}_{\text{MVU}}$ must have the smallest variance for all values of θ .¹

¹To emphasize the fact that the MVU estimator must have the smallest variance for all values of θ , B & D refer to it as *uniformly minimum variance unbiased* (UMVU).

Comments:

- Even if it exists for a particular problem, MVU estimator is not optimal in terms of minimizing the MSE and we may be able to do better.
- Unbiasedness is nice, but not the most important \implies we can relax this condition and consider biased estimators as well, e.g. by making them *asymptotically unbiased*. By relaxing the unbiasedness condition, it is possible to outperform the MVU estimators in terms of MSE, as shown in the following example.

Example 2.² Consider now estimating the variance σ^2 of independent, identically distributed (i.i.d.) zero-mean Gaussian observations, using the following family of estimators of σ^2 :

$$\underbrace{\hat{\sigma}^2}_{\hat{\sigma}^2(\mathbf{x})} = c \cdot \frac{1}{N} \sum_{n=0}^{N-1} X^2[n] \quad (4)$$

where $c > 0$ is a constant that we need to select. If we choose

²This example is based on P. Stoica and R. Moses, "On biased estimators and the unbiased Cramér-Rao lower bound," *Signal Processing*, vol. 21, pp. 349–350, 1991. A link to this article is posted on EE 527 web site

$c = 1$, $\hat{\sigma}^2|_{c=1}$ is an unbiased estimator with

$$\hat{\sigma}^2|_{c=1} = \hat{\sigma}_{\text{MVU}}^2 = \frac{1}{N} \sum_{n=0}^{N-1} X^2[n]. \quad (5)$$

Now, for the estimator family (4),

$$\mathbb{E}_{\mathbf{X}}(\hat{\sigma}^2) = c \sigma^2$$

and

$$\begin{aligned} \text{MSE}\{\hat{\sigma}^2\} &= \mathbb{E}_{\mathbf{X}}[(\hat{\sigma}^2 - \sigma^2)^2] = \mathbb{E}_{\mathbf{X}}[(\hat{\sigma}^2)^2] + (\sigma^2)^2 - 2\sigma^2 \mathbb{E}_{\mathbf{X}}(\hat{\sigma}^2) \\ &= \mathbb{E}_{\mathbf{X}}\{(\hat{\sigma}^2)^2\} + (\sigma^2)^2 (1 - 2c) \\ &= \frac{c^2}{N^2} \sum_{n_1=0}^{N-1} \sum_{n_2=0}^{N-1} \mathbb{E}_{\mathbf{X}}\{X^2[n_1]X^2[n_2]\} + (\sigma^2)^2 (1 - 2c) \\ &= \frac{c^2}{N^2} [(N^2 - N) (\sigma^2)^2 + N \cdot \underbrace{\mathbb{E}_{\mathbf{X}}\{X^4[n]\}}_{3(\sigma^2)^2}] + (\sigma^2)^2 (1 - 2c) \\ &= (\sigma^2)^2 \cdot \left[c^2 \left(1 + \frac{2}{N}\right) + (1 - 2c) \right]. \end{aligned} \quad (6)$$

To evaluate (6), we have used the following facts:

- For $n_1 \neq n_2$,

$$\mathbb{E}_{\mathbf{X}}\{X^2[n_1] X^2[n_2]\} = \mathbb{E}_{\mathbf{X}}\{X^2[n_1]\} \cdot \mathbb{E}_{\mathbf{X}}\{X^2[n_2]\} = \sigma^2 \cdot \sigma^2 = (\sigma^2)^2.$$

- For $n_1 = n_2$,

$$E_X\{X^2[n_1]X^2[n_2]\} = E_X\{X^4[n_1]\} = 3(\sigma^2)^2$$

which is the fourth-order moment of a Gaussian distribution having zero mean and variance σ^2 .

Here, (6) is minimized for

$$c_{\text{OPT}} = \frac{N}{N+2}$$

yielding the estimator

$$\hat{\sigma}_{\star}^2 = c_{\text{OPT}} \frac{1}{N} \sum_{n=0}^{N-1} X^2[n]$$

whose MSE is minimum for the family of estimators in (4):

$$\text{MSE}_{\text{MIN}} = \frac{2(\sigma^2)^2}{N+2}.$$

Comments:

- $\hat{\sigma}_{\star}^2$ is *biased* and has *smaller MSE* than the MVU estimator in (5):

$$\text{MSE}_{\text{MIN}} < \text{MSE}\{\hat{\sigma}^2\} \Big|_{c=1} = \frac{2\sigma^4}{N}.$$

- Unlike in Example 1, we are able to construct a realizable estimator in this case.
- For large N , $\hat{\sigma}_\star^2$ and $\hat{\sigma}_{\text{MVU}}^2$ are approximately the same, since $N/(N+2) \nearrow 1$ as $N \nearrow \infty$. This also implies that $\hat{\sigma}_\star^2$ is *asymptotically unbiased*.

Note: Bias considerations should not be completely dismissed. For example, we may have two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ with

$$[\text{bias}\{\hat{\theta}_1\}]^2 \ll \text{var}_X\{\hat{\theta}_1\} \quad \text{and} \quad [\text{bias}\{\hat{\theta}_2\}]^2 \ll \text{var}_X\{\hat{\theta}_2\}$$

and

$$\text{MSE}\{\hat{\theta}_1\} \approx \text{MSE}\{\hat{\theta}_2\}.$$

Hence, these two estimators are *equally good* as far as MSE is concerned. But, we may have

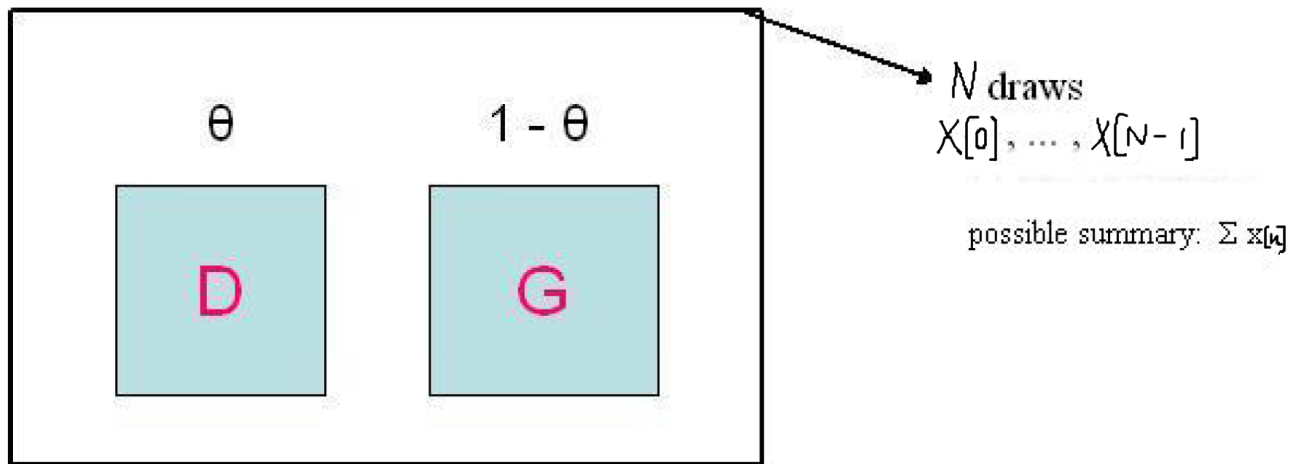
$$|\text{bias}\{\hat{\theta}_1\}| \ll |\text{bias}\{\hat{\theta}_2\}|$$

making $\hat{\theta}_1$ *more desirable* than $\hat{\theta}_2$. *Bias correction* methods have been developed for constructing estimators that have small bias, by improving (bias-correcting) estimators that have small MSE (e.g. ML estimator versus bias-corrected ML estimator). Hence, having small bias is typically a second-tier concern (compared with minimizing the MSE), but a valid one, particularly in the scenario outlined in this comment.

Sufficiency

A function $T(\mathbf{X})$ of the observations \mathbf{X} *only* is called a *statistic*.

Example: A machine produces N items in succession, with probability θ of producing a defective product. Suppose that there is no dependence in quality of the produced items.



Then, our statistical model is

$$\begin{aligned} p_{\mathbf{X} | \theta}(\mathbf{x} | \theta) &= \prod_{n=0}^{N-1} \theta^{x[n]} (1 - \theta)^{1-x[n]} \\ &= \theta^{\sum_{n=0}^{N-1} x[n]} (1 - \theta)^{N - \sum_{n=0}^{N-1} x[n]}. \end{aligned}$$

Is there a loss of information by keeping and recording only $\sum_{n=0}^{N-1} x[n]$? Answer:

{ **Yes**, we are dropping a lot of information. But
 No, in terms of inference about θ .

We typically wish to separate out any aspects of the data that are irrelevant in the context of our model. In other words, we would like to reduce the data and deal only with the statistics “whose use involves no loss of information.” For example, we could save memory and store only the reduced data. What we mean by “no loss of information” is quantified in the following definition.

Definition. $T = T(\mathbf{X})$ is a sufficient statistic for θ if the conditional distribution of \mathbf{X} given $T(\mathbf{X})$ *does not* involve θ :

$$p_{\mathbf{X} | T(\mathbf{X}), \theta}(\mathbf{x} | T(\mathbf{x}) = t, \theta) \quad \text{not a function of } \theta.$$

Think of sufficient statistics as not throwing away any useful information about θ .

Trivial example: $T(\mathbf{X}) = \mathbf{X}$, i.e. full data are always sufficient.

Example: Continue with the previous example. Here, $\mathbf{X} = [X[0], \dots, X[N-1]]^T$ is the record of N Bernoulli trials with probability θ , which can be written as

$$\Pr\{X[n] = x[n]\} = \theta^{x[n]} (1 - \theta)^{1-x[n]}$$

where $x[n]$ is 1 (defective) or 0 (not defective). Thus

$$\begin{aligned}\Pr\{\mathbf{X} = \mathbf{x}\} &= \Pr\{X[0] = x[0], \dots, X[N-1] = x[N-1]\} \\ &= \theta^t (1 - \theta)^{n-t}\end{aligned}$$

where $t = \sum_{n=0}^{N-1} x[n]$. Now, $T(\mathbf{X}) = \sum_{n=0}^{N-1} X[n]$ has a binomial distribution $\text{Bin}(N, \theta)$ and

$$\begin{aligned}p_{X|T(\mathbf{X}),\theta}(\mathbf{x} | T(\mathbf{x}) = t, \theta) &= \Pr\{\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t\} \\ &= \frac{\Pr\{\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t\}}{\binom{N}{t} \theta^t (1 - \theta)^{N-t}} \\ &= \begin{cases} 0, & \text{if } \sum_{n=0}^{N-1} x[n] \neq t \\ \frac{\theta^t (1 - \theta)^{N-t}}{\binom{N}{t} \theta^t (1 - \theta)^{N-t}} = \frac{1}{\binom{N}{t}}, & \text{otherwise} \end{cases}\end{aligned}$$

which is clearly not a function of θ . Here, we have used the general fact that

$$\{\mathbf{X} = \mathbf{x}\} \subset \{T(\mathbf{X}) = T(\mathbf{x})\}. \quad (7)$$

Thus, $T(\mathbf{X}) = \sum_{n=0}^{N-1} X[n]$ is a sufficient statistic for θ .

In general, directly checking sufficiency is difficult because we need to compute conditional distributions. Fortunately, the following theorem has conditions that are easy to verify.

Theorem. (Factorization Theorem) A statistic $T(\mathbf{X})$ is sufficient for θ if and only if there exists a function $g(t, \theta)$ and a function $h(\mathbf{x})$ such that

$$f_{X|\theta}(x|\theta) = \underbrace{g(T(x), \theta)}_{\substack{\text{parameters} \\ \text{coupled with} \\ \text{sufficient} \\ \text{statistics}}} \cdot h(x).$$

Note: $T(x)$ must be a *statistic*, a function of data x *only*.

Proof. To illustrate the idea of the proof and for simplicity, we concentrate on the discrete case. Suppose that $T(X)$ is a sufficient statistic. Then

$$\begin{aligned} p_{X|\theta}(x|\theta) &= \overbrace{\Pr\{X=x, T(X)=T(x)\}}^{P\{X=x\}, \text{ see (7)}} \\ &= \Pr\{T(X)=T(x)\} \underbrace{\Pr\{X=x | T(X)=T(x)\}}_{\substack{h(x), \\ \text{by sufficiency}}} \\ &= g(T(x), \theta) h(x). \end{aligned}$$

Conversely,

$$\begin{aligned}
 \Pr\{X = x \mid T(X) = T(x)\} &= \frac{\Pr\{X = x, T(X) = T(x)\}}{\Pr\{T(X) = T(x)\}} \\
 &= \frac{\overbrace{\Pr\{X = x\}}^{p_{X|\theta}(x|\theta)}}{\underbrace{\Pr\{T(X) = T(x)\}}_{\sum_{y:T(y)=T(x)} p_{X|\theta}(y|\theta)}} \\
 &= \frac{\overbrace{g(T(x), \theta) h(x)}^{\text{by the assumption}}}{\sum_{y:T(y)=T(x)} \underbrace{g(T(y), \theta) h(y)}_{\text{by the assumption}}} \\
 &= \frac{g(T(x), \theta) h(x)}{g(T(x), \theta) \sum_{y:T(y)=T(x)} h(y)} = \frac{h(x)}{\sum_{y:T(y)=T(x)} h(y)}
 \end{aligned}$$

which is *not* a function of θ . \square

Example: Suppose that $X[0], X[1], \dots, X[N-1]$ are i.i.d. $\mathcal{N}(a, \sigma^2)$ (conditional on a and θ). Define $\boldsymbol{\theta} = [a, \sigma^2]^T$ and

$$\mathbf{X} = \begin{bmatrix} X[0] \\ X[1] \\ \vdots \\ X[N-1] \end{bmatrix}.$$

Then

$$\begin{aligned} f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta}) &= (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - a)^2\right\} \\ &= (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{Na^2}{2\sigma^2}\right) \\ &\quad \cdot \exp\left\{-\frac{1}{2\sigma^2} \left(\sum_{n=0}^{N-1} x^2[n] - 2a \sum_{n=0}^{N-1} x[n]\right)\right\}. \end{aligned} \quad (8)$$

Applying the factorization theorem leads to the sufficient statistics for $\boldsymbol{\theta}$:

$$\mathbf{T}(\mathbf{x}) = \begin{bmatrix} \sum_{n=0}^{N-1} x[n] \\ \sum_{n=0}^{N-1} x^2[n] \end{bmatrix}.$$

Here, $h(\mathbf{x})$ is trivial: $h(\mathbf{x}) = 1$.

Define

$$\bar{x} = \frac{1}{N} \sum_{n=0}^{N-1} x[n].$$

A frequently used equivalent sufficient statistic

$$\mathbf{T}(\mathbf{x}) = \begin{bmatrix} \bar{x} \\ \frac{1}{N-1} \sum_{n=0}^{N-1} (x[n] - \bar{x})^2 \end{bmatrix}$$

can be obtained by suitably arranging the terms in the expression for $f_{\mathbf{X}|a}(\mathbf{x}|\boldsymbol{\theta})$:

$$f_{\mathbf{X}|a}(\mathbf{x}|\boldsymbol{\theta}) = (2\pi\sigma^2)^{-N/2} \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}[(x[n] - \bar{x}) + (\bar{x} - a)]^2\right\}$$

and expanding the squares in the exponent:

$$f_{\mathbf{X}|a}(\mathbf{x}|a) = (2\pi\sigma^2)^{-N/2} \cdot \exp\left(-\frac{N}{2\sigma^2}\left\{(\bar{x} - a)^2 + \frac{1}{N}\sum_{n=0}^{N-1}(x[n] - \bar{x})^2\right\}\right).$$

Example. Suppose $X[0], X[1], \dots, X[N-1]$ are i.i.d. $\mathcal{N}(a, 1)$ given a . Then, using (8) with $\sigma^2 = 1$, we obtain

$$f_{\mathbf{X}|a}(\mathbf{x}|a) = \underbrace{\exp\{N a (\bar{x} - \frac{1}{2} a)\} \cdot (2\pi)^{-\frac{1}{2}N} \exp\left(-\frac{1}{2}\sum_{n=0}^{N-1} x^2[n]\right)}_{h(\mathbf{x})}$$

and, by the factorization theorem, \bar{x} is a sufficient statistic for a .

Example: Digital Communications

$$x(t) = \underbrace{s(t)}_{\text{signal}} + \underbrace{w(t)}_{\text{noise}}$$

where the signal $s(t)$ is usually represented using orthonormal basis functions $\varphi_k(t)$:

$$s(t) = \sum_{k=1}^K \alpha_k \varphi_k(t).$$

Note: The signal $s(t)$ is unknown, but it has known structure, incorporated in this basis-function expansion. **We wish to use this structure for data reduction.**

If $\varphi_k(t)$ are orthonormal, α_k can be computed as:

$$\alpha_k = \int s(t) \varphi_k(t) dt.$$

Here, our goal at the receiver is to decide which $s(t)$ (α_k 's) has been transmitted.

What is typically done in communication receivers is the following: the received data $x(t)$ are matched to the basis functions, i.e.

$$\hat{\alpha}_k = \int x(t) \varphi_k(t) dt \quad k = 1, 2, \dots, K \quad (9)$$

are computed and utilized for demodulation.

Question: Are the $\hat{\alpha}_k$ s sufficient statistics for inference about $s(t)$ (or, more precisely, for inference on the α_k s)?

Note: In some applications, sampled data $x[n]$, $n = 0, 1, \dots, N - 1$ are available and

$$\hat{\alpha}_k = \hat{\alpha}_k(\mathbf{x}) = \sum_{n=0}^{N-1} x[n] \varphi_k[n] \quad k = 1, 2, \dots, K$$

are used to approximate the integrals in (9) (up to a scaling factor). We focus on this scenario, having in mind that we can easily switch from sums to integrals by letting the sampling interval go to zero — then N will go to infinity. Clearly, N is much larger than the number of basis functions K , i.e.

$$K \ll N.$$

In the sampled-data case, our model is

$$X[n] = \underbrace{s[n]}_{\text{signal}} + \underbrace{W[n]}_{\text{noise}}$$

where

$$s[n] = \sum_{k=1}^K \alpha_k \varphi_k[n].$$

Define

$$\mathbf{X} = \begin{bmatrix} X[0] \\ X[1] \\ \vdots \\ X[N-1] \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} W[0] \\ W[1] \\ \vdots \\ W[N-1] \end{bmatrix}$$

and

$$\boldsymbol{\mu}(\boldsymbol{\alpha}) = \begin{bmatrix} s[0] \\ s[1] \\ \vdots \\ s[N-1] \end{bmatrix} \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_K \end{bmatrix}$$

implying

$$\mathbf{X} = \boldsymbol{\mu}(\boldsymbol{\alpha}) + \mathbf{W}.$$

If the noise is additive zero-mean Gaussian with covariance matrix

$$\mathbb{E}_{\mathbf{W}}(\mathbf{W} \mathbf{W}^T) = \mathbf{C}$$

then

$$\{\mathbf{X} \mid \boldsymbol{\alpha}\} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\alpha}), \mathbf{C})$$

which is a multivariate Gaussian pdf:

$$f_{\mathbf{X} \mid \boldsymbol{\alpha}}(\mathbf{x} \mid \boldsymbol{\alpha}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{C}|}} \exp \left\{ -\frac{1}{2} [\mathbf{x} - \boldsymbol{\mu}(\boldsymbol{\alpha})]^T \mathbf{C}^{-1} [\mathbf{x} - \boldsymbol{\mu}(\boldsymbol{\alpha})] \right\} \quad (10)$$

and

$$\boldsymbol{\mu}(\boldsymbol{\alpha}) = \begin{bmatrix} s[0] \\ s[1] \\ \vdots \\ s[N-1] \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^K \alpha_k \varphi_k[0] \\ \sum_{k=1}^K \alpha_k \varphi_k[1] \\ \vdots \\ \sum_{k=1}^K \alpha_k \varphi_k[N-1] \end{bmatrix} = F \boldsymbol{\alpha}$$

where

$$F = \begin{bmatrix} \varphi_1[0] & \varphi_2[0] & \cdots & \varphi_K[0] \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1[N-1] & \varphi_2[N-1] & \cdots & \varphi_K[N-1] \end{bmatrix}$$

is an $N \times K$ matrix. So

$$f_{\mathbf{X}|\boldsymbol{\alpha}}(\mathbf{x}|\boldsymbol{\alpha}) = \frac{1}{\sqrt{|2\pi C|}} \exp\left[-\frac{1}{2}(\mathbf{x} - F\boldsymbol{\alpha})^T C^{-1}(\mathbf{x} - F\boldsymbol{\alpha})\right].$$

What are the sufficient statistics for inference on $\boldsymbol{\alpha}$? If C is *unknown*, we cannot separate out any non-trivial sufficient statistics for both $\boldsymbol{\alpha}$ and C . If C is *known*, then the vector of sufficient statistics for $\boldsymbol{\alpha}$ is

$$F^T C^{-1} \mathbf{x} \tag{11}$$

which is a $K \times 1$ vector. Since, $K \ll N$, (11) achieves dimensionality reduction compared with the raw data \mathbf{x} .

For white noise ($C = \sigma^2 I$) with known variance σ^2 , (11) simplifies to (up to a known proportionality factor):

$$F^T \mathbf{x} = \begin{bmatrix} \sum_{n=0}^{N-1} \varphi_1[n] x[n] \\ \sum_{n=0}^{N-1} \varphi_2[n] x[n] \\ \vdots \\ \sum_{n=0}^{N-1} \varphi_K[n] x[n] \end{bmatrix} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_K \end{bmatrix}.$$

If σ^2 is *unknown*, then

$$f_{\mathbf{x} | \alpha}(\mathbf{x} | \alpha) = \frac{1}{\sqrt{(2\pi\sigma^2)^N}} \cdot \exp \left[-\frac{1}{2\sigma^2} (\mathbf{x} - F\alpha)^T (\mathbf{x} - F\alpha) \right].$$

where σ^2 and α are parameters and \mathbf{x} is the data vector. Now

$$\mathbf{x}^T \mathbf{x} = \sum_{n=0}^{N-1} x^2[n] \quad \text{and} \quad F^T \mathbf{x} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_K \end{bmatrix}$$

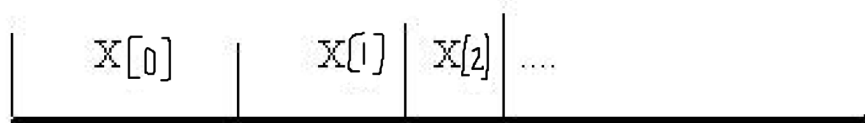
are jointly sufficient for α and σ^2 .

Examples: Computing Sufficient Statistics

Example: Suppose that elements of

$$\mathbf{X} = [X[0], X[1], \dots, X[N-1]]^T$$

are i.i.d. inter-arrival times of packets arriving at a node in a communication network.



Assume that $X[n]$, $n = 0, 1, \dots, N-1$ come from $\text{Expon}(\theta)$ distribution (conditional on θ):

$$\begin{aligned} f_{\mathbf{X}|\theta}(\mathbf{x}|\theta) &= \prod_{n=0}^{N-1} [\theta e^{-\theta x[n]} i_{[0,\infty)}(x[n])] \\ &= \theta^N \exp\left(-\theta \underbrace{\sum_{n=0}^{N-1} x[n]}_{T(\mathbf{x})}\right) i_{[0,\infty)}\left(\min_{\forall n} x[n]\right) \end{aligned}$$

where $i_A(x)$ denotes the indicator function:

$$i_A(x) = \begin{cases} 1, & x \in A, \\ 0, & \text{otherwise} \end{cases} .$$

Note that

$$\prod_{n=0}^{N-1} i_{[0,\infty)}(x[n]) = i_{[0,\infty)}(\min_{\forall n} x[n])$$

since

$$x[0], x[1], \dots, x[N-1] \geq 0 \iff \min_{\forall n} x[n] \geq 0.$$

Example: Elements of

$$\mathbf{X} = [X[0], X[1], \dots, X[N-1]]^T$$

are i.i.d. uniform(0, θ) (conditional on θ):

$$\begin{aligned} f_{\mathbf{X} | \theta}(\mathbf{x} | \theta) &= \prod_{n=0}^{N-1} \left[\frac{1}{\theta} i_{[0,\theta]}(x[n]) \right] \\ &= \frac{1}{\theta^N} \left[\prod_{n=0}^{N-1} i_{[0,\theta]}(x[n]) \right] \\ &= \underbrace{\frac{1}{\theta^N} i_{(-\infty, \theta]}(\max_{\mathbf{x}} x[n])}_{g(T(\mathbf{x}), \theta)} \cdot \underbrace{i_{[0,\infty)}(\min_{\mathbf{x}} x[n])}_{h(\mathbf{x})} \end{aligned}$$

Here, we have used the facts that

$$x[0], x[1], \dots, x[N-1] \leq \theta \iff \max_{\forall n} x[n] \leq \theta$$

and

$$x[0], x[1], \dots, x[N-1] \geq 0 \iff \min_{\forall n} x[n] \geq 0.$$

Example. Detection problem: $\theta \in \{0, 1\}$ and

$$\begin{aligned} f_{\mathbf{X}|\theta}(\mathbf{x}|\theta) &= \theta f_{\mathbf{X}|\theta}(\mathbf{x}|1) + (1-\theta) f_{\mathbf{X}|\theta}(\mathbf{x}|0) \\ &= \underbrace{\left[\theta \frac{\overbrace{f_{\mathbf{X}|\theta}(\mathbf{x}|1)}^{T(\mathbf{x})}}{f_{\mathbf{X}|\theta}(\mathbf{x}|0)} + (1-\theta) \right]}_{g(T(\mathbf{x}), \theta)} \underbrace{f_{\mathbf{X}|\theta}(\mathbf{x}|0)}_{h(\mathbf{x})} \end{aligned}$$

The likelihood ratio $T(\mathbf{x})$ is sufficient for θ . It is a very useful sufficient statistics because it is one-dimensional regardless of the nature of $f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)$. See also Poor, Example IV.C.1.

Definition. The statistic $T(\mathbf{x})$ is *minimally sufficient* if it is sufficient and provides a reduction of data greater than or equal to the data reduction achieved by any other sufficient statistic $S(\mathbf{x})$.

(Multiparameter) Exponential Family of Distributions

$$f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x}) \exp \left[\underbrace{\sum_{i=1}^d \eta_i(\boldsymbol{\theta}) T_i(\mathbf{x}) - B(\boldsymbol{\theta})}_{\text{coupling between parameters and data has a very specific form}} \right]$$

where $\boldsymbol{\theta}$ is a d -dimensional vector:

$$\boldsymbol{\theta} \in \mathbb{R}^d.$$

By the factorization theorem, $\mathbf{T}(\mathbf{X}) = [T_1(\mathbf{X}), \dots, T_d(\mathbf{X})]^T$ is sufficient for $\boldsymbol{\theta}$. It is the *natural sufficient statistic* of the family. The exponential family is important: It covers quite a few useful distributions, including some that are fairly complex; e.g. Markov random fields, used in image analysis, are virtually all in the exponential-family form.

Note that, for pdfs/pmfs in the exponential family, the size d of $\boldsymbol{\theta}$ is the *same* as the dimension of the vector of natural sufficient statistics.

For more about exponential families, see Sections 3.5.4 and 3.5.5 in Hero's notes.

Recall that the support of $f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta})$ is the set of \mathbf{x} for which

the pdf is positive:

$$A = \{\mathbf{x} \mid f_{\mathbf{X}|\theta}(\mathbf{x}|\theta) > 0\}.$$

Useful tests:

- if the support A of $f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)$ depends on θ , then $f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)$ *cannot* belong to the exponential family of distributions. For example, $U(0, \theta)$ is not a member of the exponential family. Read also Section 3.5.5 in Hero's notes.
- Keep in mind: for pdfs/pmfs in the exponential family, the size d of θ is the *same* as the dimension of the vector of natural sufficient statistics.

Suppose that we have multiple i.i.d. measurements

$$\mathbf{X} = [X[0], X[1], \dots, X[N-1]]^T$$

coming from the above pdf/pmf. Then,

$$\begin{aligned}
 f_{X|\theta}(\mathbf{x}|\theta) &= \prod_{n=0}^{N-1} \left\{ h(x[n]) \exp \left[\sum_{i=1}^d \eta_i(\boldsymbol{\theta}) T_i(x[n]) - B(\boldsymbol{\theta}) \right] \right\} \\
 &= \underbrace{\left[\prod_{n=0}^{N-1} h(x[n]) \right] \cdot \exp \left[\sum_{i=1}^d \eta_i(\boldsymbol{\theta}) \sum_{n=0}^{N-1} T_i(x[n]) - N B(\boldsymbol{\theta}) \right]}_{\text{again the exponential family}}
 \end{aligned}$$

and, hence, the vector of natural sufficient statistics is

$$\mathbf{T}(\mathbf{x}) = \left[\sum_{n=0}^{N-1} T_1(x[n]), \sum_{n=0}^{N-1} T_2(x[n]), \dots, \sum_{n=0}^{N-1} T_d(x[n]) \right]^T.$$

A Side Note on One-Parameter Canonical Exponential Family

Here is a simple special sub-family of the exponential family, *the one-parameter canonical exponential family*:

$$f_{\mathbf{X} | \eta}(\mathbf{x} | \eta) = h(\mathbf{x}) \exp \left[\underbrace{\eta}_{\substack{\text{scalar} \\ \text{canonical} \\ \text{parameter}}} T(\mathbf{x}) - A(\eta) \right]$$

where

$$A(\eta) = \ln \int \cdots \int h(\boldsymbol{\chi}) \exp[\eta T(\boldsymbol{\chi})] d\boldsymbol{\chi} \quad \text{for a pdf } f_{\mathbf{X} | \eta}(\mathbf{x} | \eta)$$

$$A(\eta) = \ln \sum_{\boldsymbol{\chi}} h(\boldsymbol{\chi}) \exp[\eta T(\boldsymbol{\chi})] \quad \text{for a pmf } p_{\mathbf{X} | \eta}(\mathbf{x} | \eta).$$

If we can compute the normalizing term $A(\eta)$ in a simple form, then it is easy to find the mean and variance of $T(X)$ given η :

$$\mathbb{E}_{\mathbf{X} | \eta}[T(\mathbf{X}) | \eta] = \frac{dA(\eta)}{d\eta}, \quad \text{var}_{\mathbf{X} | \eta}[T(\mathbf{X}) | \eta] = \frac{d^2 A(\eta)}{d\eta^2}. \quad (12)$$

Why is this useful? Here is an example. Suppose that

$X[0], X[1], \dots, X[N-1]$ are conditionally i.i.d. given θ^2 , from

$$f_{X|\theta^2}(x|\theta) = \underbrace{\frac{x}{\theta^2} \exp\left(-\frac{x^2}{2\theta^2}\right)}_{\text{Rayleigh pdf}} i_{[0,\infty)}(x).$$

Define $\mathbf{X} = [X[0], X[1], \dots, X[N-1]]^T$ and $\mathbf{x} = [x[0], x[1], \dots, x[N-1]]^T$ and write the pdf of \mathbf{X} given θ (i.e. the likelihood function of θ for data \mathbf{x}) as

$$\begin{aligned} f_{\mathbf{X}|\theta}(\mathbf{x}|\theta) &= \prod_{n=0}^{N-1} \left[\frac{x[n]}{\theta^2} \cdot \exp\left(-\frac{x^2[n]}{2\theta^2}\right) \right] \\ &= \left(\prod_{n=0}^{N-1} x[n] \right) \cdot \exp \left[\underbrace{-\frac{1}{2\theta^2}}_{\eta(\theta^2)} \cdot \underbrace{\left(\sum_{n=0}^{N-1} x^2[n] \right)}_{T(\mathbf{x})} - \underbrace{N \ln(\theta^2)}_{B(\theta^2)} \right] \end{aligned}$$

implying

$$\eta = -\frac{1}{2\theta^2}, \quad \theta^2 = -\frac{1}{2\eta}$$

and, consequently,

$$\underbrace{A(\eta)}_{B(\theta)=N \ln(\theta^2)} = N \ln\left(-\frac{1}{2\eta}\right) = -N \ln(-2\eta).$$

Therefore, the natural sufficient statistic $T(\mathbf{X}) = \sum_{n=0}^{N-1} X^2[n]$

has mean

$$\mathbb{E}_{\mathbf{X} | \eta} \left(\sum_{n=0}^{N-1} X^2[n] \mid \eta \right) = \frac{dA(\eta)}{d\eta} = -\frac{N}{\eta} = 2 N \theta^2$$

and variance

$$\text{var}_{\mathbf{X} | \eta} \left(\sum_{n=0}^{N-1} X^2[n] \mid \eta \right) = \frac{d^2 A(\eta)}{d\eta^2} = \frac{N}{\eta^2} = 4 N (\theta^2)^2.$$

Direct computation of these moments would be more complicated.