data

MDPI

Article

Information Loss Due to the Data Reduction of Sample Data from Discrete Distributions

Maryam Moghimi *,†,‡ **and Herbert W. Corley** *,‡

Center on Stochastic Modeling, Optimization, and Statistics (COSMOS), the University of Texas at Arlington, Arlington, TX 76013, USA

* Correspondence: maryam.moghimi@mavs.uta.edu (M.M.); corley@uta.edu (H.W.C.); Tel.: +1-214-971-0904 (M.M.); +1-817-272-3092 (H.W.C.)

† This paper was part of the author's doctoral dissertation of May 2020.

‡ The two authors contributed equally to this paper.

Received: 17 August 2020; Accepted: 10 September 2020; Published: 13 September 2020

Abstract: In this paper, we study the information lost when a real-valued statistic is used to reduce or summarize sample data from a discrete random variable with a one-dimensional parameter. We compare the probability that a random sample gives a particular data set to the probability of the statistic's value for this data set. We focus on sufficient statistics for the parameter of interest and develop a general formula independent of the parameter for the Shannon information lost when a data sample is reduced to such a summary statistic. We also develop a measure of entropy for this lost information that depends only on the real-valued statistic but neither the parameter nor the data. Our approach would also work for non-sufficient statistics, but the lost information and associated entropy would involve the parameter. The method is applied to three well-known discrete distributions to illustrate its implementation.

Keywords: data reduction; Shannon information; entropy; information loss

1. Introduction

We consider the data sample $\mathbf{x} = (x_1, \ldots, x_n)$ from a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ for a discrete random variable X with sample space S and one-dimensional parameter θ. A statistic $T(\mathbf{X})$ is a function of the random sample $\mathbf{X}$ for any fixed but arbitrary value of a parameter θ associated with the underlying X. Thus a statistic $T(\mathbf{X})$ is a random variable itself. Here we consider only real-valued statistics that reduce the data sample $\mathbf{x}$ to a number $T(\mathbf{x})$ that might be used to summarize $\mathbf{X}$, to characteriz X, or perhaps to estimate θ. However, data reduction is an irreversible process [1] and always involves some information loss. For instance, if $T(\mathbf{X})$ is the sample mean $\overline{X}$, the original measurements $\mathbf{x}$ cannot be reconstructed from $\overline{x}$, and some information about $\mathbf{x}$ is lost. Nonetheless, such data reduction is frequently used to make inferences.

More explicitly, our motivation for considering such situations is that $T(\mathbf{x})$ is usually communicated in practice as a summary for the data $\mathbf{x}$ but without the actual data. The question then naturally arises: how much information is lost to someone about a data sample $\mathbf{x}$ when only the value of T is available for $\mathbf{x}$, but not the data itself? To answer this question, we develop a theoretical framework for determining how much information is lost about a given data set $\mathbf{x}$ by knowing only the value of $T(\mathbf{x})$ but neither $\mathbf{x}$ itself nor the parameter θ. Our information-theoretic approach to data reduction generalizes the observation in [2] that a binomial random variable loses all the information about the order of successes in the associated sequence of Bernoulli trials. In other words, for a series of n Bernoulli trials one cannot recreate the order of successes by only knowing the number that occurred.

For any real-valued statistic T and the given sample data **x**, we decompose the total information about **X** available in **x** into the sum of (a) the information available in the reduced data $T(\mathbf{x}) = \bar{\mathbf{x}}$ and (b) the information lost in the process of data reduction. When T is a sufficient statistic for θ, this lost information is independent of θ. Moreover, by taking the expected value of this lost information over all possible data sets, we define an associated entropy measure that depends on T but neither **x** nor θ. Our approach also works for non-sufficient statistics, but the lost information and associated entropy would then involve θ. Thus θ must be estimated before computing these quantities.

The paper is organized as follows. In Section 2, we present the necessary definitions, notation, and preliminary results. In Section 3, we decompose the total information available about **X** in **x** and give various expressions for the Shannon information lost by reducing **x** to $T(\mathbf{x})$. In Section 4, we develop an entropy measure associated with this lost information. In Section 5, we present examples of our results for some standard discrete distributions and several sufficient statistics for θ. Conclusions are offered in Section 6.

2. Preliminaries

Standard definitions, notation, and results to be used in our development are now presented for completeness and accessibility. In addition, some new definitions and results are established for subsequent use. **Definition 1**, **Result 1**, **Definition 2**, and **Definition 3** can be found in [3–5] and elsewhere. The notion of a sufficient statistic is first defined.

Definition 1 (Sufficient Statistic [3]). *A statistic $T(X)$ is a sufficient statistic (SS) for the parameter θ if the probability*

$$\mathrm{P}[\mathbf{X} = \mathbf{x} \big| \mathrm{T}(\mathbf{X}) = \mathrm{T}(\mathbf{x})] \tag{1}$$

is independent of θ.

Note that P instead of P_θ is used in (1) since this probability is independent of θ. In addition, observe that (1) is not a joint conditional probability distribution for **X** since its condition changes with **x**. This observation is significant in Section 4. The fact that (1) does not involve θ can be used to prove the Fisher Factorization Theorem (FFT) below, which is the usual method for determining if a statistic is an SS for θ. In the FFT, we use the notation $\mathrm{f}(\mathbf{x}|\theta)$ to denote the joint probability mass function (pmf) of **X** evaluated at the variable **x** for a fixed value of θ.

Result 1 (FFT [3]). *The real-valued statistic $T(X)$ is sufficient for θ if and only if there exist functions $g : R^1 \to R^1$ and $h : S^n \to R^1$ such that for any sample data x and for all values of θ the joint pmf $f(x|\theta)$ of X can be factored as*

$$\mathrm{f}(\mathbf{x}|\theta) = \mathrm{g}[\mathrm{T}(\mathbf{x})|\theta] \times \mathrm{h}(\mathbf{x}) \tag{2}$$

for real-valued, nonnegative functions g on and h on S^n. The function h does not depend on θ, while g does depend on x but only through $T(x)$.

We focus on a sufficient statistic T for θ in Section 3, where we need the notion of a partition [5] defined next.

Definition 2 (Partition [3]). *Let S be the denumerable sample space of the discrete random variable X so that S^n is the denumerable sample space of the random sample X. For any statistic $T : S^n \to R^1$, let τ_T be the denumerable set $\tau_T = \{t | \exists x \in S^n$ for which $t = T(x)\}$, which is the range of T. Then T partitions the sample space S^n into the mutually exclusive and collectively exhaustive partition sets $A_t = \{x \in S^n \big| T(x) = t\}$, $\forall t \in \tau_T$.*

Figure 1 illustrates **Definition 2**.

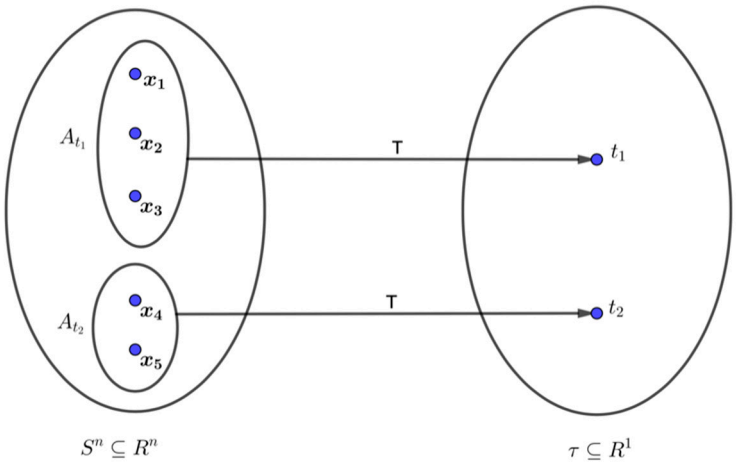

Figure 1. Partition Sets.

We also use the well-known likelihood function.

Definition 3 (Likelihood Function [3–5]). *Let x be sample data from a random sample X from a discrete random variable X with sample space S and real-valued parameter θ, and let $f(x|\theta)$ denote the joint pmf of the random sample X. For any sample data x, the likelihood function of θ is defined as*

$$\mathrm{L}(\theta|\mathbf{x}) = \mathrm{f}(\mathbf{x}|\theta). \tag{3}$$

The likelihood function $\mathrm{L}(\theta|\mathbf{x})$ in (3) is a function of the variable θ for given data $\mathbf{x}$. However, the joint pmf $\mathrm{f}(\mathbf{x}|\theta)$ as a function of $\mathbf{x}$ for fixed θ is frequently called the likelihood function as well. In this case, we also write the joint pmf as $\mathrm{L}(\mathbf{x}|\theta)$. We distinguish the two cases since $\mathrm{L}(\theta|\mathbf{x})$ is not a statistic but $\mathrm{L}(\mathbf{x}|\theta)$ is one that incorporates all available information about $\mathbf{X}$. Moreover, $\mathrm{L}(\mathbf{x}|\theta)$ is an SS for θ [4] and uniquely determines an associated SS called the likelihood kernel to be used in subsequent examples.

We next define a new concept called the likelihood kernel. As a function of $\mathbf{x}$ for fixed θ, it is shown below to be a sufficient statistic for θ and is used in Section 5 to facilitate the computation of lost information associated with other sufficient statistics T. As a function of θ for fixed $\mathbf{x}$, a possibility not considered here, the likelihood kernel may be useful in applying the likelihood principle [3,4] to make inferences about θ without resorting to the notion of equivalence classes. It would be the "simplest" factor of $\mathrm{L}(\theta|\mathbf{x})$ that can be used in a likelihood ratio comparing two values of θ.

Definition 4 (Likelihood kernel). *Let S be the sample space of X. For fixed but arbitrary θ, suppose that $L(x|\theta)$ can be factored as*

$$\mathrm{L}(\mathbf{x}|\theta) = \mathrm{K}(\mathbf{x}|\theta) \times \mathrm{R}(\mathbf{x}), \ \forall \mathbf{x} \in \mathrm{S}^{\mathrm{n}}, \tag{4}$$

where $K : S^n \to R^1$ and $R : S^n \to R^1$ have the following properties.

(a) Every nonnumerical factor of $K(\boldsymbol{x}|\theta)$ contains θ.
(b) $R(\boldsymbol{x})$ does not contain θ.
(c) For $\forall \boldsymbol{x} \in S^n$, both $K(\boldsymbol{x}|\theta) \geq 0$ and $R(\boldsymbol{x}) \geq 0$.
(d) $K(\boldsymbol{x}|\theta)$ is not divisible by any positive number except 1.

Then $K(\boldsymbol{x}|\theta)$ is defined as the likelihood kernel of $L(\boldsymbol{x}|\theta)$ and $R(\boldsymbol{x})$ as the residue of $L(\boldsymbol{x}|\theta)$.

Theorem 1. *The likelihood kernel $K(x|\theta)$ has the following properties.*

(i) $K(\boldsymbol{x}|\theta)$ *exists uniquely.*

(ii) $K(\boldsymbol{x}|\theta)$ *is an SS for* θ.

(iii) For any θ_1 *and* θ_2, *the likelihood ratio* $\frac{L(\boldsymbol{x}|\theta_1)}{L(\boldsymbol{x}|\theta_2)}$ *equals* $\frac{K(\boldsymbol{x}|\theta_1)}{K(\boldsymbol{x}|\theta_2)}$.

Proof. To prove (i), for fixed θ we first show that the likelihood kernel $K(\mathbf{x}|\theta)$ of **Definition 4** exists by construction. Since the formula for $L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$ must explicitly contain θ, the parameter θ cannot appear only in the range of $\mathbf{x}$. Hence $L(\mathbf{x}|\theta)$ as a function of $\mathbf{x}$ can be factored into $K(\mathbf{x}|\theta) \times R(\mathbf{x})$, satisfying (a) and (b) of **Definition 4**, where $K(\mathbf{x}|\theta) \geq 0,\ \forall \mathbf{x} \in S^n$, and the numerical factor of $K(\mathbf{x}|\theta)$ is either $+1$ or -1. Then $R(\mathbf{x}) \geq 0,\ \forall \mathbf{x} \in S^n$ since $K(\mathbf{x}|\theta) \geq 0,\ \forall \mathbf{x} \in S^n$, and $K(\mathbf{x}|\theta) \times R(\mathbf{x}) = f(\mathbf{x}|\theta) \geq 0$. Thus (c) is satisfied. Finally, the only positive integer that evenly divides $+1$ or -1 is 1, so (d) holds. It follows that the likelihood kernel $K(\mathbf{x}|\theta)$ and its associated $R(\mathbf{x})$ in **Definition 4** are well defined and exist.

We next show that $K(\mathbf{x}|\theta)$ as constructed above is unique. Let $K_1(\mathbf{x}|\theta)$ with residue $R_1(\mathbf{x})$ and $K_2(\mathbf{x}|\theta)$ with $R_2(\mathbf{x})$ both satisfy **Definition 4**. Thus for $j = 1, 2,\ R_j(\mathbf{x})$ does not contain θ, while every nonnumerical factor of $K_j(\mathbf{x}|\theta)$ does contain θ. It follows that $K_1(\mathbf{x}|\theta) \geq 0$ and $K_2(\mathbf{x}|\theta) \geq 0$ must be identical or else be a positive multiple of one another. Assume that $K_2(\mathbf{x}|\theta) = \lambda K_1(\mathbf{x}|\theta)$ for some $\lambda > 0$. If $\lambda \neq 1,\ K_2(\mathbf{x}|\theta)$ is divisible by a positive number other than 1 to contradict (d). Thus $K(\mathbf{x}|\theta)$ is unique.

To prove (ii), we show that this unique $K(\mathbf{x}|\theta)$ is an SS for θ. For $L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$, let $g[z] = z$ and $h(\mathbf{x}) = R(\mathbf{x})$ in (2). Then, $L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = g[K(\mathbf{x}|\theta)] \times h(\mathbf{x}) = K(\mathbf{x}|\theta) \times R(\mathbf{x})$. Thus $K(\mathbf{x}|\theta)$ is an SS by the FFT of **Result 1**. Finally, (iii) follows from **Definition 4** and the fact that the joint pmf $L(\mathbf{x}|\theta_2) \neq 0$ for $\mathbf{x} \in S^n$. □

We next discuss the notion of information to be used. Actually, probability itself is a measure of information in the sense that it captures the surprise level of an event. An observer obtains more information, i.e., surprise, if an unlikely event occurs than if a likely one does. Instead of probability, however, we use the additive measure known as Shannon information [6,7], defined as follows.

Definition 5 (Shannon Information [6,7]). *Let* $\boldsymbol{x}$ *be sample data for the random sample* $\boldsymbol{X}$ *from the discrete random variable* X *with a one-dimensional parameter* θ, *and let* $f(x|\theta)$ *be the joint pmf of* $\boldsymbol{X}$ *at* $\boldsymbol{x}$. *The Shannon information obtained from the sample data* $\boldsymbol{x}$ *is defined as*

$$I(\mathbf{x}|\theta) = -\log\ f(\mathbf{x}|\theta), \tag{5}$$

where the units of $I(\boldsymbol{x}|\theta)$ is bits if the base of the logarithm is 2, which is to be used here.

Other definitions for information have been proposed. For example, Vigo [8,9] has defined a measure of representational information. Further details on different types of information can be found in [10–16]. For Shannon information, we use its expected value over $\forall \mathbf{x} \in S^n$.

Definition 6 (Entropy [17–20]). *Under the conditions of* ***Definition 5****, the Shannon entropy* $H(\boldsymbol{X}|\theta)$ *is defined as the expected value of* $I(\boldsymbol{X}|\theta)$; *i.e.,*

$$H(\mathbf{X}|\theta) = \sum_{\mathbf{x}} f(\mathbf{x}|\theta) I(\mathbf{x}|\theta). \tag{6}$$

The general properties of Shannon entropy are given in [17–20], for example. Since entropy is the expected information over all possible random samples, it can be argued that entropy is a better measure of the available information about $\boldsymbol{X}$ *than would the Shannon information for a single data set* x, *which might not be typical [18]. We next give a method to obtain the information loss about* $\boldsymbol{X}$ *that occurs when a data set* $\boldsymbol{x}$ *is reduced to* $T(\boldsymbol{x})$. *In our approach, we focus on a sufficient statistic* T *so there will be no* θ *in (5) for the lost information below.*

3. Information Decomposition under Data Reduction by a Real-Valued Statistic

We now develop a procedure to determine how much information about $\mathbf{X}$ contained in a data set $\mathbf{x}$ is lost when the data is reduced to $T(\mathbf{x})$ by the sufficient statistic T. Consider the joint conditional probability

$$P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})], \tag{7}$$

which is identified with the probabilistic information lost about the event $\mathbf{X} = \mathbf{x}$ by the data reduction of $\mathbf{x}$ to $T(\mathbf{x})$. The notation P_θ refers to the fact that the discrete probability (7), in general, involves the parameter θ. We next express (7) using the definition of conditional probability to obtain the basis of our development. **Result 2** is given in ([3], p. 273) and proven below to illustrate the reasoning.

Result 2 [3]. *Let $\mathbf{x}$ be sample data for a random sample $\mathbf{X}$ from a discrete random variable X with sample space S and real-valued parameter θ, and let $T(\mathbf{X})$ be any real-valued statistic. Then*

$$P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})] = \frac{P_\theta[\mathbf{X} = \mathbf{x}]}{P_\theta[T(\mathbf{X}) = T(\mathbf{x})]}. \tag{8}$$

Proof. Using the definition of conditional probability, rewrite (7) as

$$\frac{P_\theta[\mathbf{X} = \mathbf{x}; T(\mathbf{X}) = T(\mathbf{x})]}{P_\theta[T(\mathbf{X}) = T(\mathbf{x})]}. \tag{9}$$

However, $T(\mathbf{X}) = T(\mathbf{x})$ whenever $\mathbf{X} = \mathbf{x}$, so (8) follows. □

Observe that if T is an SS for θ, the left side of (8) is independent of θ by the FFT and hence so is the right. Taking the negative logarithm of (8) and rearranging terms gives

$$-\log P_\theta[\mathbf{X} = \mathbf{x}] = -\log P_\theta[T(\mathbf{X}) = T(\mathbf{x})] - \log P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]. \tag{10}$$

From (8), note that $P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})] \geq P_\theta[\mathbf{X} = \mathbf{x}]$ since $P_\theta[T(\mathbf{X}) = T(\mathbf{x})] \leq 1$, so $-\log P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})] \leq -\log P_\theta[\mathbf{X} = \mathbf{x}]$. Similarly, $-\log P_\theta[T(\mathbf{X}) = T(\mathbf{x})] \leq -\log P_\theta[\mathbf{X} = \mathbf{x}]$. These facts suggest that the left side of (10) is the total Shannon information in bits about $\mathbf{X}$ contained in the sample data $\mathbf{x}$. On the right side of (10), the term $-\log P_\theta[T(\mathbf{X}) = T(\mathbf{x})]$ is considered the information about $\mathbf{X}$ contained in the reduced data summary $T(\mathbf{x})$, and the term $-\log P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]$ is identified as the information about $\mathbf{X}$ that has been lost as the result of the data reduction by $T(\mathbf{x})$.

In particular, this lost information represents a combinatorial loss in the sense that multiple $\mathbf{x}$'s may give the same value $T(\mathbf{x}) = t$, as depicted in Figure 1 above. In other words, the lost information $-\log P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]$ is a measure of the knowledge unavailable about the data sample $\mathbf{x}$ when only the reduced data summary $T(\mathbf{x})$ is known but not $\mathbf{x}$ itself. For a sufficient statistic $T(\mathbf{X})$ for θ, this lost information is independent of θ. It is a characteristic of $T(\mathbf{X})$ for the given data sample $\mathbf{x}$.

In terms of Figure 1, (10) may be described as follows. On the left of the figure is the sample space $S^n \subseteq R^n$ over which probabilities on $\mathbf{X}$ are computed. On the right is the range $\tau_T \subseteq R^1$ of T over which the probability of $T(\mathbf{X})$ are computed. T reduces the data sample $\mathbf{x}$ into $T(\mathbf{x})$, where multiple $\mathbf{x}$'s may give the same $T(\mathbf{x}) = t$. In Figure 1, the distinct data samples $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ are all reduced into the same value t_1. However, knowing that $T(\mathbf{x}) = t_1$ for some data sample $\mathbf{x}$ does not provide sufficient information to know unequivocally, for example, that $\mathbf{x} = \mathbf{x}_1$. Information is lost in the reduction. One can also say that the total information $-\log P_\theta[\mathbf{X} = \mathbf{x}]$ obtained from the left side of Figure 1 is reduced to $-\log P_\theta[T(\mathbf{X}) = T(\mathbf{x})]$ obtained from the right. The reduction in information from the left to the right side is precisely the lost information $-\log P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]$ of (10). For fixed t, it is lost due to the ambiguity as to which data sample on the left actually gave t. There is no such ambiguity when T is one-to-one.

The general decomposition of information in (10) is next summarized in **Definition 7**, where T does not need to be sufficient for θ.

Definition 7 ($\mathbf{I_{total}}$, $\mathbf{I_{reduced}}$, $\mathbf{I_{lost}}$). *Let $\boldsymbol{x}$ be sample data for a random sample $\boldsymbol{X}$ from a discrete random variable X with sample space S and real-valued parameter θ. For any real-valued statistic $T(\boldsymbol{X})$, the Shannon information about $\boldsymbol{X}$ obtained from the sample data $\boldsymbol{x}$ can be decomposed as*

$$\mathrm{I}_{\mathrm{total}}(\mathbf{x}|\theta) = \mathrm{I}_{\mathrm{reduced}}(\mathbf{x}|\theta, \mathrm{T}) + \mathrm{I}_{\mathrm{lost}}(\mathbf{x}|\theta, \mathrm{T}), \tag{11}$$

where

$$\mathrm{I}_{\mathrm{total}}(\mathbf{x}|\theta) = -\log \mathrm{P}_\theta[\mathbf{X} = \mathbf{x}] \tag{12}$$

$$\mathrm{I}_{\mathrm{reduced}}(\mathbf{x}|\theta, \mathrm{T}) = -\log \mathrm{P}_\theta[\mathrm{T}(\mathbf{X}) = \mathrm{T}(\mathbf{x})] \tag{13}$$

and

$$\mathrm{I}_{\mathrm{lost}}(\mathbf{x}|\theta, \mathrm{T}) = -\log \mathrm{P}_\theta[\mathbf{X} = \mathbf{x} \mid \mathrm{T}(\mathbf{X}) = \mathrm{T}(\mathbf{x})]. \tag{14}$$

Definition 7 *generalizes the information decomposition of [2] for a data sample $\boldsymbol{x}$ of size n from a Bernoulli random variable X. In terms of this paper, the parameter θ in [2] is the probability 0.5 of success on a single Bernoulli trial, and $T(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$. It should be noted that the notation I_{comp} in [2], which refers to compressed information, corresponds to $I_{reduced}$ in Equation (13). We use the term "data reduction" as described in [3] as opposed to "data compression" to prevent misinterpretation. In computer science, data compression refers to encoding information using fewer bits than the original representation and is often lossless.*

Both **Result 2** and **Definition 7** are valid for any real-valued statistic for **X**. The notation $\mathrm{I}_{\mathrm{total}}(\mathbf{x}|\theta)$ indicates that $\mathrm{I}_{\mathrm{total}}$ is a function of the sample data **x** for a fixed but arbitrary parameter value θ. Similarly, both $\mathrm{I}_{\mathrm{reduced}}(\mathbf{x}|\theta, \mathrm{T})$ and $\mathrm{I}_{\mathrm{lost}}(\mathbf{x}|\theta, \mathrm{T})$ are functions of **x** for fixed θ and T. However, in this paper we focus on sufficient statistics, which provide a simpler expression for $\mathrm{I}_{\mathrm{lost}}(\mathbf{x}|\theta, \mathrm{T})$ that does not involve θ. For a sufficient statistic T for θ, we use the notation $\mathrm{I}_{\mathrm{lost}}(\mathbf{x}|\mathrm{T})$ for the lost information, though $\mathrm{I}_{\mathrm{total}}(\mathbf{x}|\theta)$ and $\mathrm{I}_{\mathrm{reduced}}(\mathbf{x}|\theta, \mathrm{T})$ still require θ. The next result is an application of the FFT of **Result 1**.

Theorem 2 (Lost Information for an SS). *Let $\boldsymbol{x}$ be sample data for a random sample $\boldsymbol{X}$ from a discrete random variable X with sample space S and real-valued parameter θ. Let T be an SS for θ, let $f(\boldsymbol{x}|\theta)$ be the joint pmf of $\boldsymbol{X}$, and write $f(\boldsymbol{x}|\theta) = g[T(x)|\theta] \times h(x)$ as in **Result 1**. Then for all $\boldsymbol{x} \in S^n$*

$$\mathrm{I}_{\mathrm{lost}}(\mathbf{x}|\mathrm{T}) = -\log \frac{\mathrm{h}(\mathbf{x})}{\sum_{\mathbf{y} \epsilon \mathrm{A}_{\mathrm{T}(\mathbf{x})}} \mathrm{h}(\mathbf{y})}, \tag{15}$$

*where $A_{T(\boldsymbol{x})}$ is defined in **Definition 2** for $t = T(\boldsymbol{x})$.*

Proof. Let $\mathbf{x} \in \mathrm{S}^{\mathrm{n}}$. Then $\mathrm{f}(\mathbf{x}|\theta) > 0$ since **x** is a realization of **X**. Because T is an SS, we write (7) without θ. It now suffices to establish that

$$\mathrm{P}[\mathbf{X} = \mathbf{x}|\mathrm{T}(\mathbf{X}) = \mathrm{T}(\mathbf{x})] = \frac{\mathrm{h}(\mathbf{x})}{\sum_{\mathbf{y} \epsilon \mathrm{A}_{\mathrm{T}(\mathbf{x})}} \mathrm{h}(\mathbf{y})}, \tag{16}$$

from which (15) immediately follows. Rewrite (8) as

$$\mathrm{P}[\mathbf{X} = \mathbf{x}|\mathrm{T}(\mathbf{X}) = \mathrm{T}(\mathbf{x})] = \frac{\mathrm{P}_\theta[\mathbf{X} = \mathbf{x}]}{\mathrm{P}_\theta[\mathrm{T}(\mathbf{X}) = \mathrm{T}(\mathbf{x})]} = \frac{\mathrm{f}(\mathbf{x}|\theta)}{\sum_{\mathbf{y} \epsilon \mathrm{A}_{\mathrm{T}(\mathbf{x})}} \mathrm{f}(\mathbf{y}|\theta)}, \tag{17}$$

so from (17) and (2), then

$$P[\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})] = \frac{g[T(\mathbf{x})|\theta] \times h(\mathbf{x})}{\sum_{\mathbf{y}\epsilon A_{T(\mathbf{x})}} g[T(\mathbf{y})|\theta] \times h(\mathbf{y})}. \quad (18)$$

However, $T(\mathbf{y}) = T(\mathbf{x}),\ \forall \mathbf{y} \in A_{T(\mathbf{x})}$ in (18), so

$$P[\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})] = \frac{g[T(\mathbf{x})|\theta] \times h(\mathbf{x})}{g[T(\mathbf{x})|\theta] \times \sum_{\mathbf{y}\epsilon A_{T(\mathbf{x})}} h(\mathbf{y})},\ \forall \mathbf{x} \in S^n. \quad (19)$$

Since $f(\mathbf{x}|\theta) > 0$ and hence $g[T(\mathbf{x})|\theta] \neq 0$, this term can be canceled on the right side of (19) to yield (16). Taking the $-\log$ of (16) completes the proof. □

Now consider **Theorem 2** when each A_t is a singleton in (16), i.e., when T is a one-to-one function. In this extreme case, $P[\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})] = 1$ since $\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y}) = h(\mathbf{x})$ in the denominator of the right side of (16). Thus $I_{lost}(\mathbf{x}|T) = 0$ from which $I_{comp}(\mathbf{x}|\theta, T) = I_{total}(\mathbf{x}|\theta)$ for all $\mathbf{x}$ in S^n. Thus, the special case of a one-to-one T justifies the identification of the lost information as $I_{lost}(\mathbf{x}|\theta, T) = -\log P_\theta[\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})]$. In other words, for all data samples $\mathbf{x},\ \mathbf{y} \in S^n$, if $\mathbf{x} \neq \mathbf{y}$ whenever $T(\mathbf{x}) \neq T(\mathbf{y})$, then $P[\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})]$ is not diminished by the reduction of the singleton $A_{T(\mathbf{x})}$ to the number $T(\mathbf{x})$.

More generally, it is also true that $I_{lost}(\mathbf{x}|\theta, T) = 0$ when T is one-to-one but not sufficient for θ. In this case, write $P_\theta[\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})] = \frac{P_\theta[\mathbf{X}=\mathbf{x}]}{P_\theta[T(\mathbf{X})=T(\mathbf{x})]} = \frac{f(\mathbf{x}|\theta)}{\sum_{\mathbf{y}\epsilon A_{T(\mathbf{x})}} f(\mathbf{y}|\theta)}$. However, since T is one-to-one, then $\sum_{\mathbf{y}\epsilon A_{T(\mathbf{x})}} f(\mathbf{y}|\theta) = f(\mathbf{x}|\theta)$, $P_\theta[\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})] = 1$, and again $I_{lost}(\mathbf{x}|\theta, T) = 0$.

Now consider the other extreme case where $T(\mathbf{x}) = c$ is constant on S^n. Thus $P_\theta[\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = c] = \frac{P_\theta[\mathbf{X}=\mathbf{x}]}{P_\theta[T(\mathbf{X})=c]}$. However, $P_\theta[T(\mathbf{X}) = c] = 1$, so $P_\theta[\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = c] = P_\theta[\mathbf{X} = \mathbf{x}]$ and $I_{lost}(\mathbf{x}|\theta, T) = I_{total}(\mathbf{x}|\theta, T)$ on S^n. In this case, $I_{reduced}(\mathbf{x}|\theta, T) = 0$ because the event $T(\mathbf{x}) = c$ gives no information about $\mathbf{x}$.

We also note that $I_{lost}(\mathbf{x}|T)$ could be used as a metric to compare sufficient statistics for a given data sample $\mathbf{x}$. For example, T_1 could be regarded as better than T_2 for $\mathbf{x}$ if $I_{lost}(\mathbf{x}, T_1) < I_{lost}(\mathbf{x}, T_2)$. However, this comparison would be limited to the given $\mathbf{x}$. In Section 4, we propose but do not explore a metric based on entropy independent of a particular data sample. We next show that (16) can be simplified when T is the likelihood function.

Corollary 1 (Information Loss for Likelihood Function). *Under the assumptions of Theorem 2, if* $T(\boldsymbol{x}) = L(\boldsymbol{x}|\theta)$, then

$$I_{lost}(\mathbf{x}|L) = -\log \frac{1}{\left|A_{L(\mathbf{x}|\theta)}\right|}, \quad (20)$$

where $\left|A_{L(\boldsymbol{x}|\theta)}\right|$ *is the cardinality of the partition set* A_t *for* $t = L(\boldsymbol{x}|\theta)$.

Proof. For $T(\mathbf{x}) = L(\mathbf{x}|\theta) = f(\mathbf{x}|\theta)$ in (2), let g be the identity function and $h(\mathbf{x}) = 1$. Then substituting $h(\mathbf{x}) = 1$ into (16) gives the denominator $\sum_{\mathbf{y}\epsilon A_{L(\mathbf{x}|\theta)}} 1 = \left|A_{L(\mathbf{x}|\theta)}\right|$ to yield (20). □

We next state a reproductive property of a statistic T′ that is a one-to-one function of a sufficient statistic T for θ.

Theorem 3. *If there is a one-to-one function between a sufficient statistic T for* θ *and an arbitrary real-valued statistic T′ on* S^n*, then the following hold.*

(i) *T′ is also an SS.*
(ii) *T and T′ partition the sample space S* into the same partition sets.
(iii) $I_{lost}(x|T) = I_{lost}(x|T'), \forall x \in S^n$.

Proof. To prove (i), let u be a real-valued one-to-one function of T′ such that

$$\mathrm{T}(\mathbf{x}) = \mathrm{u}[\mathrm{T}'(\mathbf{x})]. \tag{21}$$

Since T is an SS, by Equation (2) there are real-valued functions g on R^1 and h on S^n for which

$$\mathrm{f}(\mathbf{x}|\theta) = \mathrm{g}[\mathrm{T}(\mathbf{x})|\theta] \times \mathrm{h}(\mathbf{x}). \tag{22}$$

By substituting $\mathrm{T}(\mathbf{x})$ from (21) in (22), we get

$$\mathrm{f}(\mathbf{x}|\theta) = \mathrm{g}(\mathrm{u}[\mathrm{T}'(\mathbf{x})]|\theta) \times \mathrm{h}(\mathbf{x}), \tag{23}$$

which can be rewritten as

$$\mathrm{f}(\mathbf{x}|\theta) = (\mathrm{g} \circ \mathrm{u})[\mathrm{T}'(\mathbf{x})|\theta] \times \mathrm{h}(\mathbf{x}). \tag{24}$$

Since T′ in (24) satisfies the condition of **Result 1** for $\mathrm{g}' = \mathrm{g} \circ \mathrm{u}$, T′ is an SS.

To prove (ii), we use **Definition 2**. Let T partition the sample space S^n into the mutually exclusive and collectively exhaustive sets $\mathrm{A_t} = \{\mathbf{x}|\mathrm{T}(\mathbf{x}) = \mathrm{t}\}, \ \forall \mathrm{t} \in \tau_\mathrm{T}$. By Equation (21) we can also write $\mathrm{A_t}$ as

$$\mathrm{A_t} = \{\mathbf{x}|\mathrm{u}[\mathrm{T}'(\mathbf{x})] = \mathrm{t}\}, \ \forall \mathrm{t} \in \tau_\mathrm{T}. \tag{25}$$

Since u is a one-to-one function, it has an inverse u^{-1}. Letting $\mathrm{u}^{-1}(\mathrm{t}) = \mathrm{t}'$, we apply u^{-1} to the right side of (25) and get

$$\mathrm{A_t} = \left\{\mathbf{x}\middle|\mathrm{T}'(\mathbf{x}) = \mathrm{t}'\right\}, \forall \mathrm{t}' \in \mathrm{u}(\tau_\mathrm{T}). \tag{26}$$

However, $\mathrm{u}(\tau_\mathrm{T}) = \tau_{\mathrm{T}'}$ and the cardinalities satisfy $\left|\tau_\mathrm{T}\right| = \left|\tau_{\mathrm{T}'}\right|$, so the right side of (26) is $\mathrm{A_{t'}}$ and

$$\mathrm{A_t} = \mathrm{A_{t'}}. \tag{27}$$

Finally, to get (iii) we use **Theorem 2** to calculate information lost over two statistics T and T′. Since $\mathrm{h}(\mathbf{x})$ is the same in (22) and (24) and since Equation (27) holds, we sum $\mathrm{h}(\mathbf{x})$ over the same sets in the denominator of Equation (16) for both T and T′ to give

$$\mathrm{I_{lost}}(\mathbf{x}|\mathrm{T}) = \mathrm{I_{lost}}\left(\mathbf{x}\middle|\mathrm{T}'\right) \tag{28}$$

and complete the proof. □

We next compare the information loss of the sufficient statistic $\mathrm{L}(\mathbf{x}|\theta)$ to other sufficient statistics. For the sufficient statistic $\mathrm{K}(\mathbf{x}|\theta)$, a lemma is needed.

Lemma 1. *Let* $\boldsymbol{x}$ *be any data sample for a random sample* $\boldsymbol{X}$ *from the discrete random variable X with real-valued parameter* θ. *Then* $K(\boldsymbol{x}|\theta)$ *is a function of* $L(\boldsymbol{x}|\theta)$ *and* $\tau_L \geq \tau_K$.

Proof. From ([3], p. 280), $\mathrm{K}(\mathbf{x}|\theta)$ is a function of $\mathrm{L}(\mathbf{x}|\theta)$ if and only if $\mathrm{K}(\mathbf{x}|\theta) = \mathrm{K}(\mathbf{y}|\theta)$ whenever $\mathrm{L}(\mathbf{x}|\theta) = \mathrm{L}(\mathbf{y}|\theta)$. For all data samples $\mathbf{x}$ and $\mathbf{y}$, we prove that if $\mathrm{L}(\mathbf{x}|\theta) = \mathrm{L}(\mathbf{y}|\theta)$, then $\mathrm{K}(\mathbf{x}|\theta) = \mathrm{K}(\mathbf{y}|\theta)$. Thus suppose that $\mathrm{L}(\mathbf{x}|\theta) = \mathrm{L}(\mathbf{y}|\theta)$. By **Definition 4**, we can decompose $\mathrm{L}(\mathbf{x}|\theta)$ and $\mathrm{L}(\mathbf{y}|\theta)$ into $\mathrm{K}(\mathbf{x}|\theta)\mathrm{R}(\mathbf{x})$ and $\mathrm{K}(\mathbf{y}|\theta)\mathrm{R}(\mathbf{y})$, respectively. Note that $\mathrm{K}(\mathbf{y}|\theta) \neq 0$. Otherwise, $\mathrm{L}(\mathbf{y}|\theta) = 0$ in contradiction to $\mathbf{y}$ being sample data with a nonzero probability of occurring. Now write

$$\frac{K(\mathbf{x}|\theta)}{K(\mathbf{y}\big|\theta)} = \frac{R(\mathbf{y})}{R(\mathbf{x})}. \tag{29}$$

Suppose that $K(\mathbf{x}|\theta) \neq K(\mathbf{y}\big|\theta)$ so that $\frac{K(\mathbf{x}|\theta)}{K(\mathbf{y}|\theta)} = \frac{R(\mathbf{y})}{R(\mathbf{x})} \neq 1$ in (29). From **Definition 4**, every nonnumerical factor of $K(\mathbf{x}|\theta)$ and $K(\mathbf{y}\big|\theta)$ contains θ. Moreover, neither $K(\mathbf{x}|\theta)$ nor $K(\mathbf{y}\big|\theta)$ is divisible by any positive number except the number 1. Hence, since $\frac{R(\mathbf{y})}{R(\mathbf{x})}$ does not contain θ, the nonnumerical factors of $K(\mathbf{x}|\theta)$ and $K(\mathbf{y}\big|\theta)$ must cancel in (29) and the remaining numerical factors could not be identical. Thus at least one of these factors would be divisible by a positive number other than 1 in contradiction to **Definition 4**. It now follows that $K(\mathbf{x}|\theta) = K(\mathbf{y}\big|\theta)$, so $K(\mathbf{x}|\theta)$ is some function u of $L(\mathbf{x}|\theta)$. Finally, $\tau_L \geq \tau_K$ since this function u is surjective from S^n onto its image $u(S^n)$. □

Lemma 2. *Under the conditions of* ***Lemma 1****, the sufficient statistics L and K satisfy*

$$I_{\text{reduced}}(\mathbf{x}|\theta, L) \geq I_{\text{reduced}}(\mathbf{x}|\theta, K), \ \forall \mathbf{x} \in S^n. \tag{30}$$

Proof. Let $\mathbf{x} \in S^n$ and suppose that $\mathbf{y} \in A_{L(\mathbf{x})}$. Then $L(\mathbf{y}\big|\theta) = L(\mathbf{x}|\theta)$, so it follows from **Lemma 1** that $K(\mathbf{y}\big|\theta) = K(\mathbf{x}|\theta)$ and thus $\mathbf{y} \in A_{K(\mathbf{x})}$. Hence $A_{L(\mathbf{x})} \subseteq A_{K(\mathbf{x})}$, and so

$$P_\theta[L(\mathbf{X}|\theta) = L(\mathbf{x}|\theta)] = \sum_{\mathbf{y}\epsilon A_{L(\mathbf{x})}} f(\mathbf{x}|\theta) \leq \sum_{\mathbf{y}\epsilon A_{K(\mathbf{x})}} f(\mathbf{x}|\theta) = P_\theta[K(\mathbf{X}|\theta) = K(\mathbf{x}|\theta)], \forall \mathbf{x} \in S^n. \tag{31}$$

Taking the negative log of both sides of the inequality in (31) and using (13) gives (30). □

Theorem 4. *Let* $\boldsymbol{x}$ *be sample data for a random sample* $\boldsymbol{X}$ *from a discrete random variable X with the real-valued parameter* θ*. Then for all* $\boldsymbol{x} \in S^n$*,*

$$I_{\text{lost}}(\mathbf{x}|L) \leq I_{\text{lost}}(\mathbf{x}|K). \tag{32}$$

Proof. Let $\mathbf{x} \in S^n$. Note that $I_{\text{total}}(\mathbf{x}|\theta)$ in (12) does not depend on the arbitrary sufficient statistic T of (11). Hence

$$I_{\text{total}}(\mathbf{x}|\theta) = I_{\text{reduced}}(\mathbf{x}|\theta, L) + I_{\text{lost}}(\mathbf{x}|L) = I_{\text{reduced}}(\mathbf{x}|\theta, K) + I_{\text{lost}}(\mathbf{x}|K). \tag{33}$$

Then (32) follows immediately from (30) and (33). □

As a consequence of **Theorem 3**, **Theorem 4** has an immediate corollary.

Corollary 2. *Under the conditions of* ***Theorem 4****, let T be a sufficient statistic for* θ *for which there is a one-to-one function between T and K. Then for all* $\boldsymbol{x} \in S^n$*,*

$$I_{\text{lost}}(\mathbf{x}|L) \leq I_{\text{lost}}(\mathbf{x}|T). \tag{34}$$

The question remains open as to whether (34) holds for all sufficient statistics T for θ. Regardless, the proofs of **Lemma 2** and **Theorem 4** illustrate the fact that the relation between the lost information for two statistics T and T′ is determined by the relation between their partition sets $A_t = \{\mathbf{x}|T(\mathbf{x}) = t\}$ and $B_{t'} = \left\{\mathbf{x}\middle|T'(\mathbf{x}) = t'\right\}$. For example, if for every A_t there exists a $B_{t'}$ for which $A_t \subset B_{t'}$, then the partition of S^n by the $B_{t'}$ of T′ is said to be coarser than the partition by the A_t of T. In that case, $I_{\text{lost}}(\mathbf{x}|\theta, T) \leq I_{\text{lost}}(\mathbf{x}|\theta, T')$ because each $\mathbf{x} \in S^n$ has more $\mathbf{y} \in S^n$ with $T'(\mathbf{y}) = T'(\mathbf{x})$ than there are with

$T(\mathbf{y}) = T(\mathbf{x})$. In other words, $T'(\mathbf{y}) = t'$ is at least as ambiguous as $T(\mathbf{y}) = t$ in determining the data sample giving the value of the respective statistics.

4. Entropic Loss for an SS

For a sufficient statistic T for θ, we now propose an entropy measure to characterize T by the expected lost information incurred by the reduction of $\mathbf{X}$ to $T(\mathbf{X})$. This expectation is taken over all possible data sets $\mathbf{x}$. This nonstandard entropy measure is called entropic loss, and it depends on neither a particular data set $\mathbf{x}$ nor the value of θ. Before defining this measure, we need to determine the appropriate pmf to use in taking an expectation. The following results are used.

Result 3. *Under the assumptions of* ***Theorem 2****, for any data sample let* $t = T(\boldsymbol{x})$ *and consider the partition set* A_t. Then

$$\sum_{\mathbf{x}\in A_t} P[\mathbf{X} = \mathbf{x} \,\big|\, T(\mathbf{X}) = t] = 1. \tag{35}$$

Proof. Summing (16) over $\mathbf{x} \in A_t$ yields

$$\sum_{\mathbf{x}\in A_t} P[\mathbf{X} = \mathbf{x} \,\big|\, T(\mathbf{X}) = t] = \frac{\sum_{\mathbf{x}\epsilon A_{T(\mathbf{x})}} h(\mathbf{x})}{\sum_{\mathbf{y}\epsilon A_{T(\mathbf{x})}} h(\mathbf{y})} = 1. \tag{36}$$

to give (35). □

Result 4. *Under the assumptions of* ***Theorem 2****, the sum*

$$\sum_{\mathbf{x}\in S^n} P[\mathbf{X} = \mathbf{x} \,|\, T(\mathbf{X}) = T(\mathbf{x})] = |\tau_T|. \tag{37}$$

Proof. We perform the sum on the left of (37) by first summing over $\mathbf{x} \in A_t$ for fixed t and then summing over each $t \in \tau_T$ to give

$$\sum_{\mathbf{x}\in S^n} P[\mathbf{X} = \mathbf{x} \,|\, T(\mathbf{X}) = T(\mathbf{x})] = \sum_{t\in\tau_T} \sum_{\mathbf{x}\in A_t} P[\mathbf{X} = \mathbf{x} \,\big|\, T(\mathbf{X}) = t]. \tag{38}$$

The inner series on the right side of (38) sums to one by **Result 3**. Hence, the outer sum yields $|\tau_T|$ for $\tau_T = \{t | \exists \mathbf{x} \in S^n$ for which $t = T(\mathbf{x})\}$. □

From (37), it follows that the left side of (37) is not a probability distribution on S^n unless $|\tau_T| = 1$. Moreover, $P[\mathbf{X} = \mathbf{x} \,|\, T(\mathbf{X}) = T(\mathbf{x})]$ is not a conditional probability distribution even if $|\tau_T| = 1$ since the condition $T(\mathbf{X}) = T(\mathbf{x})$ varies with $\mathbf{x}$. However, we use **Result 4** to normalize $P[\mathbf{X} = \mathbf{x} \,|\, T(\mathbf{X}) = T(\mathbf{x})]$ and obtain the appropriate pmf for calculating the expectation of $I_{lost}(\mathbf{X}|T)$.

Definition 8 (Entropic Loss). *Under the assumptions of* ***Theorem 2****, the entropic loss resulting from the data reduction by T is defined as*

$$H_{lost}(\mathbf{X}, T) = \frac{-1}{|\tau_T|} \sum_{\mathbf{x}\in S^n} P[\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})] \log P[\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})], \tag{39}$$

which from (15) and (16) can be rewritten as

$$H_{lost}(\mathbf{X}, T) = \frac{-1}{|\tau_T|} \sum_{\mathbf{x}\in S^n} \frac{h(\mathbf{x})}{\sum_{\mathbf{y}\epsilon A_{T(\mathbf{x})}} h(\mathbf{y})} \log \frac{h(\mathbf{x})}{\sum_{\mathbf{y}\epsilon A_{T(\mathbf{x})}} h(\mathbf{y})}. \tag{40}$$

Observe that (39) and (40) are independent of both $\mathbf{x}$ and θ. Indeed, for a given underlying random variable X and sample size n, $H_{lost}(\mathbf{X}, T)$ is a function only of T. Thus, $H_{lost}(\mathbf{X}, T)$ could be used as a metric to compare sufficient statistics independent of the data sample. In particular, T_1 could be regarded as better than T_2 if $H_{lost}(\mathbf{X}, T_1) < H_{lost}(\mathbf{X}, T_2)$, i.e., if the expected information loss associated with T_1 is less than that for T_2. Moreover, for a given underlying random variable X and sample size n, **Definition 8** could be extended to non-sufficient statistics. In that case, the entropic loss $H_{lost}(\mathbf{X}, T, \theta)$ would be a function of both T and θ. For a fixed θ, a non-sufficient statistic T_1 could again be considered as better than a non-sufficient T_2 if $H_{lost}(\mathbf{X}, T_1, \theta\ ) < H_{lost}(\mathbf{X}, T, \theta\ )$. Furthermore, for a given statistic T, the numerical value θ_1 could be considered as a better numerical point estimate for θ than the value θ_2 if $H_{lost}(\mathbf{X}, T, \theta_1\ ) < H_{lost}(\mathbf{X}, T, \theta_2\ )$. Similarly, $H_{lost}(\mathbf{X}, T, \theta)$ could be minimized over θ to give a best numerical point estimate for θ based on the entropic loss criterion. However, we do not pursue these possibilities here. We next compute $H_{lost}(T)$ for the sufficient statistic $T(\mathbf{X}) = L(\mathbf{X}|\theta)$.

Theorem 5 (Entropic Loss for Likelihood Function). *Under the assumptions of* ***Theorem*** **2**, *the entropic loss resulting from the data reduction by* $T(x) = L(x|\theta)$ is

$$H_{lost}(\mathbf{X}, L) = \frac{-1}{|\tau_L|} \sum_{t \in \tau_L} \log \frac{1}{|A_t|}. \tag{41}$$

Proof. From (20), write

$$H_{lost}(\mathbf{X}, L) = \frac{-1}{|\tau_L|} \sum_{\mathbf{x} \in S^n} \frac{1}{\left|A_{L(\mathbf{x})}\right|} \log \frac{1}{\left|A_{L(\mathbf{x})}\right|}. \tag{42}$$

We decompose the sum over $\mathbf{x} \in S^n$ in (42) to consecutive sums over $\mathbf{x} \in A_t$ and then $t \in \tau_T$ to get

$$H_{lost}(\mathbf{X}, L) = \frac{-1}{|\tau_L|} \sum_{t \in \tau_L} \sum_{\mathbf{x} \in A_t} \frac{1}{|A_t|} \log \frac{1}{|A_t|} = \frac{-1}{|\tau_L|} \sum_{t \in \tau_L} \frac{|A_t|}{|A_t|} \log \frac{1}{|A_t|}. \tag{43}$$

Equation (41) now follows from (43). □

Result 5. *Suppose there is a one-to-one function between two sufficient statistics T and T′ for* θ. Then

$$H_{lost}(\mathbf{X}, T) = H_{lost}(\mathbf{X}, T'). \tag{44}$$

Proof. For all $\mathbf{x} \in S^n$, $I_{lost}(\mathbf{x}|T\ ) = I_{lost}(\mathbf{x}|T'\ )$ from **Theorem 3**, so

$$-\log \frac{h(\mathbf{x})}{\sum_{\mathbf{y} \epsilon A_{T(\mathbf{x})}} h(\mathbf{y})} = -\log \frac{h(\mathbf{x})}{\sum_{\mathbf{y} \epsilon A_{T'(\mathbf{x})}} h(\mathbf{y})}, \tag{45}$$

from which

$$\frac{h(\mathbf{x})}{\sum_{\mathbf{y} \epsilon A_{T(\mathbf{x})}} h(\mathbf{y})} = \frac{h(\mathbf{x})}{\sum_{\mathbf{y} \epsilon A_{T'(\mathbf{x})}} h(\mathbf{y})}. \tag{46}$$

Thus from (45) and (46),

$$\frac{h(\mathbf{x})}{\sum_{\mathbf{y} \epsilon A_{T(\mathbf{x})}} h(\mathbf{y})} \log \frac{h(\mathbf{x})}{\sum_{\mathbf{y} \epsilon A_{T(\mathbf{x})}} h(\mathbf{y})} = \frac{h(\mathbf{x})}{\sum_{\mathbf{y} \epsilon A_{T'(\mathbf{x})}} h(\mathbf{y})} \log \frac{h(\mathbf{x})}{\sum_{\mathbf{y} \epsilon A_{T'(\mathbf{x})}} h(\mathbf{y})}. \tag{47}$$

Now summing (47) over $\mathbf{x} \in S^n$ yields

$$\sum_{\mathbf{x} \in S^n} \frac{h(\mathbf{x})}{\sum_{\mathbf{y} \epsilon A_{T(\mathbf{x})}} h(\mathbf{y})} \log \frac{h(\mathbf{x})}{\sum_{\mathbf{y} \epsilon A_{T(\mathbf{x})}} h(\mathbf{y})} = \sum_{\mathbf{x} \in S^n} \frac{h(\mathbf{x})}{\sum_{\mathbf{y} \epsilon T'(\mathbf{x})} h(\mathbf{y})} \log \frac{h(\mathbf{x})}{\sum_{\mathbf{y} \epsilon A_{T'(\mathbf{x})}} h(\mathbf{y})}\ . \tag{48}$$

However, from **Theorem 3**, $|\tau_T| = |\tau_{T'}|$. Thus dividing the left side of (48) by $-|\tau_T|$ and the right side by $-|\tau_{T'}|$ yields (44). □

5. Examples and Computational Issues

In this section, we present examples involving the discrete Poisson, binomial, and geometric distributions [21]. For each distribution, three sufficient statistics for a parameter θ are analyzed. For each such T, the quantity $I_{lost}(\mathbf{x}|T)$ does not involve θ. However, calculating $I_{lost}(\mathbf{x}|T)$ can still present computational issues, some of which are discussed below. Our examples are simple in order to focus on the definitions and results of Sections 3 and 4.

Example 1 (Poisson Distribution). *Consider the random sample* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *with the data sample* $\boldsymbol{x} = (x_1, \ldots, x_n)$ *from a Poisson random variable* X. *We consider three sufficient statistics for the parameter* $\theta > 0$. *These sufficient statistics are* $T_1(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$, *the likelihood kernel* $T_2(\boldsymbol{X}) = K(\boldsymbol{X}|\theta)$ *for fixed but arbitrary* θ, *and the likelihood function* $T_3(\boldsymbol{X}) = L(\boldsymbol{X}|\theta)$ *for fixed but arbitrary* θ. *We use* $T_1(\boldsymbol{X})$ *as a surrogate for* $T_1'(\boldsymbol{X}) = \frac{\sum_{i=1}^{n} X_i}{n}$. *Neither* $T_1(\boldsymbol{X})$ *or* $T_1'(\boldsymbol{X})$ *involves* θ *and can thus be used either to characterize* $\boldsymbol{X}$ *or to estimate* θ. *Moreover, there is an obvious one-to-one function relating* $\frac{\sum_{i=1}^{n} X_i}{n}$ *and* $\sum_{i=1}^{n} X_i$, *so* ***Theorems 3*** *and* ***5*** *establish that* $I_{lost}\left(\boldsymbol{x}\middle|T_1'\right) = I_{lost}(\boldsymbol{x}|T_1)$ *and* $H_{lost}\left(\boldsymbol{X}, T_1'\right) = H_{lost}(\boldsymbol{X}, T_1)$, *respectively. We analyze* $T_1(\boldsymbol{X})$ *because it is also Poisson, whereas* $T_1'(\boldsymbol{X})$ *is not Poisson since* $\frac{\sum_{i=1}^{n} X_i}{n}$ *is notnecessarily a nonnegative integer. In contrast to* $T_1(\boldsymbol{X})$, *both* $T_2(\boldsymbol{X})$ *and* $T_3(\boldsymbol{X})$ *contain* θ *and can only be used to characterize* $\boldsymbol{X}$. *For each of these three sufficient statistics, we develop an expression for* $I_{lost}(\boldsymbol{x}|T)$ *and describe how to obtain a numerical value. We then illustrate previous results with a realistic Poisson data sample. We present further computational results in Table 1.*

Table 1. Poisson Example.

$\mathbf{x} = (x_1, x_2, x_3)$	$T_1(\mathbf{x})$	$I_{lost}(\mathbf{x}\|T_1)$	$T_2(\mathbf{x})$	$I_{lost}(\mathbf{x}\|T_2)$	$T_3(\mathbf{x})$	$I_{lost}(\mathbf{x}\|T_3)$
(0,0,0)	0	0	$e^{-3\theta}$	0	$e^{-3\theta}$	0
(0,0,1)	1	$\log 3$	$\theta e^{-3\theta}$	$\log 3$	$\theta e^{-3\theta}$	$\log 3$
(0,1,0)						
(1,0,0)						
(1,1,0)	2	$\log \frac{9}{2}$	$\theta^2 e^{-3\theta}$	$\log \frac{9}{2}$	$\theta^2 e^{-3\theta}$	$\log 3$
(1,0,1)						
(0,1,1)						
(2,0,0)	2	$\log 9$	$\theta^2 e^{-3\theta}$	$\log 9$	$\frac{\theta^2 e^{-3\theta}}{2}$	$\log 3$
(0,2,0)						
(0,0,2)						

Case 1: Let $T_1(\mathbf{X}) = \sum_{i=1}^{n} X_i$. Observe that $T_1(\mathbf{X})$ is a sufficient statistic for θ from **Result 1** since $f(\mathbf{x}|\theta) = P_\theta[\mathbf{X} = \mathbf{x}] = \frac{\theta^{\sum_{i=1}^{n} x_i} e^{-n\theta}}{\prod_{i=1}^{n} x_i!}$ can be factored in (2) into the functions $g[T_1(\mathbf{x})|\theta] = \theta^{\sum_{i=1}^{n} x_i} e^{-n\theta}$ and $h(\mathbf{x}) = \frac{1}{\prod_{i=1}^{n} x_i!}$. Next recall that the statistic $\sum_{i=1}^{n} X_i$ has a Poisson distribution with parameter $n\theta$ [21]. Thus, $P_\theta\left[\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} x_i\right] = \frac{(n\theta)^{\sum_{i=1}^{n} x_i} e^{-n\theta}}{(\sum_{i=1}^{n} x_i)!}$, and so (8) becomes

$$P[\mathbf{X} = \mathbf{x} | \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} x_i] = \frac{1}{n^{\sum_{i=1}^{n} x_i}} \left(\begin{array}{c} \sum_{i=1}^{n} x_i \\ x_1, \ldots, x_n \end{array} \right), \tag{49}$$

where the multinomial coefficient $\left(\begin{array}{c} \sum_{i=1}^{n} x_i \\ x_1, \ldots, x_n \end{array} \right) = \frac{(\sum_{i=1}^{n} x_i)!}{\prod_{i=1}^{n} x_i!}$. It follows from (49) and (10) that

$$I_{lost}(\mathbf{x}|T_1) = -\log \left(\begin{array}{c} \sum_{i=1}^{n} x_i \\ x_1, \ldots, x_n \end{array} \right) + (\log n) \sum_{i=1}^{n} x_i, \tag{50}$$

which is also $I_{lost}\left(\mathbf{x}\middle|T_1'\right)$, as noted above.

For a data sample $(x_1, \ldots, x_n)$, the evaluation of $I_{lost}(\mathbf{x}|T_1)$ in (50) involves computing factorials [22]. For realistic data, the principal limitation to calculating them by direct multiplication is their magnitude. See [23] for a discussion. However, (50) can be approximated using either the well-known Stirling formula or the more accurate Ramanujan approximation [24]. The online multinomial coefficient calculator [25] can evaluate multinomial coefficients for when all x_i as well as n are less than approximately 50 if any $x_i = 0$ is removed from $\left(\begin{array}{c} \sum_{i=1}^{n} x_i \\ x_1, \ldots, x_n \end{array} \right)$. Such deletions do not affect the calculation since $0! = 1$.

As a numerical example, consider a data sample $\mathbf{x}$ of size $n = 34$ from a Poisson random variable X with $\theta = 3$, where

$$\mathbf{x} = (4,\ 7,\ 1,\ 3,\ 4,\ 2,\ 5,\ 0,\ 1,\ 2,\ 3,\ 6,\ 8,\ 0,\ 1,\ 2,\ 4,\ 9,\ 0,\ 2,\ 3,\ 1,\ 4,\ 2,\ 0,\ 1,\ 5,\ 6,\ 2,\ 7,\ 0,\ 1,\ 4,\ 2). \tag{51}$$

Then, $T_1(\mathbf{x}) = \sum_{i=1}^{n} x_i = 102$ from (51), and the calculator at [25] gives $\left(\begin{array}{c} \sum_{i=1}^{n} x_i \\ x_1, \ldots, x_n \end{array} \right) \approx 1.574 \times 10^{123}$ in (49) and (50). Moreover, $(\log n) \sum_{i=1}^{n} x_i = 518.915$. Hence, from (50), $I_{lost}(\mathbf{x}|T_1) = I_{lost}(\mathbf{x}|T_1') \approx 109.667$ bits. This value corresponds to 13.708 bytes at 8 bits per byte or to 0.013 kilobytes (KB) at 1024 bytes per KB. It follows from previous discussion in this example that the Shannon information lost by using the sample mean T_1' as a surrogate for $\mathbf{x}$ itself is $I_{lost}(\mathbf{x}|T_1') = I_{lost}(\mathbf{x}|T_1) \approx 0.013$ KB, which seems surprisingly small. Perhaps the small loss results partially from the fact that $T_1'(\mathbf{x}) = \bar{x} = \theta = 3$ exactly.

Case 2: Let $T_2(\mathbf{X}) = K(\mathbf{X}|\theta)$ for fixed but arbitrary $\theta > 0$. For a data sample $(x_1, \ldots, x_n)$, write

$$L(\mathbf{x}|\theta) = f(\mathbf{x}|\theta) = \frac{\theta^{\sum_{i=1}^{n} x_i} e^{-n\theta}}{\prod_{i=1}^{n} x_i!}, \tag{52}$$

from which

$$K(\mathbf{x}|\theta) = \theta^{\sum_{i=1}^{n} x_i} e^{-n\theta} \tag{53}$$

and $R(\mathbf{x}) = \frac{1}{\prod_{i=1}^{n} x_i!}$ in (4). Note that for all fixed $\theta > 0$ except $\theta = 1$, there is an obvious one-to-one function between $T_1(\mathbf{x}) = \sum_{i=1}^{n} x_i$ and (53). Hence, from **Case 1**, $I_{lost}(\mathbf{x}|K(\mathbf{x}|\theta)) = I_{lost}(\mathbf{x}|T_1) \approx 0.013$ KB from **Theorem 3** for all $\theta > 0$ except $\theta = 1$. For $\theta = 1$, $K(\mathbf{x}|\theta) = e^{-n}$ from (53) and is constant with respect to any data sample $\mathbf{x}$. Thus, $I_{reduced}(\mathbf{x}|1, K) = 0$ and $I_{lost}(\mathbf{x}|K(\mathbf{x}|1)) = I_{total}(\mathbf{x}|1, K)$. It follows that $K(\mathbf{x}|1)$ provides no information about $\mathbf{X}$.

Case 3: Let $T_3(\mathbf{X}) = L(\mathbf{X}|\theta)$ for fixed but arbitrary $\theta > 0$. We attempt to obtain $I_{lost}(\mathbf{x}|L(\mathbf{x}|\theta))$ for a data sample $\mathbf{x} = (x_1, \ldots, x_n)$ by determining $\left|A_{L(\mathbf{x}|\theta)}\right|$ and using (20). From (52), note that for all fixed $\theta > 0$ except $\theta = 1$, $\mathbf{y} \in A_{L(\mathbf{x}|\theta)}$ if and only if

$$\frac{\theta^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} = \frac{\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}. \tag{54}$$

Thus, for any fixed θ satisfying $\theta > 0$ and $\theta \neq 1$, $\mathbf{y} \in A_{L(\mathbf{x}|\theta)}$ if both $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i$ and $\prod_{i=1}^{n} y_i! = \prod_{i=1}^{n} x_i!$. However, for some $\theta > 0$ and $\theta \neq 1$, it is possible that $\mathbf{y} \in A_{L(\mathbf{x}|\theta)}$ when neither $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i$ nor $\prod_{i=1}^{n} y_i! = \prod_{i=1}^{n} x_i!$. For example, let $\theta = 2$, $\mathbf{x} = (4,1,1,0)$, and $\mathbf{y} = (3,2,0,0)$. Then, $\sum_{i=1}^{n} x_i = 6$, $\sum_{i=1}^{n} y_i = 5$, $\prod_{i=1}^{n} x_i! = 24$, and $\prod_{i=1}^{n} y_i! = 12$. However, (54) is satisfied.

This complication suggests that an efficient implicit enumeration of the $\mathbf{y}$ satisfying (54) would be required to obtain $\left|A_{L(\mathbf{x}|\theta)}\right|$ for calculating $I_{lost}(\mathbf{x}|L(\mathbf{x}|\theta))$ from (20). Using such an algorithm, a conventional computer could possibly compute $I_{lost}(\mathbf{x}|L(\mathbf{x}|\theta))$ for the numerical data and value of θ in **Case 1**, since there is now a 250 petabyte, 200 petaflop conventional computer [26]. Substantially larger problems, if not already tractable, will likely be so in the foreseeable future on quantum computers. Recently, the milestone of quantum supremacy was achieved where the various possible combinations of a certain randomly generated output were obtained in 110 s, whereas this task would have taken the above conventional supercomputer 10,000 years [27]. Regardless, for the data of **Case 1**, we have the upper bound $I_{lost}(\mathbf{x}|L(\mathbf{x}|\theta)) \leq 0.013$ KB from (32)**.**

Finally, we present some simple computational results to illustrate the relationships among T_1, T_2, T_3 with regard to the Poisson distribution. Table 1 below summarizes the results for sample data (x_1, x_2, x_3) with $\sum_{i=1}^{3} x_i \leq 2$. In particular, a complete enumeration of $A_{L(\mathbf{x}|\theta)}$ in (20) gives $I_{lost}(\mathbf{x}|L(\mathbf{x}|\theta))$.

Example 2 (Binomial Distribution). *Consider a random sample* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *from a binomial random variable* X *withparameters* m *and* θ*, where* θ *is the probability of success on any of the* m *Bernoulli trials associated with the* X_i*,* $i = 1, \ldots, n$*. Let* m *be fixed, so the only parameter is* θ*. Moreover, the sample space of the underlying random variable* X is now finite.

Case 1: $T_1(\mathbf{X}) = \sum_{i=1}^{n} X_i$. Again, $\sum_{i=1}^{n} X_i$ is an SS for θ. From [21], $\sum_{i=1}^{n} X_i$ has a binomial distribution with parameter θ for fixed nm. Hence,

$$P_\theta\left[\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} x_i\right] = \theta^{\sum_{i=1}^{n} x_i}\,\theta^{mn-\sum_{i=1}^{n} x_i}\begin{pmatrix} mn \\ \sum_{i=1}^{n} x_i \end{pmatrix} \tag{55}$$

and

$$P_\theta[\mathbf{X} = \mathbf{x}] = \theta^{\sum_{i=1}^{n} x_i}\,\theta^{mn-\sum_{i=1}^{n} x_i}\prod_{i=1}^{n}\begin{pmatrix} m \\ x_i \end{pmatrix}. \tag{56}$$

From (1), dividing (56) by (55) gives

$$P\left[\mathbf{X} = \mathbf{x}\middle|\sum_{i=1}^{n} X_i = t\right] = \frac{\prod_{i=1}^{n}\begin{pmatrix} m \\ x_i \end{pmatrix}}{\begin{pmatrix} mn \\ t \end{pmatrix}}. \tag{57}$$

By taking the $-\log$ of (57), the lost information is given as

$$I_{lost}(\mathbf{x}|T_1) = -\log\frac{\prod_{i=1}^{n}\begin{pmatrix} m \\ x_i \end{pmatrix}}{\begin{pmatrix} mn \\ t \end{pmatrix}} = -\sum_{i=1}^{n}\log\begin{pmatrix} m \\ x_i \end{pmatrix} + \log\begin{pmatrix} mn \\ t \end{pmatrix}. \tag{58}$$

Case 2: $T_2(\mathbf{X}) = K(\mathbf{X}|\theta)$. In this case, we use (16) as in **Example 1**. Write

$$L(\mathbf{x}|\theta) = f(\mathbf{x}|\theta) = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{mn-\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \binom{m}{x_i}, \tag{59}$$

from which $K(\mathbf{x}|\theta) = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{mn-\sum_{i=1}^{n} x_i}$ and $R(\mathbf{x}) = \prod_{i=1}^{n} \binom{m}{x_i}$ in (4). To factor the right side of (60) as in (2), let g be the identity function and $h(\mathbf{x}) = \prod_{i=1}^{n} \binom{m}{x_i}$. Hence,

$$I_{lost}(\mathbf{x}|T_2) = -\log \frac{\prod_{i=1}^{n} \binom{m}{x_i}}{\sum_{\mathbf{y} \in A_{K(\mathbf{x}|\theta)}} \prod_{i=1}^{n} \binom{m}{y_i}}, \tag{60}$$

and (60) gives

$$I_{lost}(\mathbf{x}|T_2) = -\sum_{i=1}^{n} \log \binom{m}{x_i} + \log \sum_{\mathbf{y} \in A_{K(\mathbf{x}|\theta)}} \prod_{i=1}^{n} \binom{m}{y_i}, \tag{61}$$

where

$$A_{K(\mathbf{x}|\theta)} = \left\{ \mathbf{y} \in S^n \,|\, \theta^{\sum_{i=1}^{n} y_i} (1-\theta)^{mn-\sum_{i=1}^{n} y_i} = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{mn-\sum_{i=1}^{n} x_i} \right\}. \tag{62}$$

From (62), for any fixed θ satisfying $0 < \theta < 1$ and $\theta \neq 1/2$, it can easily be shown that $\mathbf{y} \in A_{K(\mathbf{x}|\theta)}$ if and only if $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i$. Thus, in general, for a given $\mathbf{x}$ and fixed θ, determining $A_{K(\mathbf{x}|\theta)}$ in **Case 2** would require an enumeration of the $\mathbf{y}$ satisfying (62) to compute (61). We perform such an enumeration below for a simple example.

Case 3: $T_3(\mathbf{X}) = L(\mathbf{X}|\theta)$. For a data sample $\mathbf{x} = (x_1, \ldots, x_n)$, we now have

$$L(\mathbf{x}|\theta) = \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^{n} x_i} (1-\theta)^{mn} \prod_{i=1}^{n} \binom{m}{x_i} \tag{63}$$

with g being the identity function and $h(\mathbf{x}) = 1$ in (2). For fixed θ satisfying $0 < \theta < 1$ and $\theta \neq 1/2$, from (63) we obtain that $\mathbf{y} \in A_{L(\mathbf{x}|\theta)}$ if and only if

$$\left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^{n} y_i} \prod_{i=1}^{n} \binom{m}{y_i} = \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \binom{m}{x_i}. \tag{64}$$

As in **Case 3** of **Example 1**, developing an algorithm to use (64) and determine $\left|A_{L(\mathbf{x}|\theta)}\right|$ for calculating $I_{lost}(\mathbf{x}|L(\mathbf{x}|\theta))$ from (20) is beyond the scope of this paper.

As a simple example, consider the experiment of flipping a possibly biased coin twice ($m = 2$). The total number of heads follows a binomial distribution with the parameter θ, which is the probability of getting a head on any flip. By doing this experiment three times, we generate the random variables X_1, X_2, X_3 with possible values 0, 1, 2. Table 2 shows all the possibilities and the lost information for the statistics. The small size of this example allows the computation of I_{lost} in **Cases 2** and **3** via total enumeration.

Table 2. Binomial Example.

$x = (x_1,x_2,x_3)$	$T_1(x)$	$I_{lost}(x\|T_1)$	$T_2(x)$	$I_{lost}(x\|T_2)$	$T_3(x)$	$I_{lost}(x\|T_3)$
(0,0,0)	0	0	$(1-\theta)^6$	0	$(1-\theta)^6$	0
(0,0,1)	1	$\log 3$	$(1-\theta)^5\theta^1$	$\log 3$	$2(1-\theta)^5\theta^1$	$\log 3$
(0,1,0)						
(1,0,0)						
(1,1,0)	2	$\log \frac{15}{4}$	$(1-\theta)^4\theta^2$	$\log \frac{15}{4}$	$4(1-\theta)^4\theta^2$	$\log 3$
(1,0,1)						
(0,1,1)						
(2,0,0)	2	$\log 15$	$(1-\theta)^4\theta^2$	$\log 15$	$(1-\theta)^4\theta^2$	$\log 3$
(0,2,0)						
(0,0,2)						
(1,1,1)	3	$\log \frac{5}{2}$	$(1-\theta)^3\theta^3$	$\log \frac{5}{2}$	$8(1-\theta)^3\theta^3$	0
(2,1,0)	3	$\log 10$	$(1-\theta)^3\theta^3$	$\log 10$	$2(1-\theta)^3\theta^3$	$\log 6$
(2,0,1)						
(1,0,2)						
(1,2,0)						
(0,1,2)						
(0,2,1)						
(2,1,1)	4	$\log \frac{15}{4}$	$(1-\theta)^2\theta^4$	$\log \frac{15}{4}$	$4(1-\theta)^2\theta^4$	$\log 3$
(1,2,1)						
(1,1,2)						
(2,2,0)	4	$\log 15$	$(1-\theta)^2\theta^4$	$\log 15$	$(1-\theta)^2\theta^4$	$\log 3$
(2,0,2)						
(0,2,2)						
(2,2,1)	5	$\log 3$	$(1-\theta)^1\theta^5$	$\log 3$	$2(1-\theta)^1\theta^5$	$\log 3$
(2,1,2)						
(1,2,2)						
(2,2,2)	6	0	θ^6	0	θ^6	0

Now, using (40), we give in Table 3 the entropic losses of **Example 2** for T_1, T_2, T_3. Note that $H_{lost}(\mathbf{X}, T)$ is the same for the sum T_1 and the likelihood kernel T_2, which are related by a one-to-one function. Hence, **Result 5** is corroborated. In addition, observe that $H_{lost}(\mathbf{X}, T)$ is smallest for the likelihood function T_3.

Table 3. Entropic loss over different statistics for a binomial distribution.

$H_{lost}(X,T_1)$	$H_{lost}(X,T_2)$	$H_{lost}(X,T_3)$
1.4722	1.4722	1.2095

Example 3 (Geometric Distribution). *Consider a random sample* $\mathbf{X} = (X_1, \ldots, X_n)$ *with sample data* $x = (x_1, \ldots, x_n)$ *from a geometric random variable* X*, where the parameter* θ *is the probability of success on any of the series of independent Bernoulli trials for which* X *is the trial number on which the first success is obtained. It readily follows from [5] that*

$$P[\mathbf{X} = \mathbf{x}] = \theta^n (1-\theta)^{\sum_{i=1}^{n} x_i - n}. \tag{65}$$

Case 1: $T_1(\mathbf{X}) = \sum_{i=1}^{n} X_i$. For fixed n, $\sum_{i=1}^{n} X_i$ has a negative binomial distribution with parameter θ [21]. Hence,

$$P\left[\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} x_i\right] = \begin{pmatrix} \sum_{i=1}^{n} x_i - 1 \\ n-1 \end{pmatrix} \theta^n (1-\theta)^{\sum_{i=1}^{n} x_i - n}. \tag{66}$$

Thus $T_1(\mathbf{X}) = \sum_{i=1}^{n} X_i$ is an SS for θ since it satisfies (2) with $g[T_1(\mathbf{x})|\theta] = \theta^n (1-\theta)^{T_1(\mathbf{x})-n}$ and $h(x_1, \ldots, x_n) = \begin{pmatrix} \sum_{i=1}^{n} x_i - 1 \\ n-1 \end{pmatrix}$. Moreover, substitution of (65) and (66) into (8) gives

$$P\left[\mathbf{X} = \mathbf{x} \middle| \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} x_i\right] = \frac{1}{\begin{pmatrix} \sum_{i=1}^{n} x_i - 1 \\ n-1 \end{pmatrix}}. \tag{67}$$

Then from (14) and (67) we obtain that

$$I_{lost}(\mathbf{x}|T_1) = \log\begin{pmatrix} \sum_{i=1}^{n} x_i - 1 \\ n-1 \end{pmatrix}. \tag{68}$$

Case 2: $T_2(\mathbf{X}) = K(\mathbf{X}|\theta)$. From (66), for all $\mathbf{x} \in S^n$, $R(\mathbf{x}) = 1$ and

$$K(\mathbf{x}|\theta) = L(\mathbf{x}|\theta) = \left(\frac{\theta}{1-\theta}\right)^n (1-\theta)^{\sum_{i=1}^{n} x_i}. \tag{69}$$

Thus, for $0 < \theta < 1$, there is an obvious one-to-one function between $T_1(\mathbf{x}) = \sum_{i=1}^{n} x_i$ and $T_2(\mathbf{x}) = K(\mathbf{x}|\theta)$ in (69). Thus, from **Theorem 3**, $I_{lost}(\mathbf{x}|T_2(\mathbf{x})) = I_{lost}(\mathbf{x}|T_1)$ as given in (68).

Case 3: $T_3(\mathbf{X}) = L(\mathbf{X}|\theta)$. Since $K(\mathbf{X}|\theta) = L(\mathbf{X}|\theta)$ from (69), then

$$I_{lost}(\mathbf{x}|T_3) = \log\begin{pmatrix} \sum_{i=1}^{n} x_i - 1 \\ n-1 \end{pmatrix} \tag{70}$$

from (68). However, there is an alternate derivation of (70). For $0 < \theta < 1$ it follows from (69) that then $\mathbf{y} \in A_{L(\mathbf{x}|\theta)}$ if and only if

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i. \tag{71}$$

But for fixed positive integers $x_1, \ldots, x_n$ we have from [28] that the number of solutions $|A_{L(\mathbf{x}|\theta)}|$ to (71) in positive integers $y_1, \ldots, y_n$ is

$$\begin{pmatrix} \sum_{i=1}^{n} x_i - 1 \\ n-1 \end{pmatrix}. \tag{72}$$

Thus, (70) follows for $L(\mathbf{X}|\theta)$ from (72) and (20), so $I_{lost}(\mathbf{x}|T_1) = I_{lost}(\mathbf{x}|T_2) = I_{lost}(\mathbf{x}|T_3)$ from **Theorem 3**.

As a numerical illustration, let the random variable X denote the number of flips of a possibly biased coin until a head is obtained. Then, X has a geometric distribution, where the parameter θ is now the probability of getting a head on any flip. Suppose this experiment is performed three times yielding the sample data $\mathbf{x} = (x_1, x_2, x_3)$ shown in Table 4. $I_{lost}(\mathbf{x}|T)$ is then calculated for each of

the sufficient statistics for θ of **Example 3**. Observe that the individual statistics depend on θ while the lost information does not. Moreover, $I_{lost}(\mathbf{x}|T_1) = I_{lost}(\mathbf{x}|T_2) = I_{lost}(\mathbf{x}|T_3)$ for all data samples, as established above.

Table 4. Geometric Example.

$\mathbf{x} = (x_1,x_2,x_3)$	$T_1(\mathbf{x})$	$I_{lost}(\mathbf{x}\|T_1)$	$T_2(\mathbf{x})$	$I_{lost}(\mathbf{x}\|T_2)$	$T_3(\mathbf{x})$	$I_{lost}(\mathbf{x}\|T_3)$
(1,1,1)	3	0	θ^3	0	θ^3	0
(2,1,1)	4	$\log 3$	$\theta^3(1-\theta)$	$\log 3$	$\theta^3(1-\theta)$	$\log 3$
(1,2,1)						
(1,1,2)						
(2,2,1)	5	$\log 6$	$\theta^3(1-\theta)^2$	$\log 6$	$\theta^3(1-\theta)^2$	$\log 6$
(2,1,2)						
(1,2,2)						
(2,2,2)	6	$\log 10$	$\theta^3(1-\theta)^3$	$\log 10$	$\theta^3(1-\theta)^3$	$\log 10$

6. Conclusions

In this paper, the Shannon information obtained for a random sample **X** taken from a discrete random variable X with a single parameter θ was decomposed into two components: (i) the reduced information associated with the value of a real-valued statistic $T(\mathbf{X})$ evaluated at the data sample **x**, and (ii) the information lost by using this value as a surrogate for **x**. Information is lost because multiple data sets can give the same value of the statistic. In data analysis, the data uniquely determines the value of a statistic, but typically the value of the statistic does not uniquely determine the data yielding it. The lost information thus measures the knowledge unavailable about the data sample **x** when only the reduced data summary $T(\mathbf{x})$ is known, but not **x** itself. To eliminate the effect of θ, we focused on sufficient statistics for θ such as the sample mean. We then answered the question: how much Shannon information is lost to someone about a data sample **x** when only the value of $T(\mathbf{x})$ is available but not the data **x** itself? Our answer is independent of the parameter θ and does not require that θ be known. Our method generalizes the approach of [2] for analyzing the information contained in a sequence of Bernoulli trials.

More generally, we developed a metric associated with the value $T(\mathbf{x})$ used to summarize, represent, or characterize a given data set. Our approach and results are significant because such statistics are often communicated without the original data. One could argue that $I_{lost}(\mathbf{x}|T)$ should be communicated along with $T(\mathbf{x})$ in a manner similar to providing the margin of error associated with the results of a poll. A small $I_{lost}(\mathbf{x}|T)$ would signify that $T(\mathbf{x})$ is more informative than if $I_{lost}(\mathbf{x}|T)$ were large.

In addition, we defined the entropic loss associated with a sufficient statistic T under consideration as the expected lost information over all possible samples, to give a value dependent only on T. We noted but did not explore the possibility that entropic loss could be used as a metric to compare different sufficient statistics. Moreover, if sufficient statistics were not required, entropic loss could provide metrics on either θ or T if the other of these variables is fixed. Finally, numerical examples of our results were presented and some computational issues noted.

Author Contributions: M.M. suggested the topic after reading [2]. She shared the development of the theory and examples of this paper, as well as wrote early drafts. H.C. formulated the general decomposition here. He shared the development of the theory and examples of this paper, as well as edited the final draft. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Landauer, R. Irreversibility and heat generation in the computing process. *IBM J. Res. Dev.* **1961**, *5*, 183–191. [CrossRef]
2. Hodge, S.E.; Vieland, V.J. Information loss in binomial data due to data compression. *Entropy* **2017**, *19*, 75. [CrossRef]
3. Casella, G.; Berger, R.L. *Statistical Inference*, 2nd ed.; Cengage Learning: Delhi, India, 2002.
4. Pawitan, Y. *All Likelihood: Statistical Modeling and Inference Using Likelihood*, 1st ed.; The Clarendon Press: Oxford, UK, 2013.
5. Rohatgi, V.K.; Saleh, A.K.E. *An Introduction to Probability and Statistics*, 2nd ed.; John Wiley & Sons, Inc.: New York, NY, USA, 2001.
6. Shannon, C.E.; Weaver, W. *The Mathematical Theory of Communication*, 1st ed.; The University of Illinois Press: Urbana, IL, USA, 1964.
7. Shannon, C. A mathematical theory of communication. *Bell. Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
8. Vigo, R. Representational information: A new general notion and measure of information. *Inf. Sci.* **2011**, *181*, 4847–4859. [CrossRef]
9. Vigo, R. Complexity over uncertainty in generalized representational information theory (GRIT): A structure-sensitive general theory of information. *Information* **2013**, *4*, 1–30. [CrossRef]
10. Klir, G.J. *Uncertainty and Information: Foundations of Generalized Information Theory*, 1st ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2006.
11. Devlin, K. *Logic and Information*, 1st ed.; Cambridge University Press: Cambridge, UK, 1991.
12. Luce, R.D. Whatever happened to information theory in psychology? *Rev. Gen. Psychol.* **2003**, *7*, 183–188. [CrossRef]
13. Floridi, L. *The Philosophy of Information*, 1st ed.; Oxford University Press: Oxford, UK, 2011.
14. Garner, W.R. *The Processing of Information and Structure*, 1st ed.; Wiley: New York, NY, USA, 1974.
15. Spellerberg, I.F.; Fedor, P.J. A tribute to Claude-Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the "Shannon-Wiener" index. *Glob. Ecol. Biogeogr.* **2003**, *12*, 177–179. [CrossRef]
16. Shamir, O.; Sabato, S.; Tishby, N. Learning and generalization with the information bottleneck. *Theor. Comput. Sci.* **2010**, *411*, 2696–2711. [CrossRef]
17. Csiszár, I. Axiomatic characterizations of information measures. *Entropy* **2008**, *10*, 261–273. [CrossRef]
18. Kapur, J.N.; Kesavan, H.K. *Entropy Optimization Principles and Their Applications*, 1st ed.; Water Science and Technology Library, Springer: Dordrecht, The Netherlands, 1992; Volume 9.
19. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2006.
20. Ibekwe-SanJuan, F.; Dousa, T. *Theories of Information, Communication and Knowledge: A Multidisciplinary Approach*, 1st ed.; Springer: Dordrecht, The Netherlands, 2014.
21. Johnson, J.L. *Probability and Statistics for Computer Science*, 1st ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2008.
22. Beeler, R.A. *How to Count: An Introduction to Combinatorics and Its Applications*, 1st ed.; Springer: Cham, Switzerland, 2015.
23. Sridharan, S.; Balakrishnan, R. *Foundations of Discrete Mathematics with Algorithms and Programming*, 1st ed.; Chapman and Hall/CRC: New York, NY, USA, 2018.
24. Mortici, C. Ramanujan formula for the generalized Stirling approximation. *Appl. Math. Comput.* **2010**, *19*, 2579–2585. [CrossRef]
25. Multinomial Coefficient Calculator. Available online: https://mathcracker.com/multinomial-coefficient-calculator.php (accessed on 27 July 2019).
26. Wan, L.; Mehta, K.V.; Klasky, S.A.; Wolf, M.; Wang, H.Y.; Wang, W.H.; Li, J.C.; Lin, Z. Data management challenges of exascale scientific simulations: A case study with the Gyrokinetic Toroidal Code and ADIOS. In Proceedings of the 10th International Conference on Computational Methods, ICCM'19, Singapore, 9–13 July 2019.

27. Arute, F.; Arya, K.; Martinis, J.M. Quantum supremacy using a programmable superconducting processor. *Nature* **2019**, *574*, 505–510. [CrossRef] [PubMed]
28. Mahmoudvand, R.; Hassani, H.; Farzaneh, A.; Howell, G. The exact number of nonnegative integer solutions for a linear Diophantine inequality. *IAENG Int. J. Appl. Math.* **2010**, *40*, 5.