entropy

MDPI

Article

Information Loss due to the Compression of Sample Data from Discrete Distributions

Maryam Moghimi*[1,2], H.W. Corley[1,2]

[1] Center on Stochastic Modeling, Optimization, and Statistics (COSMOS), The University of Texas at Arlington, Arlington, TX, USA

[2] The authors contributed equally to this paper.

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

Received: date; Accepted: date; Published: date

Abstract: In this paper we study the information lost when a real-valued statistic $T(X_1, \ldots, X_n)$ is used to summarize the sample data $\mathbf{x} = (x_1, \ldots, x_n)$ of a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from a discrete random variable X with a one-dimensional parameter θ. We compare the probability that the random sample $\mathbf{X}$ yields $\mathbf{x}$ to the probability that the compressed sample $T(\mathbf{X})$ yields $T(\mathbf{x})$. The former probability measures the total information about $\mathbf{x}$, while the latter measures the compressed information about $\mathbf{x}$, both of which are expressed here as Shannon information. The difference is the information lost about $\mathbf{X}$ by its compression to $T(\mathbf{X})$. We focus on sufficient statistics for the parameter θ and develop a general formula independent of θ for this lost information as well as for an associated entropy that depends only on T. Our approach would also work for non-sufficient statistics, but the lost information and associated entropy would involve θ. Examples are presented for some standard discrete distributions.

Keywords: discrete distributions, Shannon information, lost information, sampling, data reduction, data compression, entropy, sufficient statistics, likelihood

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. Here a statistic $T(\mathbf{X})$ is a real-valued function of the random sample but not a function of any parameter θ associated with X, though θ may fixed at an arbitrary value. The data sample $\mathbf{X}$ is compressed to the summary statistic $T(\mathbf{X})$, which could be used to characterize $\mathbf{X}$ or to estimate θ. Such data compression is an irreversible process [1] and always involves some information loss. For instance, if $T(\mathbf{X}) = \bar{X}$, the original measurements $\mathbf{x}$ cannot be reconstructed from $\bar{x}$, and some information about $\mathbf{x}$ is lost. Nonetheless, such data compression is frequently used to make inferences about, for example, the true mean μ of X. Our information-theoretic approach to data compression 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.

For any real-valued statistic T and the given sample data $\mathbf{x}$, we decompose the total information about $\mathbf{X}$ available in $\mathbf{x}$ into the sum of (a) the information available in the compressed data $T(\mathbf{x}) = \bar{x}$ and (b) the information lost in the compression. 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 $\mathbf{x}$ nor θ. Our approach also works for non-sufficient statistics, but the lost information and associated entropy would then involve θ, and so θ must be estimated to 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 $\mathbf{X}$ in $\mathbf{x}$

and give various expressions for the Shannon information lost by compressing $\mathbf{x}$ 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 statistics sufficient for θ. Conclusions are offered in Section 6.

2. Preliminaries

The following definitions, notation, and results are used here. Further details can be found in [3,4] and elsewhere. An important class of statistics is first defined.

Definition 2.1 (Sufficient Statistic). A statistic $T(\mathbf{X})$ is a sufficient statistic (SS) for the parameter θ if the probability

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

is independent of θ.

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

Result 2.2 (Fisher Factorization Theorem). The real-valued statistic $T(\mathbf{X})$ is sufficient for θ if and only if there exist functions $g: R^1 \longrightarrow R^1$ and $h: S^n \longrightarrow R^1$ such that for any sample data $\mathbf{x}$ and for all values of θ the joint pmf $f(\mathbf{x}|\theta)$ of $\mathbf{X}$ can be factored as

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

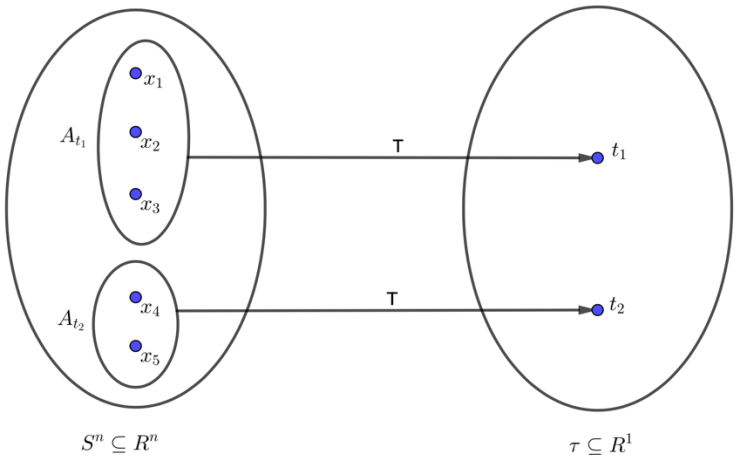
for real-valued, nonnegative functions g on R^1 and h on S^n. The function h does not depend on θ, while g does depend on $\mathbf{x}$ but only through $T(\mathbf{x})$.

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

Definition 2.3 (Partition). Let S be the denumerable sample space of the discrete random variable X, and thus let S^n be the denumerable sample space of the random sample $\mathbf{X}$. For any statistic $T: S^n \longrightarrow R^1$, let τ_T be the denumerable set $\tau_T = \{t|\exists \mathbf{x} \in S^n$ for which $t = T(\mathbf{x})\}$, which is the range of T. Then T partitions the sample space S^n into the mutually exclusive and collectively exhaustive partition sets $A_t = \{\mathbf{x} \in S^n|T(\mathbf{x}) = t\}, \forall t \in \tau_T$.

Figure 2.1 below illustrates the situation.

Figure 2.1



We also need the well-known likelihood function.

Definition 2.4 (Likelihood Function). Let $\mathbf{x}$ be sample data from a random sample $\mathbf{X}$ from a discrete random variable X with sample space S and real-valued parameter θ, and let $f(\mathbf{x}|\theta)$ denote the joint pmf of the random sample $\mathbf{X}$. For any sample data $\mathbf{x}$, the likelihood function of θ is defined as

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

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

Definition 2.5 (Likelihood kernel). Let S be the sample space of $\mathbf{X}$. For fixed θ, suppose that $L(\mathbf{x}|\theta)$ can be factored as

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

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

(a) every nonnumerical factor of $K(\mathbf{x}|\theta)$ contains θ;

(b) $R(\mathbf{x})$ does not contain θ;

(c) for $\forall \mathbf{x} \in S^n$, both $K(\mathbf{x}|\theta) \geq 0$ and $R(\mathbf{x}) \geq 0$; and

(d) $K(\mathbf{x}|\theta)$ is not divisible by any positive number except 1.

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

Theorem 2.6. The likelihood kernel $K(\mathbf{x}|\theta)$ has the following properties.

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

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

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

Proof. To prove (i), for fixed θ we first show that the likelihood kernel $K(\mathbf{x}|\theta)$ of **Definition 2.5** 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 2.5,** 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 2.5** 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 2.5.** 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 avoid (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 2.2**.

Finally, (iii) follows immediately from **Definition 2.5** and the fact that $L(\mathbf{x}|\theta_2) \neq 0$ for $\mathbf{x} \in S^n$. ■

We next discuss the notion of information to be used here. 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 2.7 (Shannon Information). Let $\mathbf{x}$ be sample data for the random sample $\mathbf{X}$ from the discrete random variable X with a one-dimensional parameter θ, and let $f(\mathbf{x}|\theta)$ be the joint pmf of $\mathbf{X}$ at $\mathbf{x}$. The Shannon information obtained from the sample data $\mathbf{x}$ is defined as

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

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

The expected information over $\forall \mathbf{x} \in S^n$ will also be used.

Definition 2.8 (Entropy). Under the conditions of **Definition 2.7**, the entropy $H(\mathbf{X}|\theta)$ is defined as the expected value of $I(\mathbf{X}|\theta)$; i.e,

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

Since entropy is the expected information over all possible random samples, it measures the available information about $\mathbf{X}$ better than would a single data set $\mathbf{x}$, which might not be typical [8]. We next give a method to obtain the information loss about $\mathbf{X}$ that occurs when a data set $\mathbf{x}$ is compressed to $T(\mathbf{x})$. In our approach, we focus on a sufficient statistic T so there will be no θ in (5) for the lost information below. However, our approach is applicable to a non-sufficient statistic as well if θ is estimated from the data.

3. Information Decomposition under Data Compression 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 compressed 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 compression 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 3.1** is given in [3, p. 273] and proved below to illustrate the reasoning.

Result 3.1. 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

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

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

Observe that if T is an SS for θ, the right side of (8) is independent of θ and hence so is the left. Now 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 compressed 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 compression 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 2.1 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 compressed data summary $T(\mathbf{x})$ is known and 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 2.1 above, the situation may be described as follows. On the left 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 compresses the data sample $\mathbf{x}$ into $T(\mathbf{x})$, where multiple $\mathbf{x}$'s may give the same $T(\mathbf{x}) = t$. In Figure 2.1 the distinct data samples $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ are all compressed into the same value t_1. But 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 compression. One can also say that the total information $-\log P_\theta[\mathbf{X} = \mathbf{x}]$ deriving from the left side of Figure 2.1 is compressed to $-\log P_\theta[T(\mathbf{X}) = T(\mathbf{x})]$ deriving from the right. The reduction of information from the left to the right side is precisely the lost information $-\log P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]$. For fixed t, it is lost due to the ambiguity as to which data sample on the left actually gave t when only t is known. There is no ambiguity when T is one-to-one.

The general decomposition of information in (10) is next summarized in **Definition 3.2**, where T does not need to be sufficient for θ.

Definition 3.2 ($I_{total}, I_{comp}, I_{lost}$). 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. For any real-valued statistic. $T(\mathbf{X})$, the Shannon information about $\mathbf{X}$ obtained from the sample data $\mathbf{x}$ can be decomposed as

$$I_{total}(\mathbf{x}|\theta) = I_{comp}(\mathbf{x}|\theta, T) + I_{lost}(\mathbf{x}|\theta, T), \tag{11}$$

where

$$I_{total}(\mathbf{x}|\theta) = -\log P_\theta[\mathbf{X} = \mathbf{x}], \tag{12}$$

$$I_{comp}(\mathbf{x}|\theta, T) = -\log P_\theta[T(\mathbf{X}) = T(\mathbf{x})], \tag{13}$$

and

$$I_{lost}(\mathbf{x}|\theta, T) = -\log P_\theta[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]. \tag{14}$$

Both **Result 3.1** and **Definition 3.2** are valid for any real-valued statistic for $\mathbf{X}$. The notation $I_{total}(\mathbf{x}|\theta)$ indicates that I_{total} is a function of the sample data $\mathbf{x}$ for a fixed but arbitrary parameter value θ. Similarly, both $I_{comp}(\mathbf{x}|\theta, T)$ and $I_{lost}(\mathbf{x}|\theta, T)$ are functions of $\mathbf{x}$ for fixed θ and T. However, in this paper we focus on sufficient statistics, which provide a simpler expression for $I_{lost}(\mathbf{x}|\theta, T)$ that does not involve θ. For a sufficient statistic T for θ, we use the notation $I_{lost}(\mathbf{x}|T)$ for the lost information, though $I_{total}(\mathbf{x}|\theta)$ and $I_{comp}(\mathbf{x}|\theta, T)$ still require θ. The next result is an application of the FFT of **Result 2.2**.

Theorem 3.3 (Lost Information for an SS). 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. Let T be an SS for θ, $f(\mathbf{x}|\theta)$ be the joint pmf of $\mathbf{X}$, and $f(\mathbf{x}|\theta) = g[T(\mathbf{x})|\theta] \times h(\mathbf{x})$ as in **Result 2.2**. Then for all $\mathbf{x} \in S^n$

$$I_{lost}(\mathbf{x}|T) = -\log \frac{h(\mathbf{x})}{\sum_{\mathbf{y} \in A_{T(\mathbf{x})}} h(\mathbf{y})}, \tag{15}$$

where $A_{T(\mathbf{x})}$ is defined in **Definition 2.3** for $t = T(\mathbf{x})$.

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

$$P[\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})] = \frac{h(\mathbf{x})}{\sum_{\mathbf{y} \in A_{T(\mathbf{x})}} h(\mathbf{y})}, \tag{16}$$

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

$$P[\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x})]=\frac{P_\theta[\mathbf{X}=\mathbf{x}]}{P_\theta[T(\mathbf{X})=T(\mathbf{x})]}=\frac{f(\mathbf{x}|\theta)}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} f(\mathbf{y}|\theta)}, \tag{17}$$

195 so from (17) and (2), then

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

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

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

197 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
198 (16). Taking $-\log$ of (16) completes the proof. ■

199 Now consider **Theorem 3.3** when each A_t is a singleton in (16), i.e., when T is a one-to-one
200 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
201 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
202 all $\mathbf{x}$ in S^n. Thus the special case of a one-to-one T justifies the identification of the lost information
203 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 $\mathbf{x}\neq\mathbf{y}$
204 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 compression of the
205 singleton $A_{T(\mathbf{x})}$ to the number $T(\mathbf{x})$.

206 More generally, it is also true that $I_{lost}(\mathbf{x}|\theta,T)=0$ when T is one-to-one but not sufficient for
207 θ. In this case, write $P_\theta[\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x})]=\frac{P_\theta[\mathbf{X}=\mathbf{x}]}{P_\theta[T(\mathbf{X})=T(\mathbf{x})]}=\frac{f(\mathbf{x}|\theta)}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} f(\mathbf{y}|\theta)}$. But since T is one-to-one,
208 $\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} f(\mathbf{y}|\theta)=f(\mathbf{x}|\theta)$, $P_\theta[\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x})]=1$, and again $I_{lost}(\mathbf{x}|\theta,T)=0$.

209 Now consider the other extreme case where $T(\mathbf{x})=c$ is constant on S^n. Then
210 $P_\theta[\mathbf{X}=\mathbf{x}|T(\mathbf{X})=c]=\frac{P_\theta[\mathbf{X}=\mathbf{x}]}{P_\theta[T(\mathbf{X})=c]}$. But $P_\theta[T(\mathbf{X})=c]=1$, so $P_\theta[\mathbf{X}=\mathbf{x}|T(\mathbf{X})=c]=P_\theta[\mathbf{X}=\mathbf{x}]$ and
211 $I_{lost}(\mathbf{x}|\theta,T)=I_{total}(\mathbf{x}|\theta,T)$ on S^n. In this case, $I_{comp}(\mathbf{x}|\theta,T)=0$ because the event $T(\mathbf{x})=c$ gives no
212 information about $\mathbf{x}$.

213 Next, in the following corollary we show that (16) can be simplified when T is the likelihood
214 function.

215 **Corollary 3.4 (Information Loss for Likelihood Function).** Under the assumptions of **Theorem**
216 **3.3**, if $T(\mathbf{x})=L(\mathbf{x}|\theta)$, then

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

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

218 **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
219 substituting $h(\mathbf{x})=1$ into (16) gives the denominator $\sum_{\mathbf{y}\in A_{L(\mathbf{x}|\theta)}} 1=\left|A_{L(\mathbf{x}|\theta)}\right|$ to yield (20). ■

220 We next state a reproductive property of a statistic T' that is a one-to-one function of a sufficient
221 statistic T for θ.

222 **Theorem 3.5.** If there is a one-to-one function between a sufficient statistic T for θ and an
223 arbitrary real-valued statistic T' on S^n, the following hold.

224 (i) T' is also an SS.

225 (ii) T and T' partition the sample space S into the same partition sets.

226 (iii) $I_{lost}(\mathbf{x}|T)=I_{lost}(\mathbf{x}|T'), \forall \mathbf{x}\in S^n$.

227 **Proof.** To prove (i), let u be a real-valued one-to-one function of T' such that

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

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

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

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

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

which can be rewritten as

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

Since T' in (24) satisfies the condition of **Result 2.2 for** $g' = g \circ u$, T' is an SS.

To prove (ii), we use **Definition 2.3**. Let T partition the sample space S^n into the mutually exclusive and collectively exhaustive sets $A_t = \{\mathbf{x}|T(\mathbf{x}) = t\}, \forall t \in \tau_T$. By equation (21) we can also write A_t as

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

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

$$A_t = \{\mathbf{x}|T'(\mathbf{x}) = t'\}, \forall t' \in u(\tau_T). \tag{26}$$

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

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

Finally, to get (iii) we use **Theorem 3.3** to calculate information lost over two statistics T and T'. Since $h(\mathbf{x})$ is the same in (22) and (24) and since equation (27) holds, we sum $h(\mathbf{x})$ over the same sets in the denominator of equation (16) for both T and T' to give

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

and complete the proof. ■

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

Lemma 3.6. Let $\mathbf{x}$ be any data sample for a random sample $\mathbf{X}$ from the discrete random variable X with real-valued parameter θ. Then $K(\mathbf{x}|\theta)$ is a function of $L(\mathbf{x}|\theta)$ and $\tau_L \geq \tau_K$.

Proof. From [3, p. 280], $K(\mathbf{x}|\theta)$ is a function of $L(\mathbf{x}|\theta)$ if and only if $K(\mathbf{x}|\theta) = K(\mathbf{y}|\theta)$ whenever $L(\mathbf{x}|\theta) = L(\mathbf{y}|\theta)$. For all data samples $\mathbf{x}$ and $\mathbf{y}$, we thus prove that if $L(\mathbf{x}|\theta) = L(\mathbf{y}|\theta)$, then $K(\mathbf{x}|\theta) = K(\mathbf{y}|\theta)$. Thus suppose that $L(\mathbf{x}|\theta) = L(\mathbf{y}|\theta)$. By **Definition 2.5** we can decompose $L(\mathbf{x}|\theta)$ and $L(\mathbf{y}|\theta)$ into $K(\mathbf{x}|\theta)R(\mathbf{x})$ and $K(\mathbf{y}|\theta)R(\mathbf{y})$, respectively. Note that $K(\mathbf{y}|\theta) \neq 0$. Otherwise $L(\mathbf{y}|\theta) = 0$ in contradiction to $\mathbf{y}$ being sample data with a nonzero probability of occurring. Write

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

Suppose that $K(\mathbf{x}|\theta) \neq K(\mathbf{y}|\theta)$ so that $\frac{K(\mathbf{x}|\theta)}{K(\mathbf{y}|\theta)} = \frac{R(\mathbf{y})}{R(\mathbf{x})} \neq 1$ in (29). From **Definition 2.5**, every nonnumerical factor of $K(\mathbf{x}|\theta)$ and $K(\mathbf{y}|\theta)$ contains θ, and neither $K(\mathbf{x}|\theta)$ nor $K(\mathbf{y}|\theta)$ 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 $K(\mathbf{y}|\theta)$ 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 2.5**. It now follows that $K(\mathbf{x}|\theta) = K(\mathbf{y}|\theta)$, 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 3.7. Under the conditions of **Lemma 3.6**, the sufficient statistics L and K satisfy

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

Proof. Let $\mathbf{x} \in S^n$ and suppose that $\mathbf{y} \in A_{L(\mathbf{x})}$. Then $L(\mathbf{y}|\theta) = L(\mathbf{x}|\theta)$, so it follows from **Lemma 3.6** that $K(\mathbf{y}|\theta) = K(\mathbf{x}|\theta)$ and thus $\mathbf{y} \in A_{K(\mathbf{x})}$. Hence $A_{L(\mathbf{x})} \subseteq A_{K(\mathbf{x})}$, and so

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

Taking the Shannon information of both sides of the inequality in (31) and using (13) gives (30). ■

Theorem 3.8. Let $\mathbf{x}$ be sample data for a random sample $\mathbf{X}$ from a discrete random variable X with the real-valued parameter θ. Then for all $\mathbf{x} \in S^n$,

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

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

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

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

As a consequence of **Theorem 3.5**, **Theorem 3.8** has an immediate corollary.

Corollary 3.9. Under the conditions of **Theorem 3.8**, let T be a sufficient statistic for θ for which there is a one-to-one function between T and K. Then for all $\mathbf{x} \in S^n$,

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

Corollary 3.9 raises the question whether (34) holds for all sufficient statistics $\mathbf{T}$ for $\boldsymbol{\theta}$ or even for all real-valued statistics $\mathbf{T}$. It is conjectured that the first conclusion is false and hence so is the second, but the question remains open. It is conceivable that notion of a minimal sufficient statistic [3] is relevant. Regardless, the proofs of **Lemma 3.7** and **Theorem 3.8** illustrate the fact that the relation between the lost information for two statistics $\mathbf{T}$ and $\mathbf{T}'$ is determined by the relation between their partition sets $\mathbf{A_t} = \{\mathbf{x}|\mathbf{T}(\mathbf{x}) = \mathbf{t}\}$ and $\mathbf{B_{t'}} = \{\mathbf{x}|\mathbf{T}'(\mathbf{x}) = \mathbf{t}'\}$. For example, if for every $\mathbf{A_t}$ there exists a $\mathbf{B_{t'}}$ for which $\mathbf{A_t} \subset \mathbf{B_{t'}}$, then the partition of $\mathbf{S^n}$ by the $\mathbf{B_{t'}}$ of $\mathbf{T}'$ is said to be coarser than the partition by the $\mathbf{A_t}$ of $\mathbf{T}$. In that case, $\mathbf{I_{lost}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{T}) \leq \mathbf{I_{lost}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{T}')$ because each $\mathbf{x} \in \mathbf{S^n}$ has more $\mathbf{y} \in \mathbf{S^n}$ with $\mathbf{T}'(\mathbf{y}) = \mathbf{T}'(\mathbf{x})$ than there are with $\mathbf{T}(\mathbf{y}) = \mathbf{T}(\mathbf{x})$. In words, $\mathbf{T}'(\mathbf{y}) = \mathbf{t}'$ is at least as ambiguous as $\mathbf{T}(\mathbf{y}) = \mathbf{t}$ 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 compressing the random sample $\mathbf{X}$ into $T(\mathbf{X})$. 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 4.1. Under the assumptions of **Theorem 3.3**, for any data sample let $t = T(\mathbf{x})$ and consider the partition set A_t. Then

$$\sum_{\mathbf{x} \in A_t} P[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t] = 1. \quad (35)$$

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

$$\sum_{\mathbf{x} \in A_t} P[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t] = \frac{\sum_{\mathbf{x} \in A_{T(\mathbf{x})}} h(\mathbf{x})}{\sum_{\mathbf{y} \in A_{T(\mathbf{x})}} h(\mathbf{y})} = 1. \quad (36)$$

to give (35). ■

Result 4.2. Under the assumptions of **Theorem 3.3**, the sum

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

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

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

The inner series on the right side of (38) sums to one by **Result 4.1**. 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, $P[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]$ 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 4.2** to normalize $P[\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})]$ and obtain the appropriate pmf for calculating the expectation of $I_{lost}(\mathbf{X}|T)$.

Definition 4.3 (Entropic Loss). Under the assumptions of **Theorem 3.3**, the entropic loss resulting from the data compression 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

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

Note that (39) and (40) are independent of both $\mathbf{x}$ and θ. Also, as noted in Section 3 for $I_{lost}(\mathbf{x}|T)$, if each $A_{T(\mathbf{x})}$ is a singleton in (40), then $H_{lost}(\mathbf{X}, T) = 0$. We now compute $H_{lost}(T)$ for the sufficient statistic $T(\mathbf{X}) = L(\mathbf{X}|\theta)$.

Theorem 4.4 (Entropic Loss for Likelihood Function). Under the assumptions of **Theorem 3.3**, the entropic loss resulting from data compression by $T(\mathbf{x}) = L(\mathbf{x}|\theta)$ 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

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

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). ■

Since $H_{lost}(\mathbf{X}, T)$ has been defined only for a sufficient statistic T for θ and is independent of θ, as well as the data sample $\mathbf{x}$. $H_{lost}(\mathbf{X}, T)$ could thus be used to compare sufficient statistics. In particular, if the sufficient statistics $T_1(\mathbf{X})$ and $T_2(\mathbf{X})$ are considered as estimators for θ, then entropic loss could serve as a metric for regarding, say, T_1 as a better estimator for θ than T_2 if $H_{lost}(\mathbf{X}, T_1) < H_{lost}(\mathbf{X}, T_2)$.

Result 4.5. If there is a one-to-one function between two sufficient statistics T and T' for θ, then they have the same entropic loss for a random sample $\mathbf{X}$; i.e.,

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

Proof. For all $\mathbf{x} \in S^n$, $I_{lost}(\mathbf{x}|T) = I_{lost}(\mathbf{x}|T')$ from **Theorem 3.5**, so

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

from which

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

Thus from (45) and (46)

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

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

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

But from **Theorem 3.5**, $|\tau_T| = |\tau_{T'}|$. Thus dividing the left side of (48) by $-|\tau_T|$ and the right side by $-|\tau_{T'}|$ yields (44). ■

Given (32), it might be anticipated that

$$H_{lost}(\mathbf{X}, L) \le H_{lost}(\mathbf{X}, K). \tag{49}$$

However, we conjecture that (49) is not always true, but we have no counterexample. If this conjecture is true, then $L(\mathbf{x}|\theta)$ would not in general have the minimum entropic loss among sufficient statistics for θ.

5. Examples and Computational Issues

In this section we present examples involving the discrete Poisson, binomial, and geometric distributions [9]. For each distribution, three sufficient statistics for some parameter θ are analyzed. Thus the right side of (8) is independent of θ, as well as the information $I_{lost}(\mathbf{x}|T)$ conveyed by the data sample $\mathbf{x}$ about $\mathbf{X}$. Even for sufficient statistics, calculating the information quantities of this paper may present computational issues, some of which are discussed in this section. Our examples are therefore simple in order to focus on the definitions and results of Sections 3 and 4.

Example 5.1 (Poisson Distribution). Consider the random sample $\mathbf{X} = (X_1, \dots, X_n)$ with the data sample $\mathbf{x} = (x_1, \dots, x_n)$ from a Poisson random variable X. We consider three sufficient statistics for the parameter $\theta > 0$. These sufficient statistics are $T_1(\mathbf{X}) = \sum_{i=1}^n X_i$, the likelihood kernel $T_2(\mathbf{X}) = K(\mathbf{X}|\theta)$ for fixed but arbitrary θ and the likelihood function $T_3(\mathbf{X}) = L(\mathbf{X}|\theta)$ for fixed but arbitrary θ. In particular, we use $T_1(\mathbf{X})$ as a surrogate for $T_1'(\mathbf{X}) = \frac{\sum_{i=1}^n X_i}{n}$. Neither $T_1(\mathbf{X})$ or $T_1'(\mathbf{X})$ involves θ and can thus be used either to characterize $\mathbf{X}$ or to estimate θ. Moreover, since there is an obvious one-to-one function relating $\frac{\sum_{i=1}^n X_i}{n}$ and $\sum_{i=1}^n X_i$, **Theorems 3.5** and **4.5** establish that $I_{lost}(\mathbf{x}|T_1') = I_{lost}(\mathbf{x}|T_1)$ and $H_{lost}(\mathbf{X}, T_1') = H_{lost}(\mathbf{X}, T_1)$, respectively. We consider $T_1(\mathbf{X})$ because it is also Poisson, whereas $T_1'(\mathbf{X})$ is not since $\frac{\sum_{i=1}^n X_i}{n}$ is not necessarily a nonnegative integer. In contrast to $T_1(\mathbf{X})$, both $T_2(\mathbf{X})$ and $T_3(\mathbf{X})$ contain θ and can only be used to characterize $\mathbf{X}$. For each of these three sufficient statistics we develop an expression for $I_{lost}(\mathbf{x}|T)$ and describe how to obtain a numerical value. We then illustrate previous results with simple data and present computational results in Table 5.1.

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 2.2** 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 $n\theta$ [9]. Thus $P_\theta[\sum_{i=1}^n X_i = \sum_{i=1}^n x_i] = \frac{(n\theta)^{\sum_{i=1}^n x_i} e^{-n\theta}}{(\sum_{i=1}^n x_i)!}$, and so (8) becomes

$$P[\mathbf{X} = \mathbf{x} | \textstyle\sum_{i=1}^n X_i = \sum_{i=1}^n x_i] = \frac{1}{n^{\sum_{i=1}^n x_i}} \binom{\sum_{i=1}^n x_i}{x_1, \dots, x_n}, \tag{50}$$

where the multinomial coefficient $\binom{\sum_{i=1}^{n} x_i}{x_1,\ldots,x_n} = \frac{(\sum_{i=1}^{n} x_i)!}{\prod_{i=1}^{n} x_i!}$. It follows from (50) and (10) that

$$I_{lost}(\mathbf{x}|T_1) = -\log\binom{\sum_{i=1}^{n} x_i}{x_1, \ldots, x_n} + (\log n)\sum_{i=1}^{n} x_i, \tag{51}$$

which is also $I_{lost}(\mathbf{x}|T_1')$.

For a data sample $(x_1, \ldots, x_n)$, the evaluation of $I_{lost}(\mathbf{x}|T_1)$ in (51) involves computing factorials. For realistic data, the principal limitation in calculating them by direct multiplication is their magnitude. See [11] for a discussion. However, (51) can be approximated using either the well-known Stirling formula or the more accurate Ramanujan approximation [12]. The online multinomial coefficient calculator [13] can evaluate multinomial coefficients for both x_i and n less than approximately 50 if any $x_i = 0$ is removed from $\binom{\sum_{i=1}^{n} x_i}{x_1,\ldots,x_n}$. 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 $\theta = 3$. On the average, $T_1(\mathbf{X}) = \sum_{i=1}^{n} X_i = n\theta = 102$, so we take $\sum_{i=1}^{n} x_i = 102$ for the data sample $\mathbf{x} = (4, 7, 1, 3, 4, 2, 5, 0, 1, 2, 3, 6, 8, 0, 1, 2, 4, 9, 0, 2, 3, 1, 4, 2, 0, 1, 5, 6, 2, 7, 0, 1, 4, 2)$. Then the calculator at [13] gives that $\binom{\sum_{i=1}^{n} x_i}{x_1,\ldots,x_n} \approx 1.574 \times 10^{123}$ in (50). Moreover, $(\log n)\sum_{i=1}^{n} x_i = 518.915$. Hence from (51), $I_{lost}(\mathbf{x}|T_1) = I_{lost}(\mathbf{x}|T_1') \approx 109.667$ bits of Shannon information. This value corresponds to 13.708 bytes at 8 bits per byte or to 0.013 kilobytes (KB) at 1024 bytes per kilobyte [14]. It thus follows from the discussion at the beginning of this example that

$$I_{lost}(\mathbf{x}|T_1) = I_{lost}(\mathbf{x}|T_1') \approx 0.013 \text{ KB}. \tag{52}$$

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

$$L(\mathbf{x}|\theta) = f(\mathbf{x}|\theta) = \frac{\theta^{\sum_{i=1}^{n} x_i}\, e^{-n\theta}}{\prod_{i=1}^{n} x_i!}, \tag{53}$$

from which

$$K(\mathbf{x}|\theta) = \theta^{\sum_{i=1}^{n} x_i}\, e^{-n\theta} \tag{54}$$

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 (54). Hence in the numerical example of **Case 1**, $I_{lost}(\mathbf{x}|K(\mathbf{x}|\theta)) = I_{lost}(\mathbf{x}|T_1) \approx 0.013$ KB from **Theorem 3.5** for all $\theta > 0$ except $\theta = 1$. For $\theta = 1$, $K(\mathbf{x}|\theta) = e^{-n}$ and is constant with respect to any data sample $\mathbf{x}$. Thus $I_{comp}(\mathbf{x}|1, K) = 0$ 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 $|A_{L(\mathbf{x}|\theta)}|$ and using (20). From (53), note that for all fixed $\theta > 0$ except $\theta = 1$, $\mathbf{y} \in A_{L(\mathbf{x}|\theta)}$ if and only if

$$\frac{\theta^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} = \frac{\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}. \tag{55}$$

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, (55) is satisfied.

Such complications suggest that an efficient implicit enumeration of the $\mathbf{y}$ satisfying (55) would be required to obtain $|A_{L(\mathbf{x}|\theta)}|$ for calculating $I_{lost}(\mathbf{x}|L(\mathbf{x}|\theta))$ from (20). Using such an algorithm, a conventional computer could probably 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 [15]. 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 seconds, whereas this task would have taken the above conventional supercomputer 10,000 years [16]. 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)**.**

We present some simple further simple computational results for the Poisson example distribution to illustrate relationships between among T_1, T_2, T_3. Table 5.1 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 $A_{L(\mathbf{x}|\theta)}$ gives $I_{lost}(\mathbf{x}|L(\mathbf{x}|\theta))$ from (20).

Table 5.1. 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)						

Example 5.2 (Binomial Distribution). Consider a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from a binomial random variable X with parameters 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 [9], $\sum_{i=1}^{n} X_i$ has a binomial distribution with parameter θ for fixed nm. Hence

$$P_\theta\left[\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} x_i\right] = \theta^{\sum_{i=1}^{n} x_i}\ \theta^{mn-\sum_{i=1}^{n} x_i}\binom{mn}{\sum_{i=1}^{n} x_i} \tag{56}$$

and

$$P_\theta[\mathbf{X} = \mathbf{x}] = \theta^{\sum_{i=1}^{n} x_i}\ \theta^{mn-\sum_{i=1}^{n} x_i}\prod_{i=1}^{n}\binom{m}{x_i}. \tag{57}$$

From (1), dividing (57) by (56) gives

$$P[\mathbf{X} = \mathbf{x}|\sum_{i=1}^{n} X_i = t] = \frac{\prod_{i=1}^{n}\binom{m}{x_i}}{\binom{mn}{t}}. \tag{58}$$

By taking the $-\log$ of (58) gives the lost information as

$$I_{lost}(\mathbf{x}|T_1) = -\log\frac{\prod_{i=1}^{n}\binom{m}{x_i}}{\binom{mn}{t}} = -\sum_{i=1}^{n}\log\binom{m}{x_i} + \log\binom{mn}{t}. \tag{59}$$

Case 2: $T_2(\mathbf{X}) = K(\mathbf{X}|\theta)$. In this case we use (16) as in **Example 5.1**. Write

$$L(\mathbf{x}|\theta) = f(\mathbf{x}|\theta) = \theta^{\sum_{i=1}^{n} x_i}\ (1-\theta)^{mn-\sum_{i=1}^{n} x_i}\prod_{i=1}^{n}\binom{m}{x_i}, \tag{60}$$

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,

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

and (61) yields

$$I_{lost}(\mathbf{x}|T_2) = -\sum_{i=1}^{n}\log\binom{m}{x_i} + \log \sum_{\mathbf{y}\in A_{K(\mathbf{x}|\theta)}}\prod_{i=1}^{n}\binom{m}{y_i}, \tag{62}$$

where

$$A_{K(\mathbf{x}|\theta)} = \left\{\mathbf{y}\in S^n \middle| \theta^{\sum_{i=1}^{n} y_i}\ (1-\theta)^{mn-\sum_{i=1}^{n} y_i} = \theta^{\sum_{i=1}^{n} x_i}\ (1-\theta)^{mn-\sum_{i=1}^{n} x_i}\right\}. \tag{63}$$

From (63), 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 (63) to compute (62). 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

$$L(\mathbf{x}|\theta) = \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^{n} x_i}(1-\theta)^{mn}\prod_{i=1}^{n}\binom{m}{x_i} \tag{64}$$

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

$$\left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^{n} y_i}\prod_{i=1}^{n}\binom{m}{y_i} = \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^{n} x_i}\prod_{i=1}^{n}\binom{m}{x_i}. \tag{65}$$

As in **Case 3** of **Example 5.1**, developing an algorithm to use (65) and determine $\left|A_{L(\mathbf{x}|\theta)}\right|$ for calculating $I_{lost}\left(\mathbf{x}\middle|L(\mathbf{x}|\theta)\right)$ 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 5.2 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 5.2. Binomial 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	$(1-\theta)^6$	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)						

(2,0,1)	3	$\log 10$	$(1-\theta)^3\theta^3$	$\log 10$	$2(1-\theta)^3\theta^3$	$\log 6$
(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 5.3 the entropic losses of **Example 5.2** 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 4.5** is corroborated. Also observe that $H_{lost}(\mathbf{X}, T)$ is smallest for the likelihood function T_3.

Table 5.3. Entropic loss over different statistics for a binomial distribution

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

Example 5.3 (Geometric Distribution). Consider a random sample $\mathbf{X} = (X_1, \dots, X_n)$ with sample data $\mathbf{x} = (x_1, \dots, x_n)$ 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

$$P[\mathbf{X} = \mathbf{x}] = \theta^n(1-\theta)^{\sum_{i=1}^n x_i - n}. \tag{66}$$

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. Hence,

$$P\left[\sum_{i=1}^n X_i = \sum_{i=1}^n x_i\right] = \binom{\sum_{i=1}^n x_i - 1}{n-1}\theta^n(1-\theta)^{\sum_{i=1}^n x_i - n}. \tag{67}$$

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 $h(x_1, \dots, x_n) = \binom{\sum_{i=1}^n x_i - 1}{n-1}$. Moreover, substitution of (66) and (67) into (8) gives

$$P\left[\mathbf{X} = \mathbf{x} \middle| \sum_{i=1}^n X_i = \sum_{i=1}^n x_i\right] = \frac{1}{\binom{\sum_{i=1}^n x_i - 1}{n-1}}. \tag{68}$$

Then from (14) and (68) we obtain that

$$I_{lost}(\mathbf{x}|T_1) = \log\binom{\sum_{i=1}^n x_i - 1}{n-1}. \tag{69}$$

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

$$K(\mathbf{x}|\theta) = L(\mathbf{x}|\theta) = \left(\frac{\theta}{1-\theta}\right)^n (1-\theta)^{\sum_{i=1}^n x_i}. \tag{70}$$

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 (70). Thus from **Theorem 3.5**, $I_{lost}(\mathbf{x}|T_2(\mathbf{x})) = I_{lost}(\mathbf{x}|T_1)$ as given in (69).

Case 3: $T_3(\mathbf{X}) = L(\mathbf{X}|\theta)$. Since $K(\mathbf{X}|\theta) = L(\mathbf{X}|\theta)$ from (70), then

$$I_{lost}(\mathbf{x}|T_3) = \log\binom{\sum_{i=1}^{n} x_i - 1}{n-1} \tag{71}$$

from (69). However, there is an alternate derivation of (71). For $0< \theta < 1$ it follows from (70) that then $\mathbf{y} \in A_{L(\mathbf{x}|\theta)}$ if and only if

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i. \tag{72}$$

But for fixed positive integers $x_1, \ldots, x_n$ we have from [17] that the number of solutions $|A_{L(\mathbf{x}|\theta)}|$ to (72) in positive integers $y_1, \ldots, y_n$ is

$$\binom{\sum_{i=1}^{n} x_i - 1}{n-1}. \tag{73}$$

Thus (71) follows for $L(\mathbf{X}|\theta)$ from (73) and (20), so $I_{lost}(\mathbf{x}|T_1) = I_{lost}(\mathbf{x}|T_2) = I_{lost}(\mathbf{x}|T_3)$ from **Theorem 3.5**.

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 with the parameter θ as the probability of getting a head on any flip. Suppose this experiment is performed three times yielding the possible sample data $\mathbf{x} = (x_1, x_2, x_3)$ shown in Table 5.4. $I_{lost}(\mathbf{x}|T)$ is then calculated for each of the sufficient statistics for θ of **Example 5.3**. 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 the sample data as established analytically above.

Table 5.4. 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. Conclusion

In this paper, the Shannon information obtained from a random sample $\mathbf{X}$ for a discrete random variable X with a single parameter θ was decomposed into two components: (i) the compressed information obtained by the value of a real-valued statistic $T(\mathbf{X})$ for the sample data $\mathbf{x}$ and (ii) the information lost by using this statistic to characterize $\mathbf{X}$. We focused on this lost information caused by multiple data sets having the same value of the statistic. This possibility is typical of data analysis, where the data uniquely determines the value of the statistic, but a value of the statistic does not uniquely determine the data yielding it. In other words, we answered the question: how much Shannon information is lost about a data sample when only the value of a sufficient statistic is known but not the original data. We also defined the entropic loss associated with a sufficient statistic T under consideration as the expected lost information over all possible samples to give a metric

dependent only on T. Our approach is applicable to any T, but we focused on sufficient statistics for θ for simplicity. Applications of our results were computationally intensive.

References

1. Landauer, R. Irreversibility and heat generation in the computing process. IBM Journal of Research and Development. 1961, 5, 183–191.
2. Hodge, S.E.; Vieland, V.J. Information loss in binomial data due to data compression. Entropy 2017, 19, 75-81.
3. Casella, G.; Berger, R.L. Statistical Inference, 2nd ed.; Cengage Learning, Delhi, India, 2002.
4. Pawitan, Y. In All Likelihood: Statistical Modeling and Inference Using Likelihood, 1st ed.; The Clarendon Press: Oxford, UK, 2013.
5. Rohatgi, V.K.; Ehsanes Saleh, A.K. An Introduction to Probability and Statistics, 2nd ed.; John Wiley & Sons, Inc., NY, USA, 2001.
6. Shannon, C.E.; Weaver, W. The Mathematical Theory of Communication, 1st ed.; The University of Illinois Press, Urbana, Illinois, 1964.
7. Shannon, C. A mathematical theory of communication. Bell. Syst. Tech. J. 1948, 27, 379–423.
8. Kapur, J.N.; Kesavan, H.K. Entropy Optimization Principles with Applications, 1st ed.; Academic Press, Inc., San Diego, CA, USA, 1992.
9. Johnson, J.L. Probability and Statistics for Computer Science, 1st ed.; John Wiley & Sons, Inc., NJ, USA, 2003.
10. Beeler, R.A. How to Count: An Introduction to Combinatorics and Its Applications, 1st ed.; Springer, Switzerland, 2015.
11. https://en.wikipedia.org/wiki/Factorial#Computation, accessed 7/1/2019.
12. Mortici, C. Ramanujan formula for the generalized Stirling approximation. Applied Mathematics and Computation 2010, 19, 2579-2585.
13. https://mathcracker.com/multinomial-coefficient-calculator.php, accessed 7/27/2019.
14. https://en.wikipedia.org/wiki/Kilobyte, accessed 7/27/2019.
15. https://en.wikipedia.org/wiki/Summit_(supercomputer), accessed 7/1/2019.
16. Arute, F.; Arya, K.; Martinis, J.M. Quantum supremacy using a programmable superconducting processor. Nature 2019, 574, 505-510.
17. Mahmoudvand, R.; Hassani, H.; Farzaneh, A.; Howell.G. The exact number of nonnegative integer solutions for a linear Diophantine inequality. IAENG International Journal of Applied Mathematics 2010, 40, 5 pages.