entropy

MDPI

Article

Universality of Logarithmic Loss in Fixed-Length Lossy Compression †

Albert No

Department of Electronic and Electrical Engineering, Hongik University, Seoul 04066, Korea; albertno@hongik.ac.kr; Tel.: +82-2-320-1649

† This paper is an extended version of our paper published in the 2015 IEEE International Symposium on Information Theory (ISIT), Hong Kong, China, 14–19 June 2015.

Received: 8 April 2019; Accepted: 8 June 2019; Published: 10 June 2019

Abstract: We established a universality of logarithmic loss over a finite alphabet as a distortion criterion in fixed-length lossy compression. For any fixed-length lossy-compression problem under an arbitrary distortion criterion, we show that there is an equivalent lossy-compression problem under logarithmic loss. The equivalence is in the strong sense that we show that finding good schemes in corresponding lossy compression under logarithmic loss is essentially equivalent to finding good schemes in the original problem. This equivalence relation also provides an algebraic structure in the reconstruction alphabet, which allows us to use known techniques in the clustering literature. Furthermore, our result naturally suggests a new clustering algorithm in the categorical data-clustering problem.

Keywords: categorical data clustering; fixed-length lossy compression; logarithmic loss; rate-distortion

1. Introduction

Logarithmic loss is a unique distortion measure in the sense that it allows a "soft" estimation (or reconstruction) of the source. Although logarithmic loss plays a crucial role in learning theory, not much work has been published regarding lossy compression until recently. A few exceptions are a line of work on multiterminal source coding [1–3], the single-shot approach to lossy source coding under logarithmic loss [4], and several universal properties of logarithmic loss in information theory [5–7]. In [4], Shkel and Verdú focused on the lossy-compression problem when the distortion measure is given by logarithmic loss. On the other hand, Jiao et al. justified logarithmic loss by showing it is the only loss function that satisfies a natural data-processing requirement [5]. Painsky and Wornell provided a universal property of logarithmic loss in the context of classification. In [7], No focused on the universal property of logarithmic loss in the successive refinement problem. We would also like to point out that the information bottleneck method [8–11] is related to lossy compression under logarithmic loss. Indeed, it is equivalent to the noisy lossy-compression problem under logarithmic loss [12].

In this paper, we present a new universal property of logarithmic loss in fixed-length lossy-compression problems. Consider an arbitrary fixed-length lossy-compression problem, where source and reconstruction alphabets $\mathcal{X}$ and $\hat{\mathcal{X}}$ are discrete. Suppose arbitrary distortion measure $d : \mathcal{X} \times \hat{\mathcal{X}}$ is given. Then, we show that there exists a corresponding fixed-length lossy-compression problem where the source alphabet remains the same, but the reconstruction alphabet is a set of distributions on $\mathcal{X}$, and the distortion measure is logarithmic loss. This implies that there is a correspondence between any fixed-length lossy-compression problem under an arbitrary distortion measure and that under logarithmic loss. The correspondence is in the following strong sense:

- optimal schemes for the two problems are the same; and
- a good scheme for one problem is also a good scheme for the other.

We are more precise about the "optimal" and "goodness" of the scheme in later sections. This finding essentially implies that it is enough to consider the lossy-compression problem under logarithmic loss.

The above correspondence provides new insights into the fixed-length lossy-compression problem. In general, the reconstruction alphabet in the lossy-compression problem does not have any well-defined operations. However, in the corresponding lossy compression under logarithmic loss, reconstruction symbols are probability distributions that have their own algebraic structure. Thus, under the corresponding setting, we can apply various techniques, such as the information geometric approach, clustering with Bregman divergence, and relaxation of the optimization problem. Furthermore, the equivalence relation suggests a new algorithm in the categorical data-clustering problem, where data are not in the continuous space.

The remainder of the paper is organized as follows. In Section 2, we revisit some of the known results of logarithmic loss and fixed-length lossy compression. Section 3 is dedicated to the equivalence between lossy compression under arbitrary distortion measures and that under logarithmic loss. In Section 4, we present the geometric interpretation of our result. We provide the log-convex relaxation of lossy compression and connection to the clustering problems in Section 5. Finally, we conclude in Section 6.

Notation: Uppercase X denotes a random variable, where $\mathcal{X}$ denotes a set of alphabet. On the other hand, lowercase x denotes a specific possible realization of random variable X, i.e., $x \in \mathcal{X}$. Similarly, X^n denotes an n-dimensional random vector $(X_1, X_2, \ldots, X_n)$ while lowercase x^n denotes a realization of X^n. The absolute value of function $|f|$ denotes a size of image of function $f : \mathcal{X} \to \mathcal{Y}$, i.e., $|\{f(x) : x \in \mathcal{X}\}|$. If it was clear from the context, we used $\sum_x$ instead of $\sum_{x\in\mathcal{X}}$. We used a natural logarithm and nats instead of bits.

2. Preliminaries

2.1. Logarithmic Loss

Suppose $\mathcal{X}$ is a finite set of discrete symbols, and $\mathcal{M}(\mathcal{X})$ is the set of probability measures on $\mathcal{X}$. For $x \in \mathcal{X}$ and $q \in \mathcal{M}(\mathcal{X})$, the definition of logarithmic loss $\ell : \mathcal{X} \times \mathcal{M}(\mathcal{X}) \to [0, \infty]$ is given by

$$\ell(x, q) = \log \frac{1}{q(x)}.$$

2.2. Fixed-Length Lossy Compression

In this section, we briefly introduce the basic settings of the fixed-length lossy-compression problem [13]. In a fixed-length lossy-compression setting, we have a source X with finite alphabet $\mathcal{X} = \{1, \ldots, r\}$ and source distribution p_X. An encoder $f : \mathcal{X} \to \{1, \ldots, M\}$ maps the source symbol to one of M messages. On the other side, a decoder $g : \{1, \ldots, M\} \to \hat{\mathcal{X}}$ maps the message to actual reconstruction $\hat{X}$, where the reconstruction alphabet is also finite $\hat{\mathcal{X}} = \{1, \ldots, s\}$. Let $d : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$ be a distortion measure between source and reconstruction.

First, we can define the code that the expected distortion is lower than a given distortion level.

Definition 1 (Average distortion criterion)**.** *An (M, D) code is a pair of an encoder f with $|f| \leq M$ and a decoder g, such that*

$$\mathbb{E}\left[d(X, g(f(X)))\right] \leq D.$$

The minimum number of codewords required to achieve average distortion not exceeding D is defined by

$$M^\star(D) = \min\{M : \exists(M, D)\ code\}.$$

Similarly, we can define the minimum achievable average distortion given number of codewords M.

$$D^\star(M) = \min\{D : \exists(M, D)\ code\}.$$

One may consider a stronger criterion that restricts the probability of exceeding a given distortion level.

Definition 2 (Excess distortion criterion)**.** *An (M, D, ϵ) code is a pair of an encoder f with $|f| \leq M$ and a decoder g such that*

$$Pr\left[d(X, g(f(X))) > D\right] \leq \epsilon.$$

The minimum number of codewords required to achieve excess distortion probability ϵ, and distortion D is defined by

$$M^\star(D, \epsilon) = \min\{M : \exists(M, D, \epsilon)\ code\}.$$

Similarly, we can define the minimum achievable excess distortion probability given target distortion D and number of codewords M.

$$\epsilon^\star(M, \epsilon) = \min\{\epsilon : \exists(M, D, \epsilon)\ code\}.$$

Given target distortion D and p_X, the information rate-distortion function is defined by

$$R(D) = \inf_{p_{\hat{X}|X}:\mathbb{E}\left[d(X,\hat{X})\right]\leq D} I(X; \hat{X}) \tag{1}$$

We make the following benign assumptions:

- There exists a unique rate-distortion function achieving conditional distribution $p^\star_{\hat{X}|X}$.
- We assume that $p^\star_{\hat{X}}(\hat{x}) > 0$ for all $\hat{x} \in \hat{\mathcal{X}}$ since we can always discard the reconstruction symbol with zero probability.
- If $d(x, \hat{x}_1) = d(x, \hat{x}_2)$ for all $x \in \mathcal{X}$, then $\hat{x}_1 = \hat{x}_2$. (If $d(x, \hat{x}_1) = d(x, \hat{x}_2)$ for all x, then, there is no difference between $\hat{x}_1$ and $\hat{x}_2$ in terms of loss. Thus, we can always discard $\hat{x}_2$ without loss of generality.)

2.3. D-Tilted Information

Define the information density of joint distribution $p_{X,\hat{X}}$ by

$$\imath_{X;\hat{X}}(x; \hat{x}) = \log \frac{p_{X,\hat{X}}(x, \hat{x})}{p_X(x) p_{\hat{X}}(\hat{x})}.$$

Then, we are ready to define D-tilted information that plays a key role in fixed-length lossy compression.

Definition 3 ([13] (Definition 6))**.** *The D-tilted information in $x \in \mathcal{X}$ is defined as*

$$\jmath_X(x, D) = \log \frac{1}{\mathbb{E}\left[\exp\left(\lambda^\star D - \lambda^\star d(x, \hat{X}^\star)\right)\right]}$$

where the expectation is with respect to the marginal distribution of $\hat{X}^\star$ *and* $\lambda^\star = -R'(D)$.

Note that $\hat{X}^\star$ is a random variable that has a marginal distribution of $p_X \times p^\star_{\hat{X}|X}$, and $R'(D)$ is the first derivative of rate-distortion function $R(D)$.

Theorem 1 ([14] (Lemma 1.4)). *For all* $\hat{x} \in \hat{\mathcal{X}}$,

$$\jmath_X(x, D) = \imath_{X;\hat{X}^\star}(x; \hat{x}) + \lambda^\star d(x, \hat{x}) - \lambda^\star D; \tag{2}$$

therefore, we have

$$R(D) = \mathbb{E}\left[\jmath(X, D)\right].$$

Let $p^\star_{X|\hat{X}}$ be the induced conditional probability from $p^\star_{\hat{X}|X}$. Then, (2) can equivalently be expressed as

$$\begin{aligned} &\log \frac{1}{p^\star_{X|\hat{X}}(x|\hat{x})} \\ &= \log \frac{1}{p_X(x)} - \jmath_X(x, D) + \lambda^\star d(x, \hat{x}) - \lambda^\star D. \end{aligned} \tag{3}$$

The following lemma shows that $p^\star_{X|\hat{X}}(\cdot|\hat{x})$ are all distinct.

Lemma 1 ([7] (Lemma 2)). *For all* $\hat{x}_1 \neq \hat{x}_2$, *there exists* $x \in \mathcal{X}$ *such that* $p^\star_{X|\hat{X}}(x|\hat{x}_1) \neq p^\star_{X|\hat{X}}(x|\hat{x}_2)$.

3. One-to-One Correspondence Between General Distortion and Logarithmic Loss

3.1. Main Results

Consider fixed-length lossy compression under arbitrary distortion $d(\cdot, \cdot)$, as described in Section 2.2. We have a source X with finite alphabet $\mathcal{X} = \{1, \ldots, r\}$, source distribution p_X, and finite reconstruction alphabet $\hat{\mathcal{X}} = \{1, \ldots, s\}$. For a fixed number of messages M, let $f^\star$ and $g^\star$ be the encoder and decoder that achieve optimal average distortion $D^\star(M)$, i.e.,

$$\mathbb{E}\left[d(X, g^\star(f^\star(X)))\right] = D^\star(M).$$

Let $p^\star_{\hat{X}|X}$ denote the rate-distortion function achieving conditional distribution at distortion $D = D^\star(M)$. In other words, $p_X \times p^\star_{\hat{X}|X}$ achieves the infimum in

$$R(D^\star(M)) = \inf_{p_{\hat{X}|X}:\mathbb{E}\left[d(X,\hat{X})\right] \leq D^\star(M)} I(X; \hat{X}). \tag{4}$$

Note that $R(D^\star(M))$ may be strictly smaller than $\log M$ in general since $R(\cdot)$ is an information rate-distortion function that does not characterize the best achievable performance for the "one-shot" setting in which $D^\star(M)$ is defined.

Now, we define the corresponding fixed-length lossy-compression problem under logarithmic loss. In the corresponding problem, source alphabet $\mathcal{X} = \{1, \ldots, r\}$, source distribution p_X, and number of messages M remain the same. However, we have a different reconstruction alphabet $\mathcal{Y} = \{p^\star_{X|\hat{X}}(\cdot|\hat{x}) : \hat{x} \in \hat{\mathcal{X}}\} \subset \mathcal{M}(\mathcal{X})$ where $p^\star$ pertains to the achiever of the infimum in Equation (4) associated with the original loss function. Recall that $\mathcal{M}(\mathcal{X})$ is the set of all probability measures on $\mathcal{X}$. Let the distortion of the corresponding problem be the logarithmic loss.

We now further connect the encoding and decoding schemes between the two problems. Suppose $f : \mathcal{X} \to \{1, \ldots, M\}$ and $g : \{1, \ldots, M\} \to \hat{\mathcal{X}}$ are an encoder and decoder pair in the original problem. When f and g are given in the original problem, we define the corresponding encoder and decoder in the corresponding problem as follows. We let the encoder be the same $f_\ell = f$, and define the decoder $g_\ell : \{1, \ldots, M\} \to \mathcal{Y}$ by

$$g_\ell(m) = p^\star_{X|\hat{X}}(\cdot|g(m)).$$

Then, f_ℓ and g_ℓ are a valid encoder and decoder pair for the corresponding fixed-length lossy-compression problem under logarithmic loss. Conversely, given f_ℓ and g_ℓ, we can find corresponding f and g because Lemma 1 guarantees that $p_{X|\hat{X}}(\cdot|\hat{x})$ are distinct.

The following result shows the relation between the corresponding schemes.

Theorem 2. *For any encoder–decoder pair (f_ℓ, g_ℓ) for the corresponding fixed-length lossy-compression problem under logarithmic loss, we have*

$$\begin{aligned}
&\mathbb{E}\left[\ell(X, g_\ell(f_\ell(X)))\right] \\
&= H(X|\hat{X}^\star) + \lambda^\star \left(\mathbb{E}\left[d(X, g(f(X)))\right] - D^\star(M)\right) \\
&\geq H(X|\hat{X}^\star)
\end{aligned}$$

where (f, g) is the corresponding encoder–decoder pair for the original lossy-compression problem. Note that $H(X|\hat{X}^\star)$ and the expectations are with respect to distribution $p_X \times p^\star_{\hat{X}|X}$. Moreover, equality holds if and only if $f_\ell = f^\star$ and $g_\ell(m) = p^\star_{X|\hat{X}}(\cdot|g^\star(m))$.

Proof. We have

$$\begin{aligned}
&\mathbb{E}\left[\ell(X, g_\ell(f_\ell(X)))\right] \\
&= \mathbb{E}\left[\ell\left(X, p^\star_{X|\hat{X}}(\cdot|g(f(X)))\right)\right] \\
&= \mathbb{E}\left[\log \frac{1}{p^\star_{X|\hat{X}}(X|g(f(X)))}\right].
\end{aligned}$$

Then, Equation (3) implies that

$$\begin{aligned}
&\mathbb{E}\left[\ell(X, g_\ell(f_\ell(X)))\right] \\
&= \mathbb{E}\left[\log \frac{1}{p_X(X)} - \jmath_X(X, D^\star(M))\right] \\
&\quad + \mathbb{E}\left[\lambda^\star d(X, g(f(X))) - \lambda^\star D^\star(M)\right] \\
&= H(X|\hat{X}^\star) + \lambda^\star \left(\mathbb{E}\left[d(X, g(f(X)))\right] - D^\star(M)\right) \qquad (5) \\
&\geq H(X|\hat{X}^\star) \qquad (6)
\end{aligned}$$

where Equation (5) is because $\mathbb{E}\left[\jmath_X(X, D^\star(M))\right] = R(D^\star(M)) = I(X; \hat{X}^\star)$ with respect to distribution $p_X \times p^\star_{\hat{X}|X}$. Inequality (6) is because $D^\star(M)$ is the minimum achievable average distortion with M codewords. Equality holds if and only if $\mathbb{E}\left[d(X, g(f(X)))\right] = D^\star(M)$, which can be achieved by the optimal scheme for the original lossy-compression problem. In other words, the equality holds if

$$\begin{aligned}
f^\star_\ell &= f^\star \\
g^\star_\ell(m) &= p^\star_{X|\hat{X}}(\cdot|g^\star(m)).
\end{aligned}$$

□

In the above theorem, distortion $D^\star(M)$ plays a critical role, which is the minimal achievable distortion in the one-shot setting. We also used $p^\star_{X|\hat{X}}$ in the corresponding problem, which is the rate-distortion-achieving conditional distribution. This might be confusing since the rate-distortion function provides the optimal rate in the asymptotic setting. However, recall that the minimal mutual information between X and $\hat{X}$ in Equation (1) is the "information" rate-distortion function. The "information" rate-distortion function is equal to the optimum rate in the asymptotic case if the source is independent and identically distributed.

On the other hand, we viewed the "information" rate-distortion function differently. We considered the one-shot setting where source X and reconstruction $\hat{X}$ are single variables. Given number of messages M, the minimal achievable distortion is given by $D^\star(M)$. Under this setting, we focused on minimal mutual information between X and $\hat{X}$ when the distortion between X and $\hat{X}$ is restricted by $D^\star(M)$. Our theorem implies that minimal achieving distribution $p^\star_{X|\hat{X}}$ provides the corresponding one-shot lossy-compression problem under logarithmic loss.

Remark 1. *In the corresponding fixed-length lossy-compression problem under logarithmic loss, the minimal achievable average distortion given number of codewords M is*

$$D^\star_\ell(M) = H(X|\hat{X}^\star)$$

where the conditional entropy is with respect to distribution $p_X \times p^\star_{\hat{X}|X}$.

Remark 2. *From now on, we denote the original lossy-compression problem under given distortion measure $d(\cdot,\cdot)$ with reconstruction alphabet $\hat{\mathcal{X}}$ by "original problem". On the other hand, we denote the corresponding lossy-compression problem under logarithmic loss with reconstruction alphabet $\mathcal{Y}$ by "corresponding problem".*

3.2. Example: Memoryless Bernoulli Source with Hamming Distortion Measure

In this section, we consider the memoryless Bernoulli source under Hamming distortion measure as an example of the above equivalence. Let $X = U^n$ be a memoryless Bernoulli source with probability α, where $\mathcal{X} = \mathcal{U}^n = \{0,1\}^n$, and reconstruction $\hat{X} = V^n$ is also an n-dimensional binary vector where $\hat{\mathcal{X}} = \mathcal{V}^n = \{0,1\}^n$. Note that block length n is fixed, so the problem is in the one-shot setting. Distortion measure d is separable Hamming distortion, i.e.,

$$d(X,\hat{X}) = d_H(U^n, V^n) = \frac{1}{n}\sum_{i=1}^n d_H(U_i, V_i)$$

where $d_H(u,v) = 1$ if $u \neq v$ and $d_H(u,v) = 0$ if $u = v$. Let M be the number of messages. Then, we are interested in optimal encoding and decoding schemes that achieve distortion $D = D^\star(M)$.

In this scenario, the information rate-distortion function is not hard to compute [15]:

$$\begin{aligned} R(D) =& \inf_{p_{\hat{X}|X}:\mathbb{E}\left[d(X,\hat{X})\right]\leq D} I(X;\hat{X}) \\ =& \inf_{p_{U^n|V^n}:\mathbb{E}[d_H(U^n,V^n)]\leq D} I(U^n;V^n) \\ =& n \inf_{p_{U|V}:\mathbb{E}[d_H(U,V)]\leq D} I(U;V) \qquad (7) \\ =& n(h_2(\alpha) - h_2(D)), \qquad (8) \end{aligned}$$

where $h_2(\cdot)$ is the binary entropy function. Let $p^\star_{U|V}$ be the distribution that achieves the infimum in Equation (7). We have an analytic formula for rate-distortion-achieving distribution $p^\star_{X|\hat{X}}$. For $x = u^n$ and $\hat{x} = v^n$, we have

$$\begin{aligned} p^{\star}_{X|\hat{X}}(x|\hat{x}) =& \prod_{i=1}^{n} p^{\star}_{U|V}(u_i|v_i) \\ =& \prod_{i=1}^{n} D^{d_H(u_i,v_i)}(1-D)^{1-d_H(u_i,v_i)} \\ =& (1-D)^n \left(\frac{D}{1-D}\right)^{nd_H(u^n,v^n)} \\ =& (1-D)^n \left(\frac{D}{1-D}\right)^{nd(x,\hat{x})}. \end{aligned}$$

Then, the corresponding problem is the rate-distortion problem under logarithmic loss where the set of reconstruction symbols is

$$\mathcal{Y} = \{p^{\star}_{X|\hat{X}}(\cdot|\hat{x}) : \hat{x} \in \mathcal{V}^n\}.$$

Remark 3. *We can rewrite Equation* (3) *in this case.*

$$\begin{aligned} \ell(x, p^{\star}_{X|\hat{X}}(\cdot|\hat{x})) =& \log \frac{1}{p^{\star}_{X|\hat{X}}(x|\hat{x})} \\ =& n \log \frac{1}{1-D} + nd(x,\hat{x}) \log \frac{1-D}{D}. \end{aligned}$$

The above equation explicitly shows the correspondence between logarithmic loss and the original distortion measure.

3.3. Discussion

3.3.1. One-to-One Correspondence

Theorem 2 implies that, for any fixed-length lossy-compression problem, we can find an equivalent problem under logarithmic loss where optimal encoding schemes are the same. Thus, without loss of generality, we can restrict our attention to the problem under logarithmic loss with reconstruction alphabet $\mathcal{Y} = \{q^{(1)}, \ldots, q^{(s)}\}$ for some $q^{(1)}, \ldots, q^{(s)} \in \mathcal{M}(\mathcal{X})$.

3.3.2. Scheme Suboptimality

Suppose f and g are a suboptimal encoder and decoder for the original fixed-length lossy-compression problem. Then, the theorem implies

$$\begin{aligned} &\mathbb{E}\left[\ell(X, g_\ell(X))\right] - H(X|\hat{X}^{\star}) \\ &= \lambda^{\star}\left(\mathbb{E}\left[d(X, g(f(X)))\right] - D^{\star}(M)\right). \end{aligned} \tag{9}$$

The left-hand side of Equation (9) is the cost of suboptimality for the corresponding lossy-compression problem. On the other hand, the right-hand side is proportional to the cost of suboptimality for the original problem. In Section 3.3.1, we discussed that the optimal schemes of the two problems coincide. Equation (9) shows stronger equivalence in which costs of suboptimalities are linearly related. This implies that a good code for one problem is also good for the other.

3.3.3. Operations on the Reconstruction Alphabet

In general, reconstruction alphabet $\hat{\mathcal{X}}$ does not have an algebraic structure. However, in the corresponding rate-distortion problem, the reconstruction alphabet is the set of probability measures

where we have natural operations such as convex combinations of elements or projection to a convex hull. We discuss such operations closer in Section 5.

3.4. Exact Performance of Optimal Scheme

In the previous section, we showed that there is a corresponding lossy-compression problem under logarithmic loss that shares the same optimal coding scheme. In this section, we investigate the exact performance of the optimal scheme for the fixed-length lossy-compression problem under logarithmic loss, when the reconstruction alphabet is the set of all probability measures on $\mathcal{X}$, i.e., $\mathcal{M}(\mathcal{X})$. (Recently, Shkel and Verdu [4] independently proposed similar results. The result was also presented in our conference version of the paper [16].) We also characterize minimal average distortion $D^\star(M)$ when we have a fixed number of messages M. Note that this is a single-letter version of ([2], [Lemma 1]). Although the optimal scheme associated with $\mathcal{M}(\mathcal{X})$ may differ from the optimal scheme with restricted reconstruction alphabets $\mathcal{Y}$, it provides an insight, as we show in Section 4. In this section, we restrict our attention to deterministic schemes. However, it is not hard to show that the same result holds even if we allow a stochastic encoder and decoder.

Let an encoder and a decoder be $f : \mathcal{X} \to \{1, \dots, M\}$ and $g : \{1, \dots, M\} \to \mathcal{M}(\mathcal{X})$ where $g(m) = q^{(m)} \in \mathcal{M}(\mathcal{X})$. Then, we have

$$\begin{aligned}
&\mathbb{E}\left[\ell(X, g(f(X)))\right] \\
&= \sum_{x \in \mathcal{X}} p_X(x) \log \frac{1}{q^{(f(x))}(x)} \\
&= H(X) + \sum_{m=1}^{M} \sum_{x \in f^{-1}(m)} p_X(x) \log \frac{p_X(x)}{q^{(m)}(x)} \\
&= H(X) + \sum_{m=1}^{M} u_m \log u_m \\
&\quad + \sum_{m=1}^{M} u_m \sum_{x \in f^{-1}(m)} \frac{p_X(x)}{u_m} \log \frac{p_X(x)/u_m}{q^{(m)}(x)},
\end{aligned}$$

where $f^{-1}(m) = \{x \in \mathcal{X} : f(x) = m\}$ and $u_m = \sum_{x \in f^{-1}(m)} p_X(x)$. Since $p_{X|f(X)}(x|m) = \frac{p_X(x)}{u_m}$ for all $x \in f^{-1}(m)$, we have

$$\begin{aligned}
&\mathbb{E}\left[\ell(X, g(f(X)))\right] \\
&= H(X) - H(f(X)) \\
&\quad + \sum_{m=1}^{M} u_m D\left(p_{X|f(X)}(\cdot|m) \,\middle\|\, q^{(m)}\right) \\
&\geq H(X) - H(f(X)).
\end{aligned}$$

Equality can be achieved by choosing $q^{(m)} = p_{X|f(X)}(\cdot|m)$, which can be done no matter what f is. Thus, we have

$$D^\star(M) = H(X) - \max_{f : |f| \leq M} H(f(X)).$$

This implies that the optimal encoder is function f that maximizes $H(f(X))$, and the optimal decoder is given by $g(m) = p_{X|f(X)}(\cdot|m)$. The above result provides a trivial lower bound:

$$D^\star(M) \geq H(X) - \log M.$$

The optimal scheme under an excess distortion criterion is given in Appendix A.

4. Geometrical Interpretation

In this section, we present another geometrical interpretation of the decoder in lossy-compression problems. Consider the original lossy-compression problem with discrete reconstruction alphabet $\hat{\mathcal{X}}$ and distortion measure $d(\cdot,\cdot)$. Suppose encoding function f is given that may or may not be optimal, where $|f| = M$. Let $A_m = f^{-1}(m) = \{x \in \mathcal{X} : f(x) = m\}$, which is the set of source symbols that are mapped to message m. Then, optimal reconstruction $g(m)$ is given by

$$g(m) = \operatorname*{argmin}_{\hat{x}\in\hat{\mathcal{X}}} \mathbb{E}\left[d(X,\hat{x}) \mid X \in A_m\right]. \tag{10}$$

Now, consider the corresponding lossy-compression problem under logarithmic loss. Recall that the set of reconstruction alphabets is given by

$$\mathcal{Y} = \{p^{\star}_{X|\hat{X}}(\cdot|\hat{x}) : \hat{x} \in \hat{\mathcal{X}}\}$$

where $\mathcal{Y} \subset \mathcal{M}(\mathcal{X})$. As we have seen in Section 3.4, the optimal reconstruction is $g^E_\ell(m) = p_{X|f(X)}(\cdot|m)$ if we have extended set of reconstruction alphabet $\mathcal{M}(\mathcal{X})$. Thus, it is natural to find the probability distribution in $\mathcal{Y}$, which is the nearest distribution from $g^E_\ell(m)$. We propose Kullback–Leibler divergence to measure the distance between probability distributions. In other words, we want to find $\tilde{g}_\ell(m) \in \mathcal{Y}$, such that

$$\tilde{g}_\ell(m) = \operatorname*{argmin}_{q\in\mathcal{Y}} D(g^E_\ell(m)\|q). \tag{11}$$

This can be viewed as projecting the optimal solution from extended set $\mathcal{M}(\mathcal{X})$ to original feasible set $\mathcal{Y}$. Since $q \in \mathcal{Y}$, there exists $\hat{x} \in \hat{\mathcal{X}}$, such that $q(\cdot) = p^{\star}_{X|\hat{X}}(\cdot|\hat{x})$. Then, the above Kullback–Leibler divergence is given by

$$\begin{aligned}
&D(p_{X|f(X)}(\cdot|m)\|p^{\star}_{X|\hat{X}}(\cdot|\hat{x}))\\
&= \sum_{x\in A_m} p_{X|f(X)}(x|m)\log\frac{p_{X|f(X)}(x|m)}{p^{\star}_{X|\hat{X}}(x|\hat{x})}\\
&= \sum_{x\in A_m} \frac{p_X(x)}{\Pr[X\in A_m]}\log\frac{p_X(x)}{\Pr[X\in A_m]p^{\star}_{X|\hat{X}}(x|\hat{x})}\\
&= \log\frac{1}{\Pr[X\in A_m]} + \sum_{x\in A_m}\frac{p_X(x)}{\Pr[X\in A_m]}\log\frac{p_X(x)}{p^{\star}_{X|\hat{X}}(x|\hat{x})}\\
&= \log\frac{1}{\Pr[X\in A_m]} + \sum_{x\in A_m}\frac{p_X(x)}{\Pr[X\in A_m]}\left(-\jmath(x,D) + \lambda^{\star}d(x,\hat{x}) - \lambda^{\star}D\right),
\end{aligned}$$

where the last equality is from Equation (2). Note that $d(x,\hat{x})$ is the only term that is a function of $\hat{x}$, and $\lambda^{\star}$ is positive. Thus, if $q(\cdot) = p_{X|\hat{X}}(\cdot|\hat{x})$ achieves the minimum in Equation (11), then $\hat{x}$ minimizes the following:

$$\sum_{x\in A_m}\frac{p_X(x)}{\Pr[X\in A_m]}d(x,\hat{x}) = \mathbb{E}\left[d(X,\hat{x}) \mid X\in A_m\right]. \tag{12}$$

Since Equation (12) coincides with Equation (10), we have

$$\tilde{g}_\ell(m) = p^{\star}_{X|\hat{X}}(\cdot|g(m)). \tag{13}$$

Remark 4. *In Section 3, we directly defined* $g_\ell(m) = p^\star_{X|\hat{X}}(\cdot|g(m))$. *However, we obtained* $\tilde{g}_\ell(m)$ *via the following two-step procedure:*

- *extend the reconstruction set from* $\mathcal{Y}$ *to* $\mathcal{M}(\mathcal{X})$, *then characterize optimal decoding functions* $g^E_\ell(m) \in \mathcal{M}(\mathcal{X})$; *and*
- *find the measure* $\tilde{g}_\ell(m) \in \mathcal{Y}$ *that is closest to* $g^E_\ell(m)$.

The above result (13) *implies that* $\tilde{g}_\ell(m) = g_\ell(m)$.

5. Log-Convex Relaxation

In the previous section, we obtained the optimal reconstruction symbol from the extended reconstruction alphabet, and projected it to the feasible set. In this section, instead of direct projection to $\mathcal{Y}$, we propose another slight extension of $\mathcal{Y}$, namely, log-convex hull. As we show in the following sections, the log-convex hull has interesting properties.

5.1. rI-Projection

Before defining the log-convex hull, we need to define the log-convex combination of probability distributions. Let p and q be probability distributions in $\mathcal{M}(\mathcal{X})$. For $0 < t < 1$, the log-convex combination of p and q is given by

$$\overline{p^t q^{1-t}}(x) = \frac{p(x)^t q(x)^{1-t}}{\sum_{\tilde{x}} p(\tilde{x})^t q(\tilde{x})^{1-t}}. \tag{14}$$

It is clear to see that $\log \overline{p^t q^{1-t}}$ is a convex combination of $\log p(x)$ and $\log q(x)$ with a normalizing constant. We can now define log-convex hull $\mathrm{logconv}(\mathcal{Y})$ that is a set of log-convex combination of probability measures in set $\mathcal{Y}$. More precisely,

$$\mathrm{logconv}(\mathcal{Y}) = \left\{ q^{(r)} \in \mathcal{M}(\mathcal{X}) : q^{(r)}(x) = \frac{1}{c(r)} \exp\left(\sum_{\hat{x}} r(\hat{x}) \log p^\star_{X|\hat{X}}(x|\hat{x}) \right) \right\}$$

where r is a weight vector (i.e., $r \in \mathcal{M}(\hat{\mathcal{X}})$), and $c(r)$ is a normalizing constant. By definition, $\mathrm{logconv}(\mathcal{Y})$ is log-convex since it contains all log-convex combinations of probability distributions in $\mathcal{Y}$.

Instead of having projection of $p_{X|f(X)}(\cdot|m)$ to $\mathcal{Y}$, we consider the projection to $\mathrm{logconv}(\mathcal{Y})$. Since $\mathrm{logconv}(\mathcal{Y})$ is log-convex, ([17], [Theorem 1]) implies that there exists unique probability distribution $q^\star_m \in \mathrm{logconv}(\mathcal{Y})$ that achieves the following minimum.

$$\min_{q \in \mathrm{logconv}(\mathcal{Y})} D\left(p_{X|f(X)}(\cdot|m) \| q \right).$$

Projection $q^\star_m$ is called an *rI*-projection of $p_{X|f(X)}(\cdot|m)$ to $\mathrm{logconv}(\mathcal{Y})$. Let $r^\star_m$ be the corresponding weights, i.e.,

$$q^\star_m = q^{(r^\star_m)}.$$

Csiszár and Matúš ([17], [Theorem 1]) showed that the *rI*-projection satisfies the following inequality for all $\hat{x} \in \hat{\mathcal{X}}$.

$$D(p_{X|f(X)}(\cdot|m) \| p^\star_{X|\hat{X}}(\cdot|\hat{x})) \geq D(p_{X|f(X)}(\cdot|m) \| q^\star_m) + D(q^\star_m \| p^\star_{X|\hat{X}}(\cdot|\hat{x})). \tag{15}$$

On the other hand, the log-convex combination of probability measures $q^{(r)}$ is called the geometric mean of probability measures [18]. The author also provided geometric compensation identity, which is given by

$$\sum_{\hat{x}} r(\hat{x}) D(p_{X|f(X)}(\cdot|m) \| p^{\star}_{X|\hat{X}}(\cdot|\hat{x})) = D(p_{X|f(X)}(\cdot|m) \| q^{(r)}) + \sum_{\hat{x}} r(\hat{x}) D(q^{(r)} \| p^{\star}_{X|\hat{X}}(\cdot|\hat{x})). \tag{16}$$

The above result holds for any $r \in \mathcal{M}(\hat{\mathcal{X}})$; therefore, Equation (16) also holds when $q^{(r)} = q^{*}_{m}$. Together with Inequality (15), we get the following result. For all $\hat{x} \in \hat{\mathcal{X}}$, if $r^{\star}_{m}(\hat{x}) \neq 0$, then

$$D(p_{X|f(X)}(\cdot|m) \| p^{\star}_{X|\hat{X}}(\cdot|\hat{x})) = D(p_{X|f(X)}(\cdot|m) \| q^{\star}_{m}) + D(q^{\star}_{m} \| p^{\star}_{X|\hat{X}}(\cdot|\hat{x})).$$

Remark 5. *The above result is similar to the projection to polytope in Euclidean space. Suppose vectors $v_1, v_2, \ldots, v_n$ form a polytope, and consider the projection from a vector w to the polytope. Let h be a projection. Then, h is a convex combination of v_i's. Thus, there exist coefficients $\{a_i\}_{1 \leq i \leq n}$, such that*

$$h = \sum_{i=1}^{n} a_i v_i$$

where $\sum_i a_i = 1$, and $a_i \geq 0$ for all i. Let $E = \{1 \leq i \leq n : a_i \neq 0\}$ be the set of indices of nonzero coefficients. Then, projection h is on the plane generated by $\{v_i\}_{i \in E}$. Thus, two vectors $w - h$ and $h - v_i$ are orthogonal for all $i \in E$. Then, Pythagorean theorem implies that, for all i, we have either $a_i = 0$ or

$$\|w - v_i\|^2 = \|w - h\|^2 + \|h - v_i\|^2.$$

5.2. Optimization

As we saw in the previous section, we want to find $q \in \text{logconv}(\mathcal{Y})$ that minimizes $D\left(p_{X|f(X)}(\cdot|m) \| q\right)$. Note that

$$\begin{aligned} D\left(p_{X|f(X)}(\cdot|m) \| q^{(r)}\right) &= \sum_{x \in \mathcal{X}} p_{X|f(X)}(x|m) \log \frac{p_{X|f(X)}(x|m)}{q^{(r)}(x)} \\ &= \sum_{x \in \mathcal{X}} p_{X|f(X)}(x|m) \log p_{X|f(X)}(x|m) + \sum_{x \in \mathcal{X}} p_{X|f(X)}(x|m) \log \frac{1}{q^{(r)}(x)}. \end{aligned}$$

Since the first term is not a function of $q^{(r)}$, it is enough to consider the second term. By the definition of $q^{(r)}$, we have

$$\begin{aligned} \sum_{x \in \mathcal{X}} p_{X|f(X)}(x|m) \log \frac{1}{q^{(r)}(x)} &= -\sum_{x \in \mathcal{X}} p_{X|f(X)}(x|m) \sum_{\hat{x}} r(\hat{x}) \log p^{\star}_{X|\hat{X}}(x|\hat{x}) + \log c(r) \\ &= -\sum_{x \in \mathcal{X}} p_{X|f(X)}(x|m) \sum_{\hat{x}} r(\hat{x}) \log p^{\star}_{X|\hat{X}}(x|\hat{x}) \\ &\quad + \log \left(\sum_{x'} \exp \left(\sum_{\hat{x}} r(\hat{x}) \log p^{\star}_{X|\hat{X}}(x'|\hat{x}) \right) \right). \end{aligned}$$

Thus, minimizing $D(p_{X|f(X)}(\cdot|m) \| q)$ is equivalent to solving the following optimization problem.

$$\begin{aligned} \min_{r \in \mathcal{M}(\hat{\mathcal{X}})} \quad & -\sum_{\hat{x}} r(\hat{x}) \sum_{x} p_{X|f(X)}(x|m) \log p^{\star}_{X|\hat{X}}(x|\hat{x}) + \log \left(\sum_{x'} \exp \left(\sum_{\hat{x}} r(\hat{x}) \log p^{\star}_{X|\hat{X}}(x'|\hat{x}) \right) \right) \\ \text{s.t.} \quad & r(\hat{x}) \geq 0 \\ & \sum_{\hat{x}} r(\hat{x}) = 1. \end{aligned}$$

Since the objective function is a convex function of $r(\hat{x})$, the above problem is a convex optimization problem that can be efficiently solved.

5.3. Relaxation in Clustering

In the corresponding lossy-compression problem under logarithmic loss, reconstruction symbols are probability measures that have a natural algebraic structure, as we discussed in Section 3.3.3. In this section, we present the benefits of such a property when we apply some known techniques from the clustering literature.

Lossy compression is closely related to the clustering problem [19–21]. Many works focused on the application of k-means clustering to a lossy-compression problem [22–24], which is an extension of the Lloyd max algorithm [25,26]. However, k-means clustering is only available when there exists a well-defined operation in $\hat{\mathcal{X}}$ (e.g., $\hat{\mathcal{X}} = \mathbb{R}^n$). This is because k-means clustering requires computing the mean of data points, which is the center of each cluster. In general lossy-compression problems, reconstruction alphabet $\hat{\mathcal{X}}$ may not have such an operation. In such cases, we may have to apply k-medoidlike clustering [27], where the center of each cluster has to be a data point. The k-medoidlike algorithm in the context of lossy compression is shown in Algorithm 1.

Algorithm 1 k-medoidlike clustering in lossy compression.

Randomly initialize $\hat{x}_1, \dots, \hat{x}_M \in \hat{\mathcal{X}}$
repeat
  Set $A_m \leftarrow \emptyset$ for all $1 \le m \le M$.
  for $x \in \mathcal{X}$ **do**
    $A_m \leftarrow A_m \cup \{x\}$ where $m = \operatorname{argmin}_{m'} d(x, \hat{x}_{m'})$
  end for
  for $m = 1$ to M **do**
    $\hat{x}_m \leftarrow \operatorname{argmin}_{\hat{x} \in \hat{\mathcal{X}}} \sum_{x \in A_m} p_X(x) d(x, \hat{x})$
  end for
until converge

On the other hand, in the corresponding problem, the reconstruction alphabet is the set of probability distributions where operations such as log-convex combinations are well-defined. This allows us to propose a k-meanslike clustering algorithm, as shown in Algorithm 2.

Algorithm 2 k-meanslike clustering in lossy compression.

Randomly initialize $r_1, \dots, r_M \in \mathcal{M}(\hat{\mathcal{X}})$
repeat
  Set $A_m \leftarrow \emptyset$ for all $1 \le m \le M$.
  for $x \in \mathcal{X}$ **do**
    $A_m \leftarrow A_m \cup \{x\}$ where $m = \operatorname{argmin}_{m'} \log \frac{1}{q^{(r_{m'})}(x)}$
    Set $f(x) = m$ where $x \in A_m$
  end for
  for $m = 1$ to M **do**
    $r_m \leftarrow \operatorname{argmin}_{r \in \mathcal{M}(\hat{\mathcal{X}})} D(p_{X|f(X)}(\cdot|m) \| q^{(r)})$
  end for
until converge

The main idea of the above algorithm is that log-convex combination $q_m^\star$ behaves like center of cluster A_m. In the clustering literature, there are many known variations of k-means clustering [28,29]. The above result shows that we can borrow those techniques and apply them to the lossy-compression problem even without any algebraic structures on the reconstruction alphabet.

5.4. Application to General Clustering Problems

The idea of the previous section can be applied to an actual clustering problem. We mainly focus on clustering categorical data where data points are not in continuous space [30–34]. Since operations such as *mean* are not well-defined in this case, it is hard to apply known data-clustering algorithms in continuous space. The key idea is that the equivalence relation with logarithmic loss allows the algebraic structure on any set. More precisely, we can transform any clustering problem to the clustering problem in continuous space and apply known techniques such as variations of k-means.

A more rigorous definition of the problem is given below. Assume that we have a finite set of data points $\mathcal{X}$, and each data point has its weight $p_X(x)$. We normalize the weights so that $\sum_x p_X(x) = 1$, and the weights may or may not be uniform. The distance between two points are given by measure $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$. Suppose we want to partition the data points into M clusters.

If we let $\hat{\mathcal{X}} = \mathcal{X}$, then the clustering problem turns out to be a lossy-compression problem under distortion measure $d(\cdot, \cdot)$, where the number of messages is M. Let $D = D^\star(M)$ be the optimal achievable distortion, and $p^\star_{\hat{X}|X}$ be the distribution that achieves rate-distortion function $R(D)$ as defined in Equation (4). Then, we can find the corresponding lossy-compression problem under logarithmic loss. Finally, we can apply clustering algorithms in continuous space such as k-means to the corresponding problem. For example, Algorithm 2 can be applied to the corresponding problem.

Remark 6. *Note that it is hard to have an exact analytic formula for $D^\star(M)$ or $p^\star_{\hat{X}|X}$. However, as we mentioned in Section 3.3.2, we do not have to find an optimal scheme under the exact problem formulation. If we can provide a good scheme of the corresponding problem with $D \approx D^\star(M)$, that should be a good enough scheme in the original problem.*

6. Conclusions

To conclude our discussion, we summarize our main contributions. We showed that for any fixed-length lossy-compression problem under an arbitrary distortion measure, there exists a corresponding lossy-compression problem under logarithmic loss where optimal schemes coincide. We also proved that a good scheme for one lossy-compression problem is also good for another problem. This equivalence provides an algebraic structure on any reconstruction alphabet that allows using various optimization techniques in lossy-compression problems, such as log-convex relaxation. Furthermore, our results naturally suggest a k-meanslike clustering algorithm in categorical data-clustering problems.

Funding: This work was supported by the National Research Foundation of Korea, funded by the Korean Government (MSIT) under Grant NRF-2017R1C1B5018298.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Optimal Scheme Under Excess Distortion Criterion

In this section, we characterize minimum number of codewords $M^\star(D, \epsilon)$ that can achieve distortion D and excess distortion probability ϵ. Let an encoder and a decoder be $f : \mathcal{X} \to \{1, \dots, M\}$ and $g : \{1, \dots, M\} \to \mathcal{M}(\mathcal{X})$ where $g(m) = q^{(m)} \in \mathcal{M}(\mathcal{X})$. Since $\ell(x, q) \le D$ is equivalent to $q(x) \ge e^{-D}$, we hav

$$\begin{aligned} 1 - p_e &= \sum_{x \in \mathcal{X}} p_X(x) \mathbf{1}\left(q^{(f(x))}(x) \ge e^{-D}\right) \\ &= \sum_{m=1}^{M} \sum_{x \in f^{-1}(m)} p_X(x) \mathbf{1}\left(q^{(m)}(x) \ge e^{-D}\right). \end{aligned}$$

However, at most, $\lfloor e^D \rfloor$ of the $q^{(m)}(x)$ can be larger than e^{-D} where $\lfloor x \rfloor$ is the largest integer that is smaller than or equal to x. Thus, we can at mos cover t $M \cdot \lfloor e^D \rfloor$ of the source symbols with M codewords. Suppose $p_X(1) \geq p_X(2) \geq \cdots \geq p_X(r)$, then the optimal scheme is

$$f(x) = \left\lceil \frac{x}{\lfloor e^D \rfloor} \right\rceil$$
$$q^{(m)}(x) = \begin{cases} 1/\lfloor e^D \rfloor & \text{if } f(x) = m \\ 0 & \text{otherwise,} \end{cases}$$

where $q^{(m)} = g(m)$ and $\lceil x \rceil$ are the smallest integer that is larger than or equal to x. The idea is that each reconstruction symbol $q^{(m)}$ covers $\lfloor e^D \rfloor$ number of source symbols by assigning probability mass $1/\lfloor e^D \rfloor$ to each of them.

The above optimal scheme satisfies

$$\begin{aligned} 1 - p_e = & \sum_{x=1}^{M \cdot \lfloor e^D \rfloor} p_X(x) \\ = & F_X\left(M \cdot \lfloor e^D \rfloor\right), \end{aligned}$$

where $F_X(\cdot)$ is the cumulative distribution function of X. This implies that the minimal error probability is

$$\epsilon^\star(M, D) = 1 - F_X\left(M \cdot \lfloor e^D \rfloor\right).$$

On the other hand, if we fix target error probability ϵ, the minimal number of codewords is

$$M^\star(D, \epsilon) = \left\lceil \frac{F_X^{-1}(1-\epsilon)}{\lfloor e^D \rfloor} \right\rceil$$

where $F_X^{-1}(y) = \underset{1 \leq x \leq r}{\operatorname{argmin}} \{x : F_X(x) \geq y\}$. Note that if we allow variable length coding without a prefix condition, the optimal coding scheme is similar to optimal nonasymptotic lossless coding introduced in [35].

References

1. Courtade, T.A.; Wesel, R.D. Multiterminal source coding with an entropy-based distortion measure. *Proc. IEEE Int. Symp. Inf. Theory. IEEE* **2011**, *2011*, 2040–2044.
2. Courtade, T.; Weissman, T. Multiterminal Source Coding Under Logarithmic Loss. *IEEE Trans. Inf. Theory* **2014**, *60*, 740–761. [CrossRef]
3. Ugur, Y.; Aguerri, I.E.; Zaidi, A. Vector Gaussian CEO problem under logarithmic loss. In Proceedings of the 2018 IEEE Information Theory Workshop, Guangzhou, China, 25–29 November 2018; pp. 1–5.
4. Shkel, Y.Y.; Verdú, S. A single-shot approach to lossy source coding under logarithmic loss. *IEEE Trans. Inf. Theory* **2018**, *64*, 129–147. [CrossRef]
5. Jiao, J.; Courtade, T.A.; Venkat, K.; Weissman, T. Justification of logarithmic loss via the benefit of side information. *IEEE Trans. Inf. Theory* **2015**, *61*, 5357–5365. [CrossRef]
6. Painsky, A.; Wornell, G.W. Bregman divergence bounds and the universality of the logarithmic loss. *arXiv* **2018**, arXiv:1810.07014.
7. No, A. Universality of Logarithmic Loss in Successive Refinement. *Entropy* **2019**, *21*, 158. [CrossRef]
8. Tishby, N. Pereira, F.; Bialek, W. The information bottleneck method. In Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 22–24 September 1999; pp. 368–377.

9. Harremoës, P.; Tishby, N. The information bottleneck revisited or how to choose a good distortion measure. In Proceedings of the 2007 IEEE International Symposium on Information Theory, Nice, France, 24–29 June 2007; pp. 566–570.
10. Gilad-Bachrach, R.; Navot, A.; Tishby, N. An information theoretic tradeoff between complexity and accuracy. In *Learning Theory and Kernel Machines*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 595–609.
11. Aguerri, I.E.; Zaidi, A. Distributed Information Bottleneck Method for Discrete and Gaussian Sources. In Proceedings of the International Zurich Seminar on Information and Communication, Zurich, Switzerland, 21–23 February 2018.
12. Kostina, V.; Verdú, S. Nonasymptotic noisy lossy source coding. *IEEE Trans. Inf. Theory* **2016**, *62*, 6111–6123. [CrossRef]
13. Kostina, V.; Verdú, S. Fixed-length lossy compression in the finite blocklength regime. *IEEE Trans. Inf. Theory* **2012**, *58*, 3309–3338. [CrossRef]
14. Csiszár, I. On an extremum problem of information theory. *Studia Scientiarum Mathematicarum Hungarica* **1974**, *9*, 57–71.
15. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2012.
16. No, A.; Weissman, T. Universality of logarithmic loss in lossy compression. In Proceedings of the 2015 IEEE International Symposium on Information Theory, Hongkong, China, 14–19 June 2015; pp. 2166–2170.
17. Csiszár, I.; Matus, F. Information projections revisited. *IEEE Trans. Inf. Theory* **2003**, *49*, 1474–1490. [CrossRef]
18. No, A. Information Geometric Approach on Most Informative Boolean Function Conjecture. *Entropy* **2018**, *20*, 688. [CrossRef]
19. Chaffee, D.L. Applications of Rate Distortion Theory to the Bandwidth Compression of Speech Signals. Ph.D. Thesis, University of California, Los Angeles, CA, USA, 1975.
20. Chen, D. On two or more dimensional optimum quantizers. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Hartford, CT, USA, 9–11 May 1977; Volume 2, pp. 640–643.
21. Gray, R.; Buzo, A.; Matsuyoma, Y.; Gray, A., Jr.; Markel, J. Source coding and speech compression. In *International Telemetering Conference Proceedings*; International Foundation for Telemetering: San Diego, CA, USA, 1978; Volume 14.
22. Linde, Y.; Buzo, A.; Gray, R. An algorithm for vector quantizer design. *IEEE Trans. Commun.* **1980**, *28*, 84–95. [CrossRef]
23. Gray, R.M.; Neuhoff, D.L. Quantization. *IEEE Trans. Inf. Theory* **1998**, *44*, 2325–2383. [CrossRef]
24. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
25. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [CrossRef]
26. Max, J. Quantizing for minimum distortion. *IRE Trans. Inf. Theory* **1960**, *6*, 7–12. [CrossRef]
27. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2009; Volume 344.
28. Phillips, S.J. Acceleration of k-means and related clustering algorithms. In Proceedings of the Workshop on Algorithm Engineering and Experimentation, San Francisco, CA, USA, 4–5 January 2002; pp. 166–177.
29. Pelleg, D.; Moore, A.W. X-means: Extending k-means with efficient estimation of the number of clusters. In Proceedings of the Seventeenth International Conference on Machine Learning, Stanford, CA, USA, 29 June–2 July 2000; Volume 1, pp. 727–734.
30. Watve, A.; Pramanik, S.; Jung, S.; Jo, B.; Kumar, S.; Sural, S. Clustering Non-Ordered Discrete Data. *J. Inf. Sci. Eng.* **2014**, *30*, 1–23.
31. Bai, L.; Liang, J.; Dang, C.; Cao, F. A novel attribute weighting algorithm for clustering high-dimensional categorical data. *Pattern Recognit.* **2011**, *44*, 2843–2861. [CrossRef]
32. Ng, R.T.; Han, J. CLARANS: A method for clustering objects for spatial data mining. *IEEE Trans. Knowl. Data Eng.* **2002**, *14*, 1003–1016. [CrossRef]
33. Ganti, V.; Gehrke, J.; Ramakrishnan, R. CACTUS—clustering categorical data using summaries. In Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 15–18 August 1999; Volume 99, pp. 73–83.

34. Kumar, S.; Sural, S.; Watve, A.; Pramanik, S. CNODE: clustering of set-valued non-ordered discrete data. *Int. J. Data Min. Model. Manag.* **2009**, *1*, 310–334. [CrossRef]
35. Kontoyiannis, I.; Verdu, S. Optimal Lossless Data Compression: Non-Asymptotics and Asymptotics. *IEEE Trans. Inf. Theory* **2014**, *60*, 777–795. [CrossRef]