*entropy*

MDPI

*Article*

# Universality of Logarithmic Loss in Successive Refinement [†]

**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: 19 December 2018; Accepted: 7 February 2019; Published: 8 February 2019

**Abstract:** We establish an universal property of logarithmic loss in the successive refinement problem. If the first decoder operates under logarithmic loss, we show that any discrete memoryless source is successively refinable under an arbitrary distortion criterion for the second decoder. Based on this result, we propose a low-complexity lossy compression algorithm for any discrete memoryless source.

**Keywords:** logarithmic loss; rate-distortion; successive refinability

## 1. Introduction

In the lossy compression problem, logarithmic loss is a criterion allowing a "soft" reconstruction of the source, a departure from the classical setting of a deterministic reconstruction. In this setting, the reconstruction alphabet is the set of probability distributions over the source alphabet. More precisely, let $x$ be the source symbol from the source alphabet $\mathcal{X}$, and $q(\cdot)$ be the reconstruction symbol which is the probability measure on $\mathcal{X}$. Then the logarithmic loss is given by

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

Clearly, if the reconstruction $q(\cdot)$ has a small probability on the true source symbol $x$, the amount of loss will be large.

Although logarithmic loss plays a crucial role in the theory of learning and prediction, relatively little work has been done in the context of lossy compression, notwithstanding the two-encoder multi-terminal source coding problem under logarithmic loss [1,2], or recent work on the single-shot approach to lossy source coding under logarithmic loss [3]. Note that lossy compression under logarithmic loss is closely related to the information bottleneck method [4–6]. In this paper, we focus on universal properties of logarithmic loss in the context of successive refinement.

Successive refinement is a network lossy compression problem where one encoder wishes to describe the source to two decoders [7,8]. Instead of having two separate coding schemes, the successive refinement encoder designs a code for the decoder with a weaker link, and sends extra information to the second decoder on top of the message of the first decoder. In general, successive refinement coding cannot do as well as two separate encoding schemes optimized for the respective decoders. However, if we can achieve the point-to-point optimum rates using successive refinement coding, we say the source is successively refinable.

Although necessary and sufficient conditions of successive refinability is known [7,8], proving (or disproving) successive refinability of the source is not a simple task. Equitz and Cover [7] found a discrete source that is not successively refinable using Gerrish problem [9]. Chow and Berger found a continuous source that is not successively refinable using Gaussian mixture [10]. Lastras and Berger showed that all sources are nearly successively refinable [11]. However, still only a few sources are known to be successively refinable. In this paper, we show that any discrete memoryless source is successively refinable as long as the weaker link employs logarithmic loss, regardless of the distortion criterion used for the stronger link.

In the second part of the paper, we show that this result can be useful to design a lossy compression algorithm with low complexity. Recently, the idea of successive refinement is applied to reduce the complexity of point-to-point lossy compression algorithm. Venkataramanan et al. proposed a new lossy compression for Gaussian source where the codewords are linear combination of sub-codewords [12]. No and Weissman also proposed a low-complexity lossy compression algorithm for Gaussian source using extreme value theory [13]. Both algorithms are successively describing source and achieve low complexity. Roughly speaking, successive refinement algorithm provides a smaller size of codebook. For example, the naive random coding scheme has a codebook of size $e^{nR}$ when the blocklength is $n$ and the rate is $R$. On the other hand, if we can design a successive refinement scheme with half rate in the weaker link, then the size of codebook is $e^{nR/2}$ each. Thus, the overall codebook size is $2e^{nR/2}$. The above idea can be generalized to successive refinement scheme with $L$ decoders [12,14]

The universal property of logarithmic loss in successive refinement implies that, for any point-to-point lossy compression of discrete memoryless source, we can insert a virtual intermediate decoder (weaker link) under logarithmic loss without losing any rates at the actual decoder (stronger link). As we discussed, this property allows us to design a lossy compression algorithm with low-complexity for any discrete source and distortion pair. Note that previous works only focused on specific source and distortion pair such as binary source with Hamming distortion.

The remainder of the paper is organized as follows. In Section 2, we revisit some of the known results pertaining to logarithmic loss. Section 3 is dedicated to successive refinement under logarithmic loss in the weaker link. In Section 4, we propose a low complexity compression scheme that can be applied to any discrete lossy compression problem. Finally, we conclude in Section 5.

*Notation*: $X^n$ denotes an $n$-dimensional random vector $(X_1, X_2, \ldots, X_n)$ while $x^n$ denotes a specific possible realization of the random vector $X^n$. $\mathcal{X}$ denotes a support of random variable $X$. Also, $Q$ denotes a random probability mass function while $q$ denotes a specific probability mass function. We use natural logarithm and nats instead of bits.

## 2. Preliminaries

### 2.1. Successive Refinability

In this section, we review the successive refinement problem with two decoders. Let the source $X^n$ be i.i.d. random vector with distribution $p_X$. The encoder wants to describe $X^n$ to two decoders by sending a pair of messages $(m_1, m_2)$ where $1 \leq m_i \leq M_i$ for $i \in \{1,2\}$. The first decoder reconstructs $\hat{X}_1^n(m_1) \in \hat{\mathcal{X}}_1^n$ based only on the first message $m_1$. The second decoder reconstructs $\hat{X}_2^n(m_1, m_2) \in \hat{\mathcal{X}}_2^n$ based on both $m_1$ and $m_2$. The setting is described in Figure 1.

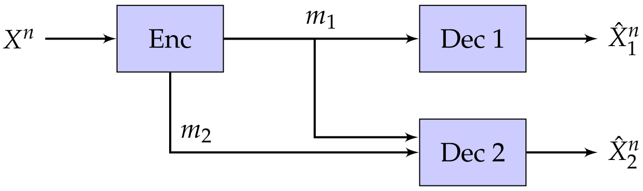

**Figure 1.** Successive Refinement.

Let $d_i(\cdot,\cdot) : \mathcal{X} \times \hat{\mathcal{X}}_i \to [0,\infty)$ be a distortion measure for $i$-th decoder. The rates of code $(R_1, R_2)$ are simply defined as

$$\begin{aligned} R_1 &= \frac{1}{n} \log M_1 \\ R_2 &= \frac{1}{n} \log M_1 M_2. \end{aligned}$$

An $(n, R_1, R_2, D_1, D_2, \epsilon)$-successive refinement code is a coding scheme with block length $n$ and excess distortion probability $\epsilon$ where rates are $(R_1, R_2)$ and target distortions are $(D_1, D_2)$. Since we have two decoders, the excess distortion probability is defined by $\Pr\left[d_i(X^n, \hat{X}_i^n) > D_i \text{ for some } i\right]$.

**Definition 1.** *A rate-distortion tuple $(R_1, R_2, D_1, D_2)$ is said to be achievable if there is a family of $(n, R_1^{(n)}, R_2^{(n)}, D_1, D_2, \epsilon^{(n)})$-successive refinement code where*

$$\begin{aligned} &\lim_{n\to\infty} R_i^{(n)} = R_i \textit{ for all } i, \\ &\lim_{n\to\infty} \epsilon^{(n)} = 0. \end{aligned}$$

For some special cases, both decoders can achieve the point-to-point optimum rates simultaneously.

**Definition 2.** *Let $R_i(D_i)$ denote the rate-distortion function of the i-th decoder for $i \in \{1, 2\}$. If the rate-distortion tuple $(R_1(D_1), R_2(D_2), D_1, D_2)$ is achievable, then we say the source is successively refinable at $(D_1, D_2)$. If the source is successively refinable at $(D_1, D_2)$ for all $D_1, D_2$, then we say the source is successively refinable.*

The following theorem provides a necessary and sufficient condition of successive refinable sources.

**Theorem 1** ([7,8])**.** *A source is successively refinable at $(D_1, D_2)$ if and only if there exists a conditional distribution $p_{\hat{X}_1, \hat{X}_2|X}$ such that $X - \hat{X}_2 - \hat{X}_1$ forms a Markov chain and*

$$\begin{aligned} R_i(D_i) &= I(X; \hat{X}_i) \\ \mathbb{E}\left[d_i(X, \hat{X}_i)\right] &\le D_i \end{aligned}$$

*for $i \in \{1, 2\}$.*

Note that the above results of successive refinability can easily be generalized to the case of $k$ decoders.

### 2.2. Logarithmic Loss

Let $\mathcal{X}$ be a set of discrete source symbols ($|\mathcal{X}| < \infty$), and $\mathcal{M}(\mathcal{X})$ be the set of probability measures on $\mathcal{X}$. Logarithmic loss $\ell : \mathcal{X} \times \mathcal{M}(\mathcal{X}) \to [0, \infty]$ is defined by

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

for $x \in \mathcal{X}$ and $q \in \mathcal{M}(\mathcal{X})$. Logarithmic loss between $n$-tuples is defined by

$$\ell_n(x^n, q^n) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{1}{q_i(x_i)},$$

i.e., the symbol-by-symbol extension of the single letter loss.

Let $X^n$ be the discrete memoryless source with distribution $p_X$. Consider the lossy compression problem under logarithmic loss where the reconstruction alphabet is $\mathcal{M}(\mathcal{X})$. The rate-distortion function is given by

$$\begin{aligned} R(D) = & \inf_{p_{Q|X}:\mathbb{E}[\ell(X,Q)]\le D} I(X;Q) \\ =& H(X) - D. \end{aligned}$$

The following lemma provides a property of the rate-distortion function achieving conditional distribution.

**Lemma 1.** *The rate-distortion function achieving conditional distribution $p_{Q^\star|X}$ satisfies*

$$p_{X|Q^\star}(\cdot|q) = q \tag{1}$$
$$H(X|Q^\star) = D \tag{2}$$

*for $p_{Q^\star}$ almost every $q \in \mathcal{M}(\mathcal{X})$. Conversely, if $p_{Q|X}$ satisfies (1) and (2), then it is a rate-distortion function achieving conditional distribution, i.e.,*

$$\begin{aligned} I(X;Q) =& R(D) = H(X) - D \\ \mathbb{E}\left[\ell(X,Q)\right] =& D. \end{aligned}$$

The key idea is that we can replace $Q$ by $p_{X|Q}(\cdot|Q)$, and have lower rate and distortion, i.e.,

$$\begin{aligned} I(X;Q) \ge& I(X; p_{X|Q}(\cdot|Q)) \\ \mathbb{E}\left[\ell(X,Q)\right] \ge& \mathbb{E}\left[\ell(X, p_{X|Q}(\cdot|Q)\right], \end{aligned}$$

which directly implies (1).

Interestingly, since the rate-distortion function in this case is a straight line, a simple time sharing scheme achieves the optimal rate-distortion trade-off. More precisely, the encoder losslessly compresses only the first $\frac{H(X)-D}{H(X)}$ fraction of the source sequence components. Then, the decoder perfectly recovers those losslessly compressed components and uses $p_X$ as its reconstruction for the remaining part. The resulting scheme obviously achieves distortion $D$ with rate $H(X) - D$.

Furthermore, this simple scheme directly implies successive refinability of the source. For $D_1 > D_2$, suppose the encoder losslessly compresses the first $\frac{H(X)-D_2}{H(X)}$ fraction of the source. Then, the first decoder can perfectly reconstruct $\frac{H(X)-D_1}{H(X)}$ fraction of the source with the message of rate $H(X) - D_1$ and distortion $D_1$ while the second decoder can achieve distortion $D_2$ with rate $H(X) - D_2$. Since both decoders can achieve the best rate-distortion pair, it follows that any discrete memoryless source under logarithmic loss is successively refinable.

We can formally prove successive refinability of discrete memoryless source under logarithmic loss using Theorem 1. I.e., by finding random probability mass functions $Q_1, Q_2 \in \mathcal{M}(\mathcal{X})$ that satisfy

$$I(X;Q_1) = H(X) - D_1, \quad \mathbb{E}\left[\ell(X,Q_1)\right] = D_1, \tag{3}$$
$$I(X;Q_2) = H(X) - D_2, \quad \mathbb{E}\left[\ell(X,Q_2)\right] = D_2, \tag{4}$$

where $X - Q_2 - Q_1$ forms a Markov chain.

Let $e_x$ be a deterministic probability mass function (pmf) in $\mathcal{M}(\mathcal{X})$ that has a unit mass at $x$. In other words,

$$e_x(\tilde{x}) = \begin{cases} 1 & \text{if } \tilde{x} = x \\ 0 & \text{otherwise.} \end{cases}$$

Then, consider random pmfs $Q_1, Q_2 \in \{e_x : x \in \mathcal{X}\} \cup \{p_X\}$. Since the support of $Q_1$ and $Q_2$ is finite, we can define the following conditional pmfs.

$$p_{Q_2|X}(q_2|x) = \begin{cases} \frac{H(X)-D_2}{H(X)} & \text{if } q_2 = e_x \\ \frac{D_2}{H(X)} & \text{if } q_2 = p_X \\ 0 & \text{otherwise} \end{cases}$$

$$p_{Q_1|Q_2}(q_1|q_2) = \begin{cases} \frac{H(X)-D_1}{H(X)-D_2} & \text{if } q_1 = q_2 = e_x \text{ for some } x \\ \frac{D_1-D_2}{H(X)-D_2} & \text{if } q_1 = p_X \text{ and } q_2 = e_x \text{ for some } x \\ 1 & \text{if } q_1 = q_2 = p_X \\ 0 & \text{otherwise.} \end{cases}$$

It is not hard to show that the above conditional pmfs satisfies (3) and (4).

## 3. Successive Refinability

*Main Results*

Consider the successive refinement problem with a discrete memoryless source as described in Section 2.1. Specifically, we are interested in the case where the first decoder is under logarithmic loss and the second decoder is under some arbitrary distortion measure $d(\cdot, \cdot)$. We only have a following benign assumption that if $d(x, \hat{x}_1) = d(x, \hat{x}_2)$ for all $x$, then $\hat{x}_1 = \hat{x}_2$. This is not a hard restriction since if $\hat{x}_1$ and $\hat{x}_2$ have the same distortion values for all $x$, then there is no reason to have both reconstruction symbols.

The following theorem shows that any discrete memoryless source is successive refinable as long as the weaker link is under logarithmic loss. This implies an universal property of logarithmic loss in the context of successive refinement.

**Theorem 2.** *Let the source be arbitrary discrete memoryless. Suppose the distortion criterion of the first decoder is logarithmic loss while that of the second decoder is an arbitrary distortion criterion $d : \mathcal{X} \times \hat{\mathcal{X}} \rightarrow [0, \infty]$. Then the source is successively refinable.*

**Proof.** The source is successively refinable at $(D_1, D_2)$ if and only if there exists a $X - \hat{X} - Q$ such that

$$\begin{aligned} I(X;Q) =& R_1(D_1), \quad \mathbb{E}\left[\ell(X,Q)\right] \leq D_1 \\ I(X;\hat{X}) =& R_2(D_2), \quad \mathbb{E}\left[d(X,\hat{X})\right] \leq D_2. \end{aligned}$$

Let $p_{\hat{X}^\star|X}$ be the conditional distribution for the second decoder that achieves the informational rate-distortion function $R_2(D_2)$. i.e.,

$$I(X;\hat{X}^\star) = R_2(D_2), \quad \mathbb{E}\left[d(X,\hat{X}^\star)\right] = D_2.$$

Since the weaker link is under logarithmic loss, we have $R_1(D_1) \leq R_2(D_2)$. This implies that $H(X) - D_1 \leq H(X) - H(X|\hat{X}^\star)$. Thus, we can assume $H(X|\hat{X}^\star) \leq D_1$ throughout the proof. For simplicity, we further have a benign assumption that there is no $\hat{x}$ such that $p_X(x) = p_{X|\hat{X}^\star}(x|\hat{x})$ for all $x$. (See Remark 1 for the case where such $\hat{x}$ exists.)

Without loss of generality, suppose $\hat{\mathcal{X}} = \{0, 1, \ldots, s-1\}$. Consider a random variable $Y \in \mathcal{Y} = \{0, 1, \ldots, s\}$ with the following pmf for some $0 \le \epsilon \le 1$:

$$p_Y(y) = \begin{cases} (1-\epsilon) p_{\hat{X}^\star}(y) & \text{if } y \le s-1 \\ \epsilon & \text{if } y = s. \end{cases}$$

The conditional distribution is given by

$$p_{\hat{X}^\star|Y}(\hat{x}|y) = \begin{cases} 1 & \text{if } \hat{x} = y \le s-1 \\ 0 & \text{if } \hat{x} \neq y \le s-1 \\ p_{\hat{X}^\star}(\hat{x}) & \text{if } y = s. \end{cases}$$

The joint distribution of $X, \hat{X}^\star, Y$ is given by

$$p_{X,\hat{X}^\star,Y}(x, \hat{x}, y) = p_{X,\hat{X}^\star}(x, \hat{x}) p_{Y|\hat{X}^\star}(y|\hat{x}).$$

It is clear that $H(X|Y) = H(X|\hat{X}^\star)$ if $\epsilon = 0$ and $H(X|Y) = H(X)$ if $\epsilon = 1$. Since $H(X|\hat{X}^\star) \le D_1$, there exists an $0 \le \epsilon \le 1$ such that $H(X|Y) = D_1$.

We are now ready to define the Markov chain. Let $Q = p_{X|Y}(\cdot|Y)$ and $q^{(y)} = p_{X|Y}(\cdot|y)$ for all $y \in \mathcal{Y}$. The following lemma implies that there is a one-to-one mapping between $q$ and $y$.

**Lemma 2.** *If $p_{X|Y}(x|y_1) = p_{X|Y}(x|y_2)$ for all $x \in \mathcal{X}$, then $y_1 = y_2$.*

The proof of lemma is given in Appendix A. Since $Q = p_{X|Y}(\cdot|Y)$ is a one-to-one mapping, we have

$$I(X; Q) = I(X; Y) = H(X) - D_1 = R_1(D_1).$$

Also, we have

$$\mathbb{E}\left[\ell(X, Q)\right] = \mathbb{E}\left[\log \frac{1}{p_{X|Y}(X|Y)}\right] = H(X|Y) = D_1.$$

Furthermore, $X - \hat{X}^\star - Q$ forms a Markov chain since $X - \hat{X}^\star - Y$ forms a Markov chain. This concludes the proof. □

The key idea of the theorem is that (1) is the only loose required condition for the rate-distortion function achieving conditional distribution. Thus, for any distortion criterion in the second stage, we are able to choose an appropriate distribution $p_{X,\hat{X},Q}$ that satisfies both (1) and the condition for successive refinability.

**Remark 1.** *The assumption $p_{X|\hat{X}^\star}(\cdot|\hat{x}) \neq p_X(\cdot)$ for all $\hat{x}$ is not necessary. Appendix B shows another joint distribution $p_{X,\hat{X}^\star,Y}$ that satisfies conditions for successive refinability when the above assumption does not hold.*

*The distribution in the above proof is one simple example that has a single parameter $\epsilon$, but we can always find other distributions that satisfy the condition for successive refinability. In the next section, we propose totally different distribution that achieves a Markov chain $X - \hat{X}^\star - Y$ with $H(X|Y) = D_1$. This implies that the above proof does not rely on the assumption.*

**Remark 2.** *In the proof, we used random variable $Y$ to define $Q = p_{X|Y}(\cdot|Y)$. On the other hand, if the joint distribution $p_{X,\hat{X}^\star,Q}$ satisfies conditions of successive refinability, there exists a random variable $Y$ where*

*$X - \hat{X}^\star - Y$ forms a Markov chain and $Q = p_{X|Y}(\cdot|Y)$. This is simply because we can set $Y = Q$, which implies $p_{X|Y}(\cdot|Y) = p_{X|Q}(\cdot|Q) = Q$.*

Theorem 2 can be generalized to successive refinement problem with $K$ intermediate decoders. Consider random variables $Y_k \in \mathcal{Y}$ for $1 \leq k \leq K$ such that $X - \hat{X}^\star - Y_K - \cdots - Y_1$ forms a Markov chain and the joint distribution of $X, \hat{X}^\star, Y_1, \ldots, Y_K$ is given by

$$\begin{aligned} &p_{X,\hat{X}^\star,Y_1,\ldots,Y_K}(x, \hat{x}, y_1, \ldots, y_K) \\ &= p_{X,\hat{X}^\star}(x,\hat{x}) p_{Y_1|\hat{X}^\star}(y_1|\hat{x}) \prod_{k=1}^{K-1} p_{Y_{k+1}|Y_k}(y_{k+1}|y_k) \end{aligned}$$

where $H(X|Y_k) = D_k$. Similar to the proof of Theorem 2, we can show that $Q_k = p_{X|Y_k}(\cdot|Y_k)$ for all $1 \leq k \leq K$ satisfy the condition for successive refinability (where posterior distributions $p_{X|Y_k}(\cdot|y_k)$ should be distinct for all $y_k \in \mathcal{Y}$ to guarantee one-to-one correspondence). Thus, we can conclude that any discrete memoryless source with $K$ intermediate decoders is successively refinable as long as all the intermediate decoders are under logarithmic loss.

## 4. Toward Lossy Compression with Low Complexity

As we mentioned in Remark 1, the choice of joint distribution $p_{X,\hat{X}^\star,Q}$ in the proof of Theorem 2 is not unique. In this section, we propose another joint distribution $p_{X,\hat{X}^\star,Q}$ that satisfies the conditions for successive refinability. It naturally suggests a new lossy compression algorithm which we will discuss in Section 4.3.

### *4.1. Rate-Distortion Achieving Joint Distribution: Small $D_1$*

Recall that $H(X|\hat{X}^\star) \leq D_1$. We first consider the case where $D_1$ is not too large so that $D_1$ is close to $H(X|\hat{X}^\star)$. We will clarify this later. For simplicity, we further assume that $p_{\hat{X}^\star}(0) \geq \cdots \geq p_{\hat{X}^\star}(s-1)$. Consider a random variable $Z_\epsilon^{(s)} \in \hat{\mathcal{X}}$ with the following pmf for some $0 \leq \epsilon \leq (s-1)\min_{\hat{x}} p_{\hat{X}^\star}(\hat{x})$

$$p_{Z_\epsilon^{(s)}}(z) = \begin{cases} 1-\epsilon & \text{if } z = 0 \\ \frac{\epsilon}{s-1} & \text{if } 1 \leq z \leq s-1. \end{cases}$$

If it is clear from context, we simply use $Z \equiv Z_\epsilon^{(s)}$ for the sake of notation. We further define a random variable $Y$ that is independent to $Z$ such that $\hat{X}^\star = Y \oplus_s Z$, where $\oplus_s$ denotes a sum modulo $s$. This can be achieved by following pmf and conditional pmf.

$$\begin{aligned} p_Y(y) &= \frac{p_{\hat{X}^\star}(y) - \frac{\epsilon}{s-1}}{1 - \frac{s}{s-1}\epsilon} \qquad (5) \\ p_{\hat{X}^\star|Y}(\hat{x}|y) &= \begin{cases} 1-\epsilon & \text{if } \hat{x} = y \\ \frac{\epsilon}{s-1} & \text{if } \hat{x} \neq y. \end{cases} \end{aligned}$$

If $\epsilon = 0$, we have $H(X|Y) = H(X|\hat{X}^\star)$. Also, it is clear that $H(X|Y)$ increases as $\epsilon$ increases. Since we assume that $D_1$ is not too large, there exists $0 \leq \epsilon \leq (s-1)\min p_{\hat{X}^\star}(\hat{x})$ such that $H(X|Y) = D_1$. We will discuss about the case of general $D_1$ in Section 4.2. The joint distribution of $X, \hat{X}^\star, Y$ is given by

$$p_{X,\hat{X}^\star,Y}(x,\hat{x},y) = p_{X,\hat{X}^\star}(x,\hat{x}) p_{Y|\hat{X}^\star}(y|\hat{x}).$$

We are now ready to define the Markov chain. Let $Q = p_{X|Y}(\cdot|Y)$ and $q^{(y)} = p_{X|Y}(\cdot|y)$ for all $y \in \mathcal{Y}$ where $\mathcal{Y} = \hat{\mathcal{X}} = \{0, 1, \ldots s-1\}$. For simplicity, we assume that $p_{X|Y}(\cdot|y_1)$ and $p_{X|Y}(\cdot|y_2)$ are not the same for all $y_1 \neq y_2$. Since $Q = p_{X|Y}(\cdot|Y)$ is a one-to-one mapping, we have

$$I(X;Q) = I(X;Y) = H(X) - D_1 = R_1(D_1).$$

Also, we have

$$\mathbb{E}\left[\ell(X,Q)\right] = \mathbb{E}\left[\log \frac{1}{p_{X|Y}(X|Y)}\right] = H(X|Y) = D_1.$$

Furthermore, $X - \hat{X}^\star - Q$ forms a Markov chain since $X - \hat{X}^\star - Y$ forms a Markov chain. Thus, the above construction of joint distribution $p_{X,\hat{X}^\star,Q}$ satisfies the conditions for successive refinability.

### 4.2. Rate-Distortion Achieving Joint Distribution: General $D_1$

The joint distribution in the previous section only works for small $D_1$. It is because $\epsilon$ has a natural upper-bound from (5) which is $\epsilon \leq (s-1)\min p_{\hat{X}^\star}(\hat{x})$. In this section, we generalize the proof in the previous section for general $D_1$. The key observation is that if we pick the maximum $\epsilon = (s-1)\min p_{\hat{X}^\star}(\hat{x})$, then $p_Y(s-1) = 0$. This implies that we can focus on the smaller set of reconstruction alphabet $\mathcal{Y} = \{0, 1, \ldots s-2\}$.

Let $U_s = \hat{X}^\star$, and define random variables $\{U_k : 1 \leq k \leq s-1\}$ recursively. More precisely, we define the random variable $U_{k-1}$ based on $U_k$ for $2 \leq k \leq s$.

$$U_k = U_{k-1} \oplus_k Z^{(k)}_{\epsilon_k}$$
$$p_{Z^{(k)}_{\epsilon_k}}(z) = \begin{cases} 1-\epsilon_k & \text{if } z = 0 \\ \frac{\epsilon_k}{k-1} & \text{if } 1 \leq z \leq k-1 \end{cases}$$

where

$$\epsilon_k = (k-1)\min_u p_{U_k}(u).$$

Similar to the definition of $Y$, we assume $U_{k-1}$ and $Z^{(k)}_{\epsilon_k}$ are independent, and $\oplus_k$ denotes modulo $k$ sum. Each time step, the alphabet size of $U_k$ decreases by one. Thus, we have $0 \leq U_k \leq k-1$, and therefore $U_1 = 0$ with probability 1. Furthermore, we have

$$H(X|U_s) \leq H(X|U_{s-1}) \leq \cdots \leq H(X|U_1) = H(X).$$

For $H(X|\hat{X}^\star) \leq D_1 < H(x)$, there exists $k$ such that $H(X|U_k) > D_1 \geq H(X|U_{k-1})$. Thus, there exists $Y$ that satisfies $H(X|Y) = D_1$ and $U_k = Y \oplus_k Z^{(k)}_\epsilon$ for some $0 \leq \epsilon \leq \epsilon_k$. This implies that

$$\hat{X}^\star = Z^{(s)}_{\epsilon_s} \oplus_s \left[Z^{(s-1)}_{\epsilon_{s-1}} \oplus_{s-1} \cdots \left(Z^{(k+1)}_{\epsilon_{k+1}} \oplus_{k+1} \left(Z^{(k)}_{\epsilon} \oplus_k Y\right)\right)\right].$$

Similar to the previous section, we assume that $p_{X|Y}(\cdot|y_1) \neq p_{X|Y}(\cdot|y_2)$ if $y_1 \neq y_2$. Then, we can set $Q = p_{X|Y}(\cdot|Y)$ which satisfies the conditions for successive refinability.

### 4.3. Iterative Lossy Compression Algorithm

The joint distribution from the previous section naturally suggests a simple successive refinement scheme. Consider the lossy compression problem where the source is i.i.d. $\sim p_X$ and the distortion measure is $d : \mathcal{X} \times \hat{\mathcal{X}} \to [0,\infty)$. Let $D$ be the target distortion, and $R > R(D)$ be the rate of the scheme where $R(D)$ is the rate-distortion function. Let $p_{X,\hat{X}^\star}$ be the rate-distortion achieving distribution.

For block length $n$, we propose a new lossy compression scheme that mimics successive refinement with $s-1$ decoders. Similar to the previous section, let $\epsilon_k = (k-1)\min_u p_{U_k}(u)$ and

$$\begin{aligned} \hat{X}^\star = U_s &= U_{s-1} \oplus_s Z^{(s)}_{\epsilon_s} \\ U_{s-1} &= U_{s-2} \oplus_{s-1} Z^{(s-1)}_{\epsilon_{s-1}} \\ &\vdots \\ U_2 &= U_1 \oplus_2 Z^{(2)}_{\epsilon_2}. \end{aligned}$$

We further let $R_{k-1} > I(X;U_k) - I(X;U_{k-1})$ for $2 \le k \le s$ that satisfy $R = \sum_{k=2}^{s} R_{k-1}$. Now, we are ready to describe our coding scheme. Generate a sub-codebook $\mathcal{C}_1 = \{z^n(1,m) : 1 \le m \le e^{R_1}\}$ where each sequence is generated according to $Z^n \sim$ i.i.d. $p_{Z^{(2)}_{\epsilon_2}}$ for all $m$. Similarly, generate sub-codebooks $\mathcal{C}_k = \{z^n(k,m) : 1 \le m \le e^{nR_k}\}$ for $2 \le k \le s-1$ where each sequence is generated according to $Z^n \sim$ i.i.d. $p_{Z^{(k+1)}_{\epsilon_{k+1}}}$ for all $m$.

Upon observing $x^n \in \mathcal{X}^n$, the encoder finds $m_1 \in \mathcal{C}_1$ that minimizes $d_1(x^n, z^n(1,m_1))$ where the distortion measure $d_1(\cdot,\cdot)$ is defined as follows.

$$d_1(x^n, z^n) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{1}{p_{X|U_2}(x_i|z_i)}.$$

Note that $d_1(x,z)$ is simply the logarithmic loss between $x$ and $p_{X|U_2}(\cdot|z)$.

Similarly, for $2 \le k \le s-1$, the encoder iteratively finds $m_k \in \mathcal{C}_k$ that minimizes $d_k\left(x^n, [[z^n(1,m_1) \oplus_3 \cdots \oplus_k z^n(k-1,m_{k-1})] \oplus_{k+1} z^n(k,m_k)]\right)$ where

$$d_k(x^n, z^n) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{1}{p_{X|U_{k+1}}(x_i|z_i)}.$$

Upon receiving $m_1, m_2, \ldots, m_{s-1}$, the decoder reconstructs

$$\hat{x}^n = [[z^n(1,m_1) \oplus_3 z^n(2,m_2)] \oplus \cdots \oplus_s z^n(s-1,m_{s-1})].$$

Suppose $R_1 \approx R_2 \approx \cdots \approx R_{s-1} \approx \frac{R}{s-1}$, and $L = s-1$. Similar to [12,14], this scheme has two main advantages compare to naive random coding scheme. First, the number of codewords in the proposed scheme is $L \cdot e^{nR/L}$, while the naive scheme requires $e^{nR}$ codewords. Also, in each iteration, the encoder finds the best codeword among $e^{nR/L}$ sub-codewords. Thus, the overall complexity is $L \cdot e^{nR/L}$ as well. On the other hand, the naive scheme requires $e^{nR}$ complexity.

**Remark 3.** *The proposed scheme constructs $\hat{X}^n$ from binary sequences. The reconstruction after each stage can be viewed as*

$$u_k^n(m_1, \ldots m_{k-1}) = [[z^n(1,m_1) \oplus_3 \cdots] \oplus_k z^n(k-1,m_{k-1})]$$

*where $0 \le u_k \le k-1$. Thus, the decoder starts from binary sequence $u_2^n(m_1)$, and the alphabet size increases by 1 at each iteration. After $(s-1)$-th iteration, it reaches the final reconstruction $\hat{X}^n$ where the size of alphabet is $s$.*

## 5. Conclusions

To conclude our discussion, we summarize our main contributions. In the context of successive refinement problem, we showed another universal property of logarithmic loss that any discrete memoryless source is successively refinable as long as the intermediate decoders operate under logarithmic loss. We applied the result to the point-to-point lossy compression problem and proposed a lossy compression scheme with lower complexity.

**Funding:** This work was supported by 2017 Hongik University Research Fund.

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

## Appendix A. Proof of Lemma 2

For $y \leq s-1$,

$$\begin{aligned} p_{X|Y}(x|y) =& \sum_{\hat{x}} p_{X|\hat{X}^\star}(x|\hat{x}) p_{\hat{X}^\star|Y}(\hat{x}|y) \\ =& p_{X|\hat{X}^\star}(x|y). \end{aligned}$$

On the other hand, for $y = s$,

$$\begin{aligned} p_{X|Y}(x|y) =& \sum_{\hat{x}} p_{X|\hat{X}^\star}(x|\hat{x}) p_{\hat{X}^\star|Y}(\hat{x}|y) \\ =& \sum_{\hat{x}} p_{X|\hat{X}^\star}(x|\hat{x}) p_{\hat{X}^\star}(\hat{x}) \\ =& p_X(x). \end{aligned}$$

Let $\jmath_X(x, D)$ be $d$-tilted information [15]:

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

where $\lambda^\star = -R'(D)$ and the expectation is with respect to marginal distribution $p_{\hat{X}^\star}$. Csiszár [16] showed that for $p_{\hat{X}^\star}$-almost every $\hat{x}$,

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

If $p_{X|\hat{X}^\star}(x|\hat{x}_1) = p_{X|\hat{X}^\star}(x|\hat{x}_2)$ for all $x$, it implies that $d(x,\hat{x}_1) = d(x,\hat{x}_2)$ for all $x$ which contradicts our assumption. On the other hand, if $p_{X|\hat{X}^\star}(x|\hat{x}_1) = p_X(x)$ for all $x$, it also contradicts our assumption. Thus, $p_{X|Y}(\cdot|y)$ are different from each other for all $0 \leq y \leq s$.

## Appendix B. Proof of the Special Case of Theorem 2

Similar to the main proof of Theorem 2, we assume $\hat{\mathcal{X}} = \mathcal{Y} = \{0, 1, \ldots, s-1\}$. Suppose there exists $\hat{x}$ such that $p_{X|\hat{X}^\star}(x|\hat{x}) = p_X(x)$ for all $x$. Without loss of generality, we assume $\hat{x} = 0$, i.e., $p_{X|\hat{X}^\star}(x|0) = p_X(x)$ for all $x$.

Consider a random variable $Y \in \mathcal{Y}$ with the following conditional pmf for some $0 \leq \epsilon \leq 1$:

$$p_{Y|\hat{X}^\star}(y|\hat{x}) = \begin{cases} 1 & \text{if } \hat{x} = y = 0 \\ \epsilon & \text{if } \hat{x} \neq 0 \text{ and } y = 0 \\ 1-\epsilon & \text{if } \hat{x} = y \neq 0. \\ 0 & \text{otherwise.} \end{cases}$$

It is clear that $H(X|Y) = H(X|\hat{X}^\star)$ if $\epsilon = 0$ and $H(X|Y) = H(X)$ if $\epsilon = 1$. Since $H(X|\hat{X}^\star) \leq D_1$, there exists an $0 \leq \epsilon \leq 1$ such that $H(X|Y) = D_1$. We also have $Q = p_{X|Y}(\cdot|Y)$ and $q^{(y)} = p_{X|Y}(\cdot|y)$ for all $y \in \mathcal{Y}$. The following lemma implies the one-to-one mapping between $q$ and $y$.

**Lemma A1.** *If $p_{X|Y}(x|y_1) = p_{X|Y}(x|y_2)$ for all $x \in \mathcal{X}$, then $y_1 = y_2$.*

**Proof.** If $y = 0$, the conditional distribution $p_{\hat{X}^\star|Y}(\hat{x}|y)$ is given by

$$p_{\hat{X}^\star|Y}(\hat{x}|y) = \begin{cases} \frac{p_{\hat{X}^\star}(0)}{(1-\epsilon)p_{\hat{X}^\star}(0)+\epsilon} & \text{if } \hat{x} = 0 \\ \frac{\epsilon \cdot p_{\hat{X}^\star}(\hat{x})}{(1-\epsilon)p_{\hat{X}^\star}(0)+\epsilon} & \text{if } \hat{x} \neq 0. \end{cases}$$

Then,

$$\begin{aligned} p_{X|Y}(x|y) =& \sum_{\hat{x}} p_{X|\hat{X}^\star}(x|\hat{x}) p_{\hat{X}^\star|Y}(\hat{x}|0) \\ =& p_{X|\hat{X}^\star}(x|0) \frac{p_{\hat{X}^\star}(0)}{(1-\epsilon)p_{\hat{X}^\star}(0)+\epsilon} + \sum_{\hat{x}\neq 0} p_{X|\hat{X}^\star}(x|\hat{x}) \frac{\epsilon \cdot p_{\hat{X}^\star}(\hat{x})}{(1-\epsilon)p_{\hat{X}^\star}(0)+\epsilon} \\ =& p_{X|\hat{X}^\star}(x|0) \frac{(1-\epsilon)\cdot p_{\hat{X}^\star}(0)}{(1-\epsilon)p_{\hat{X}^\star}(0)+\epsilon} + p_X(x) \frac{\epsilon}{(1-\epsilon)p_{\hat{X}^\star}(0)+\epsilon} \\ =& p_{X|\hat{X}^\star}(x|0) \end{aligned}$$

where the last equality is because $p_{X|\hat{X}^\star}(x|0) = p_X(x)$ for all $x$. In other words, $p_{X|Y}(x|0) = p_{X|\hat{X}^\star}(x|0)$.

On the other hand, if $y \neq 0$, the conditional distribution $p_{\hat{X}^\star|Y}(\hat{x}|y)$ is given by

$$p_{\hat{X}^\star|Y}(\hat{x}|y) = \begin{cases} 1 & \text{if } \hat{x} = y \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$\begin{aligned} p_{X|Y}(x|y) =& \sum_{\hat{x}} p_{X|\hat{X}^\star}(x|\hat{x}) p_{\hat{X}^\star|Y}(\hat{x}|y) \\ =& p_{X|\hat{X}^\star}(x|y). \end{aligned}$$

As we have seen in Appendix A, $p_{X|\hat{X}^\star}(\cdot|\hat{x}_1)$ cannot be equal to $p_{X|\hat{X}^\star}(\cdot|\hat{x}_2)$ if $\hat{x}_1 \neq \hat{x}_2$. Since $p_{X|Y}(x|y) = p_{X|\hat{X}^\star}(x|y)$ for all $x$, we can say that $p_{X|Y}(x|y_1) = p_{X|Y}(x|y_2)$ for all $x$ implies $y_1 = y_2$. □

The remaining part of the proof is exactly the same as the main proof.

## References

1. Courtade, T.A.; Wesel, R.D. Multiterminal source coding with an entropy-based distortion measure. In Proceedings of the 2011 IEEE International Symposium on Information Theory Proceedings, St. Petersburg, Russia, 31 July–5 August 2011; pp. 2040–2044.
2. Courtade, T.; Weissman, T. Multiterminal Source Coding Under Logarithmic Loss. *IEEE Trans. Inf. Theory* **2014**, *60*, 740–761. [CrossRef]
3. 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]
4. 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.
5. 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.
6. 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.
7. Equitz, W.H.; Cover, T.M. Successive refinement of information. *IEEE Trans. Inf. Theory* **1991**, *37*, 269–275. [CrossRef]
8. Koshelev, V. Hierarchical coding of discrete sources. *Probl. Peredachi Inf.* **1980**, *16*, 31–49.
9. Gerrish, A.M. Estimation of Information Rates. Ph.D. Thesis, Yale University, New Haven, CT, USA, 1963.
10. Chow, J.; Berger, T. Failure of successive refinement for symmetric Gaussian mixtures. *IEEE Trans. Inf. Theory* **1997**, *43*, 350–352. [CrossRef]
11. Lastras, L.; Berger, T. All sources are nearly successively refinable. *IEEE Trans. Inf. Theory* **2001**, *47*, 918–926. [CrossRef]
12. Venkataramanan, R.; Sarkar, T.; Tatikonda, S. Lossy Compression via Sparse Linear Regression: Computationally Efficient Encoding and Decoding. *IEEE Trans. Inf. Theory* **2014**, *60*, 3265–3278. [CrossRef]
13. No, A.; Weissman, T. Rateless lossy compression via the extremes. *IEEE Trans. Inf. Theory* **2016**, *62*, 5484–5495. [CrossRef] [PubMed]
14. No, A.; Ingber, A.; Weissman, T. Strong successive refinability and rate-distortion-complexity tradeoff. *IEEE Trans. Inf. Theory* **2016**, *62*, 3618–3635. [CrossRef]
15. Kostina, V.; Verdú, S. Fixed-length lossy compression in the finite blocklength regime. *IEEE Trans. Inf. Theory* **2012**, *58*, 3309–3338. [CrossRef]
16. Csiszár, I. On an extremum problem of information theory. *Stud. Sci. Math. Hung.* **1974**, *9*, 57–71.