A Coding Theorem for f-Separable Distortion Measures

Yanina Shkel
Princeton University NJ 08544, USA
Email: yshkel@princeton.edu

Sergio Verdú
Princeton University NJ 08544, USA
Email: verdu@princeton.edu

***Abstract*—Rate-distortion theory is a branch of information theory that provides theoretical foundation for lossy data compression. In this setting, the decompressed data need not match original data exactly; however, it must be reconstructed with a prescribed fidelity, which is modeled by a distortion measure. An ubiquitous assumption in rate-distortion literature is that such distortion measures are *separable*: that is, the distortion measure can be expressed as an arithmetic average of single-letter distortions. Such set up gives nice theoretical results at the expense of a very restrictive model. Separable distortion measures are linear functions of single-letter distortions; real-world distortion measures rarely have such nice structure. In this work we relax the separability assumption and propose f-separable distortion measures, which are well suited to model non-linear penalties. We prove a rate-distortion coding theorem for stationary ergodic sources with f-separable distortion measures, and provide some illustrative examples of the resulting rate-distortion functions.**

I. Introduction

Rate-distortion theory, a branch of information theory that studies models for lossy data compression, was introduced by Claude Shannon in [1]. The approach of [1] is to model the information source with distribution P_X on $\mathcal{X}$, a reconstruction alphabet $\hat{\mathcal{X}}$, and a distortion measure $\mathsf{d}\colon \mathcal{X}\times\hat{\mathcal{X}} \to [0,\infty)$. When the information source produces a sequence of n realizations, the source P_{X^n} is defined on $\mathcal{X}^n$ with reconstruction alphabet $\hat{\mathcal{X}}^n$, where $\mathcal{X}^n$ and $\hat{\mathcal{X}}^n$ are n-fold Cartesian products of $\mathcal{X}$ and $\hat{\mathcal{X}}$. In that case, [1] extended the notion of a single-letter distortion measure to the n-letter distortion measure, $\mathsf{d}^n\colon \mathcal{X}^n \times \hat{\mathcal{X}}^n \to [0,\infty)$, by taking an arithmetic average of single-letter distortions,

$$\mathsf{d}^n(x^n,\hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{d}(x_i,\hat{x}_i). \tag{1}$$

Distortion measures that satisfy (1) are referred to as *separable* (also additive, per-letter, averaging); the separability assumption has been ubiquitous throughout rate-distortion literature ever since its inception in [1].

On the one hand, the separability assumption is quite natural and allows for a tractable characterization of the fundamental trade-off between the rate of compression and the average distortion. For example, in the case when X^n is a stationary and memoryless source the rate-distortion function, which captures this trade-off, admits a simple characterization:

$$\mathcal{R}(d) = \inf_{P_{\hat{X}|X}\colon \mathbb{E}\left[\mathsf{d}(X,\hat{X})\right]\le d} I(X;\hat{X}). \tag{2}$$

On the other hand, the separability assumption is very restrictive as it only models distortion penalties that are *linear* functions of the per-letter distortions in the source reproduction. Real-world distortion measures, however, may be highly *non-linear*; it is desirable to have a theory that also accommodates non-linear distortion measures. To this end, we propose the following definition:

Definition 1 (f-separable distortion measure)**.** *Let $\mathsf{f}(z)$ be a continuous, increasing function on $[0,\infty)$. An n-letter distortion measure $\mathsf{d}^n(\cdot,\cdot)$ is f-separable with respect to a single-letter distortion $\mathsf{d}(\cdot,\cdot)$ if it can be written as*

$$\mathsf{d}^n(x^n,\hat{x}^n) = \mathsf{f}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathsf{f}\left(\mathsf{d}(x_i,\hat{x}_i)\right)\right). \tag{3}$$

For $\mathsf{f}(z)=z$ this is the classical separable distortion set up. By selecting f appropriately, it is possible to model a large class of non-linear distortion measures, see Figure 1 for illustrative examples.

In this work, we characterize the rate-distortion function for stationary and ergodic information sources with f-separable distortion measures. In the special case of memoryless and stationary sources we obtain the following intuitive result:

$$\mathcal{R}_{\mathsf{f}}(d) = \inf_{P_{\hat{X}|X}\colon \mathbb{E}\left[\mathsf{f}\left(\mathsf{d}(X,\hat{X})\right)\right]\le \mathsf{f}(d)} I(X;\hat{X}). \tag{4}$$

A pleasing implication of this result is that much of rate-distortion theory (e.g. the Blahut-Arimoto algorithm) developed since [1] can be leveraged to work under the far more general f-separable assumption.

The rest of this paper is structured as follows. The remainder of Section I overviews related work: Section I-A provides the intuition behind Definition 1, Section I-B reviews related work in other compression problems, and Section I-C connects f-separable distortion measures with sub-additive distortion measures. Section II formally sets up the problem. Section III presents our main result, Theorem 2, as well as some illustrative examples. Additional discussion about problem formulation and sub-additive distortion measures is given in Section IV. We conclude the paper in Section V.

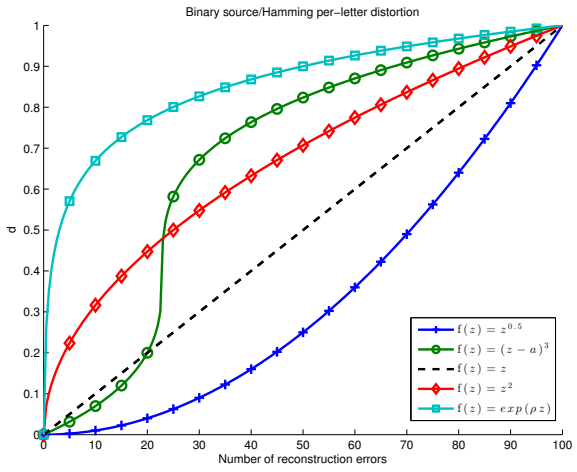

Fig. 1. The number of reconstruction errors for an information source with 100 bits vs. the penalty assessed by f-separable distortion measures based on the Hamming single-letter distortion. The $\mathsf{f}(z) = z$ plot corresponds to the separable distortion. The f-separable assumption accommodates all of the other plots, and many more, with the appropriate choice of the function f.

A. Generalized f-mean and Rényi entropy

To understand the intuition behind Definition 1, consider aggregating n numbers $\{z_1, \ldots, z_n\}$ by defining a function

$$\mathsf{M}_n(\mathbf{z}) = \mathsf{f}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathsf{f}(z_i)\right) \tag{5}$$

where f is a continuous, increasing function on the range of $\{z_1, \ldots, z_n\}$. It is easy to see that (5) satisfies the following properties:

1) $\mathsf{M}_n(\mathbf{z})$ is continuous and monotonically increasing in each z_i.
2) $\mathsf{M}_n(\mathbf{z})$ is a symmetric function of each z_i.
3) If $z_i = z$ for all i, then $\mathsf{M}_n(\mathbf{z}) = z$.
4) For any $m \leq n$

$$\mathsf{M}_n = (\mathsf{M}_m(\mathbf{z}_1^m), \ldots, \mathsf{M}_m(\mathbf{z}_1^m), z_{m+1}, \ldots, z_n). \tag{6}$$

Moreover, it is shown in [2] that any sequence of functions M_n that satisfies these properties must have the form of equation (5) for some continuous, increasing f. The function M_n is referred to as 'Kolmogorov mean', 'quazi-arithmetic mean', or 'generalized f-mean'. The most prominent examples are geometric mean, $\mathsf{f}(z) = \log z$, and root mean square, $\mathsf{f}(z) = z^2$.

The main insight behind Definition 1 is to define an n-letter distortion measure to be an f-mean of single-letter distortions. The f-separable distortion measures include all n-letter distortion measures that satisfy the above properties, with the last property saying that the non-linear 'shape' of distortion measure (cf. Figure 1) is independent of n.

Finally, we note that Rényi also arrived at his well-known family of entropies [3] by taking an f-mean of the information random variable:

$$H_\alpha(X) = \mathsf{f}_\alpha^{-1}\mathbb{E}\left[\mathsf{f}_\alpha(\imath_X(X))\right], \quad \alpha \in (0,1)\cup(1,\infty) \tag{7}$$

where the information random variable is

$$\imath_X(x) = \log\frac{1}{P_X(x)}. \tag{8}$$

Rényi limited his consideration to functions of the form $\mathsf{f}_\alpha(z) = \exp\{(1-\alpha)z\}$ in order to ensure that entropy is additive for independent random variables.

B. Compression with non-linear cost

Source coding with non-linear cost has already been explored in the variable-length lossless compression setting. Let $\ell(x)$ denote the length of the encoding of x by a given variable length code. Campbell [4], [5] proposed minimizing a cost function of the form

$$\mathsf{f}^{-1}\mathbb{E}\left[\mathsf{f}(\ell(X))\right], \tag{9}$$

instead of the usual expected length. The main result of [4], [5] is that for

$$\mathsf{f}_t(z) = \exp\{tz\}, \quad t \in (-1,0)\cup(0,\infty), \tag{10}$$

the fundamental limit of such set up is Rényi entropy of order $\alpha = \frac{1}{t+1}$. For more general f, this problem was handled by Kieffer [6], who showed that (9) has a fundamental limit for a large class of functions f. That limit is Rényi entropy of order $\alpha = \frac{1}{t+1}$ with

$$t = \lim_{z\to\infty}\frac{\mathsf{f}''(z)}{\mathsf{f}'(z)}. \tag{11}$$

More recently, a number of works [7]–[9] studied related source coding paradigms, such as guessing and task encoding. These works also focused on exponential functions given in (10); in [7], [8] Rényi entropy is shown to be a fundamental limit yet again. The main focus of [4]–[9] has been on memoryless sources.

C. Sub-additive distortion measures

A notable departure from the separability assumption in rate-distortion theory is sub-additive distortion measures discussed in [10]. Namely, a distortion measure is sub-additive if

$$\mathsf{d}^n(x^n, \hat{x}^n) \leq \frac{1}{n}\sum_{i=1}^{n}\mathsf{d}(x_i, \hat{x}_i). \tag{12}$$

In the present setting, an f-separable distortion measure is sub-additive if f is concave:

$$\mathsf{d}^n(x^n, \hat{x}^n) = \mathsf{f}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathsf{f}\left(\mathsf{d}(x_i, \hat{x}_i)\right)\right) \leq \frac{1}{n}\sum_{i=1}^{n}\mathsf{d}(x_i, \hat{x}_i). \tag{13}$$

Thus, the results for sub-additive distortion measures, such as the convexity of the rate-distortion function, are applicable to f-separable distortion measures when f is concave.

II. Preliminaries

Let X be a random variable defined on $\mathcal{X}$ with distribution P_X, with reconstruction alphabet $\hat{\mathcal{X}}$, and a distortion measure $\mathsf{d}\colon \mathcal{X} \times \hat{\mathcal{X}} \to [0,\infty)$. Let $\mathcal{M} = \{1,\dots,M\}$ be the message set.

Definition 2 (Lossy source code)**.** *A lossy source code* (g,c) *is a pair of mappings,*

$$\mathsf{g}\colon \mathcal{X} \to \mathcal{M} \tag{14}$$

$$\mathsf{c}\colon \mathcal{M} \to \hat{\mathcal{X}}. \tag{15}$$

A lossy source-code (g,c) is an (M,d)-lossy source code on $(\mathcal{X},\hat{\mathcal{X}},\mathsf{d})$ if

$$\mathbb{E}\left[\mathsf{d}(X,\mathsf{c}(\mathsf{g}(X))\right] \le d. \tag{16}$$

A lossy source code (g,c) is an (M,d,ϵ)-lossy source code on $(\mathcal{X},\hat{\mathcal{X}},\mathsf{d})$ if

$$\mathbb{P}\left[\mathsf{d}(X,\mathsf{c}(\mathsf{g}(X)) > d\right] \le \epsilon. \tag{17}$$

Definition 3. *An information source* $\mathbf{X}$ *is a stochastic process*

$$\mathbf{X} = \{X^n = (X_1,\dots,X_n)\}_{n=1}^{\infty}. \tag{18}$$

If (g,c) is an (M,d)-lossy source code for X^n on $(\mathcal{X}^n,\hat{\mathcal{X}}^n,\mathsf{d}^n)$, we say (g,c) is an (n,M,d)-lossy source code. Likewise, an (M,d,ϵ)-lossy source code for X^n on $(\mathcal{X}^n,\hat{\mathcal{X}}^n,\mathsf{d}^n)$ is an (n,M,d,ϵ)-lossy source code.

Recall f-separable distortion measures introduced in Definition 1. We remark that it is sufficient to consider f which are increasing and continuous on $[a_{\min},a_{\max}]$,

$$a_{\min} = \inf_{x\in\mathcal{X},\hat{x}\in\hat{\mathcal{X}}} \mathsf{d}(x,\hat{x}), \tag{19}$$

$$a_{\max} = \sup_{x\in\mathcal{X},\hat{x}\in\hat{\mathcal{X}}} \mathsf{d}(x,\hat{x}). \tag{20}$$

A. Rate-distortion function (average distortion)

Definition 4. *Let a sequence of distortion measures* $\{\mathsf{d}^n\}$ *be given. The rate-distortion pair* (R,d) *is achievable if there exists a sequence of* (n,M_n,d_n)*-lossy source codes such that*

$$\limsup_{n\to\infty} \frac{1}{n}\log M_n \le R, \quad \textit{and} \quad \limsup_{n\to\infty} d_n \le d.$$

Our main object of study is the following rate-distortion function with respect to f-separable distortion measures.

Definition 5. *Let* $\{\mathsf{d}^n\}$ *be a sequence of* f*-separable distortion measures. Then,*

$$\mathcal{R}_{\mathsf{f}}(d) = \inf\{R\colon (R,d) \textit{ is achievable}\}. \tag{21}$$

In the case when f is the identity, we omit the subscript f and simply write $\mathcal{R}(d)$.

B. Rate-distortion function (excess distortion)

It is useful to consider the rate-distortion function for f-separable distortion measures under the excess distortion paradigm.

Definition 6. *Let a sequence of distortion measures* $\{\mathsf{d}^n\}$ *be given. The rate-distortion pair* (R,d) *is (excess distortion) achievable if for any* $\gamma > 0$ *there exists a sequence of* $(n,M_n,d+\gamma,\epsilon_n)$*-lossy source codes such that*

$$\limsup_{n\to\infty} \frac{1}{n}\log M_n \le R, \quad \textit{and} \quad \limsup_{n\to\infty} \epsilon_n = 0.$$

Definition 7. *Let* $\{\mathsf{d}^n\}$ *be a sequence of* f*-separable distortion measures. Then,*

$$\mathcal{R}'_{\mathsf{f}}(d) = \inf\{R\colon (R,d) \textit{ is (excess distortion) achievable}\}. \tag{22}$$

Characterizing the rate-distortion function is particularly simple under the excess distortion paradigm, as shown in the following lemma.

Lemma 1. *Let the single-letter distortion* d *and an increasing, continuous* f *be given. Then,*

$$\mathcal{R}'_{\mathsf{f}}(d) = \tilde{\mathcal{R}}'(\mathsf{f}(d)) \tag{23}$$

where $\tilde{\mathcal{R}}'(d)$ *is computed with respect to* $\tilde{\mathsf{d}}(x,\hat{x}) = \mathsf{f}(\mathsf{d}(x,\hat{x}))$.

Proof. Let $\{\mathsf{d}^n\}$ be a sequence of f-separable distortions based on $\mathsf{d}(\cdot,\cdot)$ and let $\left\{\tilde{\mathsf{d}}^n\right\}$ be a sequence of separable distortion measures based on $\tilde{\mathsf{d}}(\cdot,\cdot) = \mathsf{f}(\mathsf{d}(\cdot,\cdot))$.

Since f is increasing and continuous at d, then for any $\gamma > 0$ there exists $0 < \tilde{\gamma}$ such that

$$\mathsf{f}(d+\gamma) - \mathsf{f}(d) = \tilde{\gamma}. \tag{24}$$

The reverse is also true by continuity of f: for any $\tilde{\gamma} > 0$ there exists $\gamma > 0$ such that (24) is satisfied.

Any source code $(\mathsf{g}_n,\mathsf{c}_n)$ is an $(n,M_n,d+\gamma,\epsilon_n)$-lossless code under f-separable distortion d^n if and only if $(\mathsf{g}_n,\mathsf{c}_n)$ is also an $(n,M_n,\mathsf{f}(d)+\tilde{\gamma},\epsilon_n)$-lossless code under separable distortion $\tilde{\mathsf{d}}^n$. Indeed,

$$\epsilon_n \ge \mathbb{P}\left[\mathsf{d}^n(X^n,\mathsf{c}_n\left(\mathsf{g}_n(X^n)\right)) \ge d+\gamma\right] \tag{25}$$

$$= \mathbb{P}\left[\mathsf{f}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathsf{f}(\mathsf{d}(X_i,\hat{X}_i))\right) \ge d+\gamma\right] \tag{26}$$

$$= \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\mathsf{f}(\mathsf{d}(X_i,\hat{X}_i)) \ge \mathsf{f}(d+\gamma)\right] \tag{27}$$

$$= \mathbb{P}\left[\tilde{\mathsf{d}}^n(X^n,\mathsf{c}\left(\mathsf{g}(X^n)\right)) \ge \mathsf{f}(d)+\tilde{\gamma}\right] \tag{28}$$

where $\hat{X}^n = \mathsf{c}_n\left(\mathsf{g}_n(X^n)\right)$. It follows that (R,d) is (excess distortion) achievable with respect to $\{\mathsf{d}^n\}$ if and only if $(R,\mathsf{f}(d))$ is (excess distortion) achievable with respect to $\left\{\tilde{\mathsf{d}}^n\right\}$. The lemma statement follows from this observation and Definition 6. □

III. Main Result

In this section we make the following standard assumptions.

1) $\mathbf{X}$ is a stationary and ergodic source.
2) The single-letter distortion function $\mathsf{d}(\cdot,\cdot)$ and the continuous and increasing function $\mathsf{f}(\cdot)$ are such that

$$\inf_{\hat{x}\in\hat{\mathcal{X}}} \mathbb{E}\left[\mathsf{f}(\mathsf{d}(X,\hat{x}))\right] < \infty. \tag{29}$$

3) For each $d>0$, there exists a countable subset $\{\hat{x}_i\}$ of $\hat{\mathcal{X}}$ and a countable measurable partition $\{E_i\}$ of $\mathcal{X}$ such that $\mathsf{d}(x,\hat{x}_i)\le d$, $x\in E_i$ for each $\hat{x}_i$, and

$$-\sum_i P_{X_1}(E_i)\log P_{X_1}(E_i) < \infty. \tag{30}$$

Theorem 2. *Under the stated assumptions, the rate-distortion function is given by*

$$\mathcal{R}_{\mathsf{f}}(d) = \tilde{\mathcal{R}}(\mathsf{f}(d)) \tag{31}$$

where

$$\tilde{\mathcal{R}}(\mathsf{f}(d)) = \lim_{n\to\infty} \inf_{P_{\hat{X}^n|X^n}:\ \frac{1}{n}\sum_{i=1}^n \mathbb{E}\tilde{\mathsf{d}}(X_i,Y_i)\le \mathsf{f}(d)} \frac{1}{n} I(X^n;\hat{X}^n) \tag{32}$$

is the rate-distortion function computed with respect to the separable distortion measure given by $\tilde{\mathsf{d}}(x,\hat{x}) = \mathsf{f}(\mathsf{d}(x,\hat{x}))$.

For stationary memoryless sources (31) particularizes to

$$\mathcal{R}_{\mathsf{f}}(d) = \inf_{P_{\hat{X}|X}:\ \mathbb{E}\left[\mathsf{f}\left(\mathsf{d}(X,\hat{X})\right)\right]\le \mathsf{f}(d)} I(X;\hat{X}). \tag{33}$$

Proof. Equations (32) and (33) are widely known in literature (see, for example, [10]–[12]); it remains to show (31). Under the stated assumptions,

$$\mathcal{R}_{\mathsf{f}}(d) \overset{(a)}{\le} \mathcal{R}'_{\mathsf{f}}(d) \overset{(b)}{=} \tilde{\mathcal{R}}'(\mathsf{f}(d)) \overset{(c)}{=} \tilde{\mathcal{R}}(\mathsf{f}(d)) \tag{34}$$

where (a) follows from assumption (2) and Theorem 4 in the Appendix, (b) is shown in Lemma 1, and (c) is due to [13] (see also [12, Theorem 5.9.1]). The other direction,

$$\mathcal{R}_{\mathsf{f}}(d) \ge \tilde{\mathcal{R}}(\mathsf{f}(d)) \tag{35}$$

is a consequence of the strong converse by Kieffer [14], see Lemma 6 in the Appendix. □

Example 1 (BMS, Hamming)**.** *Let* $\mathbf{X}$ *be the binary memoryless source. That is,* $\mathcal{X}=\hat{\mathcal{X}}=\{0,1\}$*,* X_i *is a Bernoulli*(p) *random variable, and* $\mathsf{d}(\cdot,\cdot)$ *is the usual Hamming distortion measure. Then, for any continuous increasing* $\mathsf{f}(\cdot)$ *and* $p\le\frac{1}{2}$*,*

$$\mathcal{R}_{\mathsf{f}}(d) = \left\{ \begin{array}{cc} h(p) - h\left(\frac{\mathsf{f}(d)-\mathsf{f}(0)}{\mathsf{f}(1)-\mathsf{f}(0)}\right), & \frac{\mathsf{f}(d)-\mathsf{f}(0)}{\mathsf{f}(1)-\mathsf{f}(0)} < p \\ 0, & o.w. \end{array} \right\}.$$

The result follows from a series of obvious equalities,

$$\mathcal{R}_{\mathsf{f}}(d) = \inf_{P_{\hat{X}|X}:\mathbb{E}\left[\mathsf{f}\left(\mathsf{d}(X,\hat{X})\right)\right]\le\mathsf{f}(d)} I(X;\hat{X}) \tag{36}$$

$$= \inf_{P_{\hat{X}|X}:\ \frac{\mathbb{E}\left[\mathsf{f}\left(\mathsf{d}(X,\hat{X})\right)\right]-\mathsf{f}(0)}{\mathsf{f}(1)-\mathsf{f}(0)}\le\frac{\mathsf{f}(d)-\mathsf{f}(0)}{\mathsf{f}(1)-\mathsf{f}(0)}} I(X;\hat{X}) \tag{37}$$

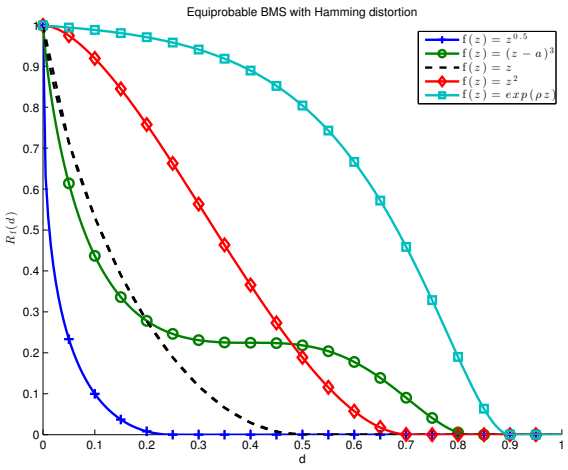

Fig. 2. $\mathcal{R}_{\mathsf{f}}(d)$ for the Binary Memoryless Source with $p=0.5$. Compare these to the f-separable distortion measures plotted for the binary source with Hamming distortion in Figure 1.

$$= \inf_{P_{\hat{X}|X}:\ \mathbb{E}\left[\frac{\mathsf{f}(\mathsf{d}(X,\hat{X}))-\mathsf{f}(0)}{\mathsf{f}(1)-\mathsf{f}(0)}\right]\le\frac{\mathsf{f}(d)-\mathsf{f}(0)}{\mathsf{f}(1)-\mathsf{f}(0)}} I(X;\hat{X}) \tag{38}$$

$$= \inf_{P_{\hat{X}|X}:\ \mathbb{E}\left[\mathsf{d}(X,\hat{X})\right]\le\frac{\mathsf{f}(d)-\mathsf{f}(0)}{\mathsf{f}(1)-\mathsf{f}(0)}} I(X;\hat{X}) \tag{39}$$

$$= \mathcal{R}\left(\frac{\mathsf{f}(d)-\mathsf{f}(0)}{\mathsf{f}(1)-\mathsf{f}(0)}\right). \tag{40}$$

The rate-distortion function given in Example 1 is plotted in Figure 2 for different functions f. Observe that for concave f (i.e. subadditive distortion) the resulting rate-distortion function is convex, which is consistent with [10]. However, for f that are not concave, the rate-distortion function is not always convex. This is a simple example of the rate-distortion function not being convex in general. We also remark that the simple derivation in Example 1 could be applied to any source for which the single-letter distortion measure can take on only two values.

Theorem 2 shows that for well-behaved stationary ergodic sources $\mathcal{R}_{\mathsf{f}}(d)$ admits a simple characterization. According to Lemma 1, the same characterization holds within excess distortion paradigm without the stationary and ergodic assumption. The next example shows that, in general, $\mathcal{R}_{\mathsf{f}}(d)\ne\tilde{\mathcal{R}}(\mathsf{f}(d)$ within the average distortion paradigm. Thus, assumption (1) is necessary for Theorem 2 to hold.

Example 2 (Mixed Source)**.** *Fix* $\lambda\in(0,1)$ *and let the source* $\mathbf{X}$ *be a mixture of two i.i.d. sources,*

$$P_{X^n}(x^n) = \lambda\prod_{i=1}^n P^1(x_i) + (1-\lambda)\prod_{i=1}^n P^2(x_i). \tag{41}$$

We can alternatively express $\mathbf{X}$ *as*

$$X^n = ZX_1^n + (1-Z)X_2^n \tag{42}$$

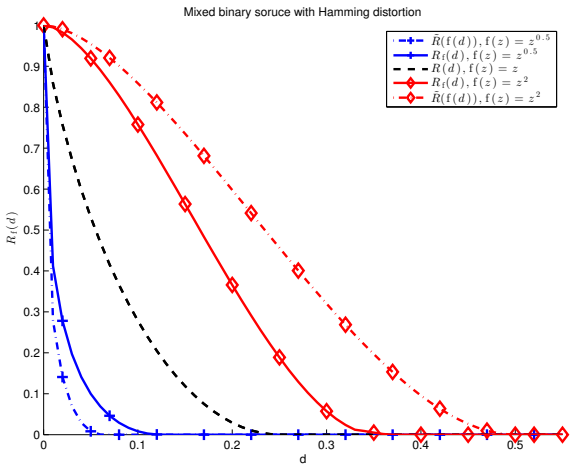

Fig. 3. Mixed binary source with $p_1 = 0.5$, $p_2 = 0.001$, and $\lambda = 0.5$. Three examples of f-separable rate-distortion functions are given. For $\mathsf{f}(z) = z$, the relation $\mathcal{R}(d) = \tilde{\mathcal{R}}(d)$ follows immediately. When f is not the identity, $\mathcal{R}_{\mathsf{f}}(d) \neq \tilde{\mathcal{R}}(\mathsf{f}(d))$ in general for non-ergodic sources.

where Z is a Bernoulli(λ) random variable. Then, the rate-distortion function for the mixture source (41) and continuous increasing f is given in Lemma 7 in the Appendix. Namely,

$$\mathcal{R}_{\mathsf{f}}(d) = \min_{(d_1,d_2):\ \lambda d_1 + (1-\lambda) d_2 \le d} \max\left(\mathcal{R}^1_{\mathsf{f}}(d_1), \mathcal{R}^2_{\mathsf{f}}(d_2)\right) \quad (43)$$

where $\mathcal{R}^1_{\mathsf{f}}(d)$ and $\mathcal{R}^2_{\mathsf{f}}(d)$ are the rate-distortion functions for DMSs given by P^1 and P^2, respectively. Likewise,

$$\tilde{\mathcal{R}}\left(\mathsf{f}(d)\right) = \min_{(d_1,d_2):\ \lambda d_1 + (1-\lambda) d_2 \le \mathsf{f}(d)} \max\left(\tilde{\mathcal{R}}^1(d_1), \tilde{\mathcal{R}}^2(d_2)\right). \quad (44)$$

As shown in Figure 3, equations (43) and (44) are not equal in general.

IV. Discussion

A. Sub-additive distortion measures

Recall that an f-separable distortion measure is sub-additive if f is concave (cf. Section I-C) . Clearly, not all f-separable distortion measures are sub-additive, and not all sub-additive distortion measures are f-separable. An examplar of a sub-additive distortion measure (which is not f-separable) given in [10, Chapter 5.2] is

$$\mathsf{d}^n(x^n, \hat{x}^n) = \frac{1}{n}\left(\sum_{i=1}^n \mathsf{d}(x_i, \hat{x}_i)^q\right)^{1/q}, \quad q > 1. \quad (45)$$

The sub-additivity of (45) follows from Minkowski inequality. Compare this to a sub-additive, f-separable distortion measure given by

$$\mathsf{d}^n(x^n, \hat{x}^n) = \left(\frac{1}{n}\sum_{i=1}^n \mathsf{d}(x_i, \hat{x}_i)^q\right)^{1/q}, \quad 0 \le q \le 1. \quad (46)$$

Observe that the discrepancy between different ranges of q in (45) and (46) has to do with the scaling $\frac{1}{n}$ factor.

Consider a binary source with Hamming distortion and let $x^n = \mathbf{0}^n, \hat{x}^n = \mathbf{1}^n$. Rewriting (45) we obtain

$$\mathsf{d}^n(x^n, \hat{x}^n) = \frac{1}{n^{(q-1)/q}}\left(\frac{1}{n}\sum_{i=1}^n \mathsf{d}(x_i, \hat{x}_i)^q\right)^{1/q} \quad (47)$$

and

$$\lim_{n\to\infty} \mathsf{d}^n(x^n, \hat{x}^n) = \lim_{n\to\infty} \frac{1}{n^{(q-1)/q}}\left(\frac{1}{n}\sum_{i=1}^n \mathsf{d}(0,1)^q\right)^{1/q} \quad (48)$$

$$= \lim_{n\to\infty} \frac{1}{n^{(q-1)/q}}\left(\frac{1}{n}\sum_{i=1}^n 1\right)^{1/q} \quad (49)$$

$$= \lim_{n\to\infty} \frac{1}{n^{(q-1)/q}} = 0. \quad (50)$$

In the binary example, the limiting distortion of (45) is zero even when the reconstruction of x^n gets every single symbol wrong. It is easy to check that example (45) is similarly degenerate in many cases of interest. The distortion measure given by (46), on the other hand, is an example of non-trivial sub-additive distortion measure, as can be seen in Figure 2 for $q = \frac{1}{2}$.

B. A consequence of Theorem 2

In light of discussion in Section I-A, an alert reader my consider modifying equation (16) to

$$\mathsf{f}^{-1}\left(\mathbb{E}\left[\mathsf{f}\left(\mathsf{d}(X, \mathsf{c}(\mathsf{g}(X)))\right)\right]\right) \le d, \quad (51)$$

and studying the (M, d)-lossy source codes under this new paradigm. Call the corresponding rate-distortion function $\mathcal{R}^{\mathsf{f}}(d)$ and assume that n-letter distortion measures are separable. Thus, at block length n the constraint (51) is

$$\mathbb{E}\left[\mathsf{f}\left(\frac{1}{n}\sum_{i=1}^n \mathsf{d}(X_i, \hat{X}_i)\right)\right] \le \mathsf{f}(d) \quad (52)$$

where $\hat{X} = \mathsf{c}(\mathsf{g}(X))$. This is equivalent to the following constraints:

$$\mathbb{E}\left[\mathsf{f}\left(\frac{1}{n}\sum_{i=1}^n \mathsf{f}^{-1}(\tilde{\mathsf{d}}(X_i, \hat{X}_i))\right)\right] \le \mathsf{f}(d) \quad (53)$$

$$\text{and} \quad \mathbb{E}\left[\tilde{\mathsf{d}}^n(X_i, \hat{X}_i))\right] \le \mathsf{f}(d) \quad (54)$$

where $\tilde{\mathsf{d}}^n$ is an f^{-1}-separable distortion measure. Putting these observations together with Theorem 2 yields

$$\mathcal{R}^{\mathsf{f}}(d) = \tilde{\mathcal{R}}_{\mathsf{f}^{-1}}(\mathsf{f}(d)) = \mathcal{R}(\mathsf{f}^{-1}(\mathsf{f}(d))) = \mathcal{R}(d). \quad (55)$$

A consequence of Theorem 2 is that the rate distortion function remains unchanged under this new paradigm.

V. Conclusion

This paper proposes f-separable distortion measures as a good model for non-linear distortion penalties. The rate-distortion function for f-separable distortion measures is characterized in terms of separable rate-distortion function with respect to a new single-letter distortion measure, $\mathsf{f}(\mathsf{d}(\cdot,\cdot))$. This characterization is straightforward for the excess distortion paradigm, as seen in Lemma 1. The proof is more involved for the average distortion paradigm, as seen in Theorem 2. An important implication of Theorem 2 is that many prominant results in rate-distortion literature (e.g. Blahut-Arimoto algorithm) can be leveraged to work for f-separable distortion measures.

Finally, we mention that a similar generalization is well-suited for channels with non-linear costs. That is, we say that b^n is an f-separable cost function if it can be written as

$$\mathsf{b}^n(x^n) = \mathsf{f}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathsf{f}\left(\mathsf{b}(x_i)\right)\right). \tag{56}$$

With this generalization we can state the following theorem.

Theorem 3 (Channels with cost)**.** *The capacity of a stationary memoryless channel given by $P_{Y|X}$ and f-separable cost function based on single-letter function $\mathsf{b}(x)$ is*

$$C_{\mathsf{f}}(\beta) = \sup_{P_X\,:\,\mathbb{E}[\mathsf{f}(\mathsf{b}(X))]\le\mathsf{f}(\beta)} I(X;Y). \tag{57}$$

The proof of Theorem 3 is left for the journal version of this paper.

Appendix

A. Lemmas for Theorem 2

Theorem 4 can be distilled from several proofs in literature. We state it here, with proof, for completeness; it is given in its present form in [15]. The condition of Theorem 4 applies when the source satisfies assumptions (1)-(3) in Section III. This is a consequence of the ergodic theorem and continuity of f.

Theorem 4. *Suppose that the source and distortion measure are such that for any $\gamma > 0$ there exists $0 < \triangle_\gamma < \infty$ and a sequence $b_1, b_2, \ldots$ such that*

$$\mathbb{E}\left[\mathsf{d}^n(X^n, b^n)1\{\triangle_\gamma < \mathsf{d}^n(X^n, b^n)\}\right] \le_n \gamma. \tag{58}$$

If a rate-distortion pair (R, d) is achievable under the excess distortion criterion, it is achievable under the average distortion criterion.

Proof. Choose $\gamma > 0$. Suppose there is a code $(\mathsf{g}^n, \mathsf{c}^n)$ with M codewords that achieves

$$\lim_{n\to\infty}\mathbb{P}\left[\mathsf{d}^n(X^n, \mathsf{c}^n(\mathsf{g}^n(X))) > d+\gamma\right] = 0. \tag{59}$$

We construct a new code $(\hat{\mathsf{g}}^n, \hat{\mathsf{c}}^n)$ with $M+1$ codewords:

$$\hat{\mathsf{c}}^n(m) = \begin{cases} b^n, & \text{if } m = 0 \\ \mathsf{c}^n(m), & \text{if } m = 1,\ldots,M, \end{cases} \tag{60}$$

$$\hat{\mathsf{g}}^n(x^n) = \begin{cases} 0, & \text{if } \mathsf{d}^n(x^n, \mathsf{c}^n(\mathsf{g}^n(x^n))) > \mathsf{d}^n(x^n, b^n) \\ \mathsf{g}^n(x^n), & \text{if } \mathsf{d}^n(x^n, \mathsf{c}^n(\mathsf{g}^n(x^n))) \le \mathsf{d}^n(x^n, b^n). \end{cases} \tag{61}$$

Then

$$\mathsf{d}^n(x^n, \hat{\mathsf{c}}^n(\hat{\mathsf{g}}^n(x^n))) = \min\left\{\mathsf{d}^n(x^n, \mathsf{c}^n(\mathsf{g}^n(x^n))), \mathsf{d}^n(x^n, b^n)\right\}. \tag{62}$$

For brevity denote,

$$V_n = \mathsf{d}^n(X^n, \hat{\mathsf{c}}^n(\hat{\mathsf{g}}^n(X^n))) \tag{63}$$

$$W_n = \mathsf{d}^n(X^n, \mathsf{c}^n(\mathsf{g}^n(X^n))) \tag{64}$$

$$Z_n = \mathsf{d}^n(X^n, b^n). \tag{65}$$

Then,

$$\begin{aligned}
&\mathbb{E}\left[V_n\right] \le \mathbb{E}\left[V_n 1\{V_n \le d+\gamma\}\right] \\
&+ \mathbb{E}\left[V_n 1\{d+\gamma < V_n \le \triangle_\gamma\}\right] + \mathbb{E}\left[V_n 1\{\triangle_\gamma < V_n\}\right] && (66)\\
&\le d+\gamma+\triangle_\gamma\mathbb{P}\left[1\{d+\gamma<V_n\}\right] + \mathbb{E}\left[V_n 1\{\triangle_\gamma < V_n\}\right] && (67)\\
&\le d+\gamma+\triangle_\gamma\mathbb{P}\left[1\{d+\gamma<W_n\}\right] + \mathbb{E}\left[Z_n 1\{\triangle_\gamma < Z_n\}\right] && (68)\\
&\le_n d+2\gamma+\mathbb{E}\left[Z_n 1\{\triangle_\gamma < Z_n\}\right] && (69)\\
&\le_n d+3\gamma && (70)
\end{aligned}$$

which gives the result. □

The following theorem is essentially shown in [14, Theorem 1].

Theorem 5 (Kieffer)**.** *Let $\mathbf{X}$ be an information source satisfying conditions (1)-(3) in Section III, with f being the identity. Let d^n be separable. Given an arbitrary sequence of (n, M_n, d, ϵ_n)-lossy source codes, if*

$$\lim_{n\to\infty}\frac{1}{n}\log M_n < \mathcal{R}(d) \tag{71}$$

then

$$\lim_{n\to\infty}\epsilon_n = 1. \tag{72}$$

An important implication of Theorem 5 for f-separable rate-distortion functions is given in the following lemma.

Lemma 6. *Let $\mathbf{X}$ be an information source satisfying conditions (1)-(3) in Section III. Then,*

$$\mathcal{R}_{\mathsf{f}}(d) \ge \tilde{\mathcal{R}}(\mathsf{f}(d)) \tag{73}$$

Proof. If $\tilde{\mathcal{R}}(\mathsf{f}(d)) = 0$, we are done. Suppose $\tilde{\mathcal{R}}(\mathsf{f}(d)) > 0$. Assume there exists a sequence $\{(\mathsf{g}_n, \mathsf{c}_n)\}_{n=1}^{\infty}$ of (n, M_n, d_n)-lossy source codes (under f-separable distortion) with

$$\limsup_{n\to\infty}\frac{1}{n}\log M_n < \tilde{\mathcal{R}}(\mathsf{f}(d)) \tag{74}$$

and

$$\limsup_{n\to\infty} d_n \le d. \tag{75}$$

Since $\tilde{\mathcal{R}}(\mathsf{f}(d))$ is continuous and decreasing, there exists some $\gamma > 0$ such that

$$\lim_{n\to\infty} \frac{1}{n} \log M_n < \tilde{\mathcal{R}}(\mathsf{f}(d+\gamma)) < \tilde{\mathcal{R}}(\mathsf{f}(d)). \tag{76}$$

For every n, the $(\mathsf{g}_n, \mathsf{c}_n)$ lossy source code is also an $(n, M_n, d+\gamma, \epsilon_n)$-lossy source code for some $\epsilon_n \in [0,1]$ and f-separable d^n. It is also an $(n, M_n, \mathsf{f}(d+\gamma), \epsilon_n)$-lossy source code with respect to separable distortion $\mathsf{d}^n(\cdot,\cdot)$. We can therefore apply Theorem 5 to obtain

$$\lim_{n\to\infty} \epsilon_n = 1. \tag{77}$$

Thus,

$$d_n \geq \mathbb{E}\left[\mathsf{d}^n(X^n, \mathsf{c}^n(\mathsf{g}^n(X^n)))\right] \geq \epsilon_n(d+\gamma) >_n d + \frac{\gamma}{2}. \tag{78}$$

The result follows since we obtained a contradiction with (75). □

B. Rate-distortion function for a mixed source

Lemma 7. *The rate-distortion function with respect to* f*-separable distortion for the mixture source (41) is given by*

$$\mathcal{R}_{\mathsf{f}}(d) = \min_{(d_1,d_2)\colon \lambda d_1 + (1-\lambda) d_2 \leq d} \max\left(\mathcal{R}^1_{\mathsf{f}}(d_1), \mathcal{R}^2_{\mathsf{f}}(d_2)\right) \tag{79}$$

where $\mathcal{R}^1_{\mathsf{f}}(d)$ *and* $\mathcal{R}^2_{\mathsf{f}}(d)$ *are the rate-distortion functions with respect to* f*-separable distortion for stationary memoryless source given by* P^1 *and* P^2*, respectively.*

Proof. Observe that,

$$M^*_{\mathsf{f}}(d) \geq \min_{(d_1,d_2)\in\mathcal{D}} \max\left(M^1_{\mathsf{f}}(d_1), M^2_{\mathsf{f}}(d_2)\right) \tag{80}$$

$$M^*_{\mathsf{f}}(d) \leq \min_{(d_1,d_2)\in\mathcal{D}} \max\left(2M^1_{\mathsf{f}}(d_1), 2M^2_{\mathsf{f}}(d_2)\right) \tag{81}$$

where

$$\mathcal{D} = \left\{(d_1,d_2)\colon \lambda d_1 + (1-\lambda) d_2 \leq d\right\}, \tag{82}$$

$M^1_{\mathsf{f}}(d_1)$ and $M^*_{2,\mathsf{f}}(d_1)$ are the non-asymptotic limits for P^1 and P^2, respectively. Indeed, the upper bound follows by designing optimal codes for P_1 and P_2 separately, and then combining them to give

$$M^*_{\mathsf{f}}(d) \leq \min_{(d_1,d_2)\in\mathcal{D}} \left(M^1_{\mathsf{f}}(d_1) + M^2_{\mathsf{f}}(d_2)\right) \tag{83}$$

$$\leq \min_{(d_1,d_2)\in\mathcal{D}} \max\left(2M^1_{\mathsf{f}}(d_1), 2M^2_{\mathsf{f}}(d_2)\right). \tag{84}$$

The lower bound follows by the following argument. Fix an (M,d)-lossy source code (f-separable distortion), (g,c). Define

$$d_1 = \mathbb{E}\left[\mathsf{d}^n(X^n, \mathsf{c}(\mathsf{g}(X^n)))|Z=0\right], \tag{85}$$

$$d_2 = \mathbb{E}\left[\mathsf{d}^n(X^n, \mathsf{c}(\mathsf{g}(X^n)))|Z=1\right]. \tag{86}$$

Clearly, $(d_1,d_2) \in \mathcal{D}$. It also follows that

$$M \geq M^1_{\mathsf{f}}(d_1) \tag{87}$$

since (g,c) is an (M,d_1)-lossy source code (f-separable distortion) code for X_1^n. Likewise,

$$M \geq M^2_{\mathsf{f}}(d_2) \tag{88}$$

which proves the lower bound. The result follows directly from (81). □

REFERENCES

[1] N. Sloane and A. Wyner, "Coding theorems for a discrete source with a fidelity criterion institute of radio engineers, international convention record, vol. 7, 1959." in *Claude E. Shannon:Collected Papers*. Wiley-IEEE Press, 1993, pp. 325–350. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5311476

[2] V. Tikhomirov, "On the notion of mean," in *Selected Works of A. N. Kolmogorov*, ser. Mathematics and Its Applications (Soviet Series), V. Tikhomirov, Ed. Springer Netherlands, 1991, vol. 25, pp. 144–146. [Online]. Available: http://dx.doi.org/10.1007/978-94-011-3030-1_17

[3] A. Rényi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. University of California Press, 1961, pp. 547–561.

[4] L. Campbell, "A coding theorem and Rényi's entropy," *Information and Control*, vol. 8, no. 4, pp. 423 – 429, 1965. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0019995865903323

[5] ——, "Definition of entropy by means of a coding problem," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 6, no. 2, pp. 113–118, 1966. [Online]. Available: http://dx.doi.org/10.1007/BF00537132

[6] J. Kieffer, "Variable-length source coding with a cost depending only on the code word length," *Information and Control*, vol. 41, no. 2, pp. 136 – 146, 1979. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0019995879905217

[7] E. Arikan, "An inequality on guessing and its application to sequential decoding," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 99–105, Jan 1996.

[8] C. Bunte and A. Lapidoth, "Encoding tasks and Rényi entropy," *IEEE Transactions on Information Theory*, vol. 60, no. 9, pp. 5065–5076, Sept 2014.

[9] E. Arikan and N. Merhav, "Guessing subject to distortion," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1041–1056, May 1998.

[10] R. M. Gray, *Entropy and Information Theory*, 2nd ed. Springer, 2011.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, 2006.

[12] T. S. Han, *Information-Spectrum Methods in Information Theory*. Springer Berlin Heidelberg, Feb 2003.

[13] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 63–86, Jan 1996.

[14] J. Kieffer, "Strong converses in source coding relative to a fidelity criterion," *IEEE Transactions on Information Theory*, vol. 37, no. 2, pp. 257–262, Mar 1991.

[15] S. Verdú, "ELE528: Information theory lecture notes," 2015, Princeton University.