A single-shot approach to lossy source coding under logarithmic loss

Yanina Shkel *Member, IEEE*, and Sergio Verdú *Fellow, IEEE*

***Abstract*—This paper considers the problem of lossy source coding with a specific distortion measure: logarithmic loss. The focus of this paper is on the single-shot approach which exposes crisply the connection between lossless source coding with list decoding and lossy source coding with log-loss. Fixed-length and variable length bounds are presented. Fixed-length bounds include the single-shot fundamental limit for average as well as excess distortion. Variable-length bounds include the single-shot fundamental limit for average as well as excess length. Two multi-terminal problems are addressed: coding with side information (Wyner-Ziv), and multiple descriptions coding. In both cases, the application of the Shannon-McMillan Theorem to the single-shot bounds yields the rate-distortion function and the rate distortion-region for stationary ergodic sources.**

Keywords: Lossy data compression; single-shot approach; logarithmic loss; Shannon theory; rate-distortion function; Wyner-Ziv coding; multiple-descriptions coding.

I. Introduction

We study the problem of lossy source coding with a specific distortion measure: logarithmic loss (log-loss). A widely used loss function in learning theory, the log-loss distortion measure was introduced in the context of rate-distortion theory by Courtade et al. in [1], [2]. As pointed out in [2], it is a natural distortion measure in settings where the reconstructions are allowed to be *soft*, i.e. the decompressor outputs a distribution, rather than a distorted sample path.

Unlike [1], [2], our focus is on the single-shot approach to source coding with log-loss. In this setting, the source to be compressed is a random variable X with either finite or countably infinite alphabet $\mathcal{X}$. The reconstruction alphabet is $\mathcal{P}(\mathcal{X})$: the set of all probability mass functions on $\mathcal{X}$.

Definition 1. *The log-loss distortion between $x \in \mathcal{X}$ and its reconstruction $\hat{P} \in \mathcal{P}(\mathcal{X})$ is given by*[1]

$$\mathsf{d}(x, \hat{P}) = \log \frac{1}{\hat{P}(x)}. \tag{1}$$

As usual, the single-shot setting readily results in the asymptotic fundamental limits for stationary ergodic sources for various settings such as point-to-point, Wyner-Ziv, and multiple description coding. To that end, it is useful to comment on the problem formulation in the n-letter setting. The approach we adopt is to consider the source X^n over $\mathcal{X}^n$ with the reconstruction alphabet $\mathcal{P}(\mathcal{X}^n)$, which contains all measures over $\mathcal{X}^n$. The n-letter distortion measure is then defined as the direct generalization of (1):

$$\mathsf{d}_n(x^n, \hat{P}^n) = \frac{1}{n}\mathsf{d}(x^n, \hat{P}^n(x^n)) = \frac{1}{n} \log \frac{1}{\hat{P}^n(x^n)}. \tag{2}$$

Alternatively, following the standard approach in information theory, [2] considers the source X^n over $\mathcal{X}^n$ with a restricted reproduction alphabet which is the n-fold Cartesian product of $\mathcal{P}(\mathcal{X})$; therefore, only the product measures, $\hat{P}^n(x^n) = \hat{P}_1(x_1) \times \cdots \times \hat{P}_n(x_n)$, are considered as possible decompressor outputs.

The approach in (2) is the natural n-shot extension for the single-shot setting in which there is no notion of blocklength. While it is shown in [2] that the asymptotic limit for memoryless stationary sources is the same for both approaches, this need not hold in the non-asymptotic setting or when the source has memory. In general, (2) allows for simple and intuitive bounds, as well as better performance than the standard approach. Moreover, one of the more compelling reasons to study source coding with logarithmic loss is its popularity as a loss function in learning theory. The setting given by (2) is the one used there, see for example [3], and thus is of greater interest to us.

Finally we remark that even the most general asymptotic results in the rate-distortion literature assume that the output alphabet at blocklength n is an n-fold product of the single-letter alphabet. Therefore, the log-loss distortion measure falls outside the standard paradigm when the output alphabet is taken to be $\mathcal{P}(\mathcal{X}^n)$. The new single-shot achievability and converse bounds, tailored specifically for log-loss, let us immediately obtain the asymptotic fundamental limits which do apply to the setting in (2).

The rest of this paper is structured as follows. Section II introduces notation and a basic auxiliary result. In Sections III and IV, we present point-to-point bounds for fixed-length and variable-length coding, respectively. This includes average and excess distortion point-to-point bounds for fixed-length coding, as well as average and excess length bounds for variable-length coding. In Section V, we provide non-asymptotic bounds for source coding with side information and the multiple descriptions problem. Section VI applies the Shannon-McMillan Theorem to the single-shot bounds in order to obtain the rate-distortion function and the rate distortion-region for stationary ergodic sources in each of the paradigms considered in the paper.

This work was supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

This paper was presented in part at the 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain.

[1]Throughout the paper all logarithms and exponent functions have an arbitrary common base.

II. Preliminaries

Let X be a random variable defined on a finite or countably infinite alphabet $\mathcal{X}$ with distribution P_X. Given a random variable Y jointly distributed with X we define information and conditional information as,

$$\imath_X(a) = \log \frac{1}{P_X(a)}, \tag{3}$$

and

$$\imath_{X|Y}(a|b) = \log \frac{1}{P_{X|Y}(a|b)}, \tag{4}$$

respectively. The entropy and conditional entropy are given by,

$$H(X) = \mathbb{E}\left[\imath_X(X)\right], \tag{5}$$

and

$$H(X|Y) = \mathbb{E}\left[\imath_{X|Y}(X|Y)\right], \tag{6}$$

respectively.

Using (3), the log-loss distortion (1) can be written as

$$\mathsf{d}(x, \hat{P}) = \imath_{\hat{P}}(x). \tag{7}$$

Thus, we can interpret the log-loss distortion measure as the remaining uncertainty about x given $\hat{P}$[2]. Moreover, when X^n is a stationary and memoryless source, the log-loss rate-distortion function is known to be [1], [2]

$$R(d) = H(P) - d. \tag{8}$$

This kind of linear relationship between rate and distortion appears in many asymptotic and non-asymptotic fundamental limits of lossy compression with log-loss.

Given a general distortion measure $\mathsf{d} : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$ we say that $\hat{x} \in \hat{\mathcal{X}}$ *d-covers* $a \in \mathcal{X}$ if $\mathsf{d}(a, \hat{x}) \le d$. For the log-loss distortion measure, P d-covers a whenever

$$P(a) \ge \exp(-d). \tag{9}$$

The following simple lemma is key to characterizing the fundamental limits in Sections III and IV, as well as drawing connections between these limits and list decoding.

Lemma 1. *Let a distribution P on $\mathcal{X}$ be given and define $\mathcal{S} = \{a \in \mathcal{X} : P \text{ } d\text{-covers } a\}$. Then, for any $P \in \mathcal{P}(\mathcal{X})$,*

$$|\mathcal{S}| \le \lfloor \exp(d) \rfloor. \tag{10}$$

Proof. The result follows from

$$1 \ge \sum_{a \in \mathcal{S}} P(a) \ge \sum_{a \in \mathcal{S}} \exp(-d) = |\mathcal{S}| \exp(-d) \tag{11}$$

where the second inequality follows from (9). □

The converse of Lemma 1 holds as well. Given any set $\mathcal{S} \in \mathcal{X}$ such that $|\mathcal{S}| \le \lfloor \exp(d) \rfloor$ it is possible to produce P which d-covers $\mathcal{S}$.

Fix an integer $L > 0$, $P_X \in \mathcal{P}(\mathcal{X})$, and label without loss of generality $\mathcal{X} = \{1, \dots, |\mathcal{X}|\}$ (where we are allowing $|\mathcal{X}| = \infty$), such that $P_X(a) \ge P_X(b)$ for $a \le b$. The fundamental limits given in Sections III and IV are closely related to the distribution of $\lceil \frac{X}{L} \rceil$, which is supported on $\left\{1, 2, \dots, \left\lceil \frac{|\mathcal{X}|}{L} \right\rceil\right\}$ with the probability mass function

$$P_{\lceil \frac{X}{L} \rceil}(a) = \begin{cases} \sum_{b=(a-1)L+1}^{aL} P_X(b), & a \le \left\lceil \frac{|\mathcal{X}|}{L} \right\rceil - 1, \\ \sum_{b=(a-1)L+1}^{|\mathcal{X}|} P_X(b), & a = \left\lceil \frac{|\mathcal{X}|}{L} \right\rceil \end{cases}. \tag{12}$$

Example 1. A fair coin is tossed until it shows heads. Let X be the number of tosses of the coin. Then,

$$P_X(a) = \frac{1}{2^a}, a \in \{1, 2, \dots\} \tag{13}$$

and

$$P_{\lceil \frac{X}{L} \rceil}(a) = \frac{2^L - 1}{2^{aL}}, a \in \{1, 2, \dots\}. \tag{14}$$

III. Point-to-point: fixed-length coding

We begin with bounds for the average distortion single-shot fundamental limit which turns out to be the only fundamental limit studied in Sections III and IV without a direct connection to list-decoding and the distribution of $\lceil \frac{X}{L} \rceil$. Next, we use Lemma 1 to characterize the excess distortion fundamental limit of X and show that it coincides with the error probability for almost lossless compression of $\lceil \frac{X}{L} \rceil$, where $L = \lfloor \exp(d) \rfloor$. We conclude this section by comparing the excess distortion fundamental limit for log-loss with general purpose excess distortion bounds specialized to log-loss.

A. Fundamental limits

For a positive integer M denote, $\mathcal{M} = \{1, \dots, M\}$. A fixed-length lossy source code (f, c) of size M is a pair of mappings,

$$\mathsf{f} \colon \mathcal{X} \to \mathcal{M}, \quad \mathsf{c} \colon \mathcal{M} \to \mathcal{P}(\mathcal{X}). \tag{15}$$

A lossy source-code (f, c) is an (M, d)-lossy source code if

$$\mathbb{E}\left[\mathsf{d}(X, \mathsf{c}(\mathsf{f}(X)))\right] \le d, \text{ and } |\mathcal{M}| \le M. \tag{16}$$

It is an (M, d, ϵ)-lossy source code if

$$\mathbb{P}\left[\mathsf{d}(X, \mathsf{c}(\mathsf{f}(X))) > d\right] \le \epsilon \text{ and } |\mathcal{M}| \le M. \tag{17}$$

The single-shot fundamental limits are defined as,

$$d_{\mathrm{f}}^*(M) = \inf\{d \colon \exists (M, d)\text{-lossy source code}\}, \tag{18}$$

$$\epsilon_{\mathrm{f}}^*(M, d) = \inf\{\epsilon \colon \exists (M, d, \epsilon)\text{-lossy source code}\}. \tag{19}$$

B. Bounds for average distortion

The next lemma is a single-shot version of [2, Lemma 1]. We provide the proof here for completeness.

Lemma 2. *Fix P_{XU} on $\mathcal{X} \times \mathcal{M}$ and a decompressor $\mathsf{c} : \mathcal{M} \to \mathcal{P}(\mathcal{X})$. Then,*

$$H(X|U) \le \mathbb{E}\left[\mathsf{d}(X, \mathsf{c}(U))\right] \tag{20}$$

with equality if $\mathsf{c}(u) = P_{X|U=u}$ for all u such that $P_U(u) > 0$.

Proof. Fix arbitrary $u \in \mathcal{U}$. Then, letting $\hat{P}_u = \mathsf{c}(u)$,

$$H(X|U = u) = \mathbb{E}\left[\imath_{X|U}(X|u)\right] \tag{21}$$

[2]Log-loss is also known as the "self-information loss" in literature.

$$= \sum_{x\in\mathcal{X}} P_{X|U=u}(x)\log\frac{1}{\hat{P}_u(x)} - \sum_{x\in\mathcal{X}} P_{X|U=u}(x)\log\frac{P_{X|U=u}(x)}{\hat{P}_u(x)} \quad (22)$$

$$= \sum_{x\in\mathcal{X}} P_{X|U=u}(x)\log\frac{1}{\hat{P}_u(x)} - D(P_{X|U=u}\|\hat{P}_u). \quad (23)$$

Equation (20) follows from the non-negativity of conditional relative entropy, which is zero if and only if $\mathsf{c}(\cdot)$ is as specified in the statement of the result. □

Lemma 2 immediately yields the following single-shot converse for the fundamental limit with average distortion.

Theorem 3 (Average distortion converse)**.**

$$d_{\mathrm{f}}^*(M) \geq [H(X) - \log M]^+ \quad (24)$$

Proof. Fix an arbitrary (M, d)-lossy source code and let $U = \mathsf{f}(X)$ be the compressor output. Then,

$$H(X) - d \leq H(X) - H(X|U) \quad (25)$$
$$= I(X;U) \quad (26)$$
$$\leq \log M \quad (27)$$

where (25) follows from Lemma 2, (27) results from the fact that U takes at most M values. □

The next example shows that the two n-shot settings introduced in Section I do generally yield different non-asymptotic fundamental limits.

Example 2. Let X^n be n samples of a Markov process on $\{0,1\}$ defined by

$$P_{X_1}(0) = P_{X_1}(1) = \frac{1}{2} \quad (28)$$

and

$$P_{X_i|X_{i-1}}(0|1) = P_{X_i|X_{i-1}}(1|0) = \epsilon \quad (29)$$
$$P_{X_i|X_{i-1}}(0|0) = P_{X_i|X_{i-1}}(1|1) = 1-\epsilon \quad (30)$$

for $i \in \{2,\ldots,n\}$ with $\epsilon = 0.11$. Then,

$$d_{\mathrm{f}}^*(1) = H(X^n) = \frac{n+1}{2}. \quad (31)$$

We can similarly define $\tilde{d}_{\mathrm{f}}^*(M)$ to be the non-asymptotic fundamental limit for the case when the output alphabet is restricted to be a set of all product distributions. Applying Lemma 2 we obtain

$$\tilde{d}_{\mathrm{f}}^*(1) \geq \sum_1^n H(X_i) = n \quad (32)$$

and thus $\tilde{d}_{\mathrm{f}}^*(1) > d_{\mathrm{f}}^*(1)$ for $n > 1$.

Theorem 3 provides intuition about the structure of a lossy source code with good average distortion. Namely, equality holds in (24) if the following three conditions are satisfied, 1) the encoder is deterministic, 2) the compressor output is equiprobable, 3) the decoder maps each message to the posterior distribution of the source given the message. Theorem 4 uses this intuition to come up with an upper bound on $d_{\mathrm{f}}^*(M)$.

Theorem 4 (Average distortion achievability)**.** *If $|\mathcal{X}| \leq M$, then $d_{\mathrm{f}}^*(M) = 0$. Otherwise,*

$$d_{\mathrm{f}}^*(M) \quad (33)$$
$$\leq \mathbb{E}\left[1\left\{\imath_X(X) > \log M\right\}\log\left(1+\exp\left(\imath_X(X) - \log M\right)\right)\right]$$
$$\leq \mathbb{E}\left[1\left\{\imath_X(X) > \log M\right\}\left(\imath_X(X) - \log\frac{M}{2}\right)\right]. \quad (34)$$

Proof. Assume $|\mathcal{X}| \leq M$. Let, $\mathsf{f}(a) = a$ and $\mathsf{c}(a) = \delta_a$, where δ_a denotes the point mass at a. Then, for every $a \in \mathcal{X}$,

$$\mathsf{d}(a, \mathsf{c}(\mathsf{f}(a))) = \log\frac{1}{\delta_a(a)} = 0, \quad (35)$$

and the average distortion is zero. Assume $|\mathcal{X}| > M$. The result is proved by means of the following greedy construction.

Compressor: Assume without loss of generality that $\mathcal{X} = \{1,\ldots,|\mathcal{X}|\}$ and $P_X(a) \geq P_X(b)$ for $a \leq b$. The compressor is defined sequentially in $|\mathcal{X}|$ steps. At the first M steps, $\mathsf{f}(a) = a$. At steps $a = M+1,\ldots$ assign

$$\mathsf{f}(a) = \operatorname*{arg\,min}_{m\in\mathcal{M}} \sum_{b=1}^{a-1} P_X(b) 1\{\mathsf{f}(b) = m\}. \quad (36)$$

In other words, a is encoded to a message that has accrued the smallest total probability so far. In particular, if all the masses in P_X are different, then $\mathsf{f}(M+1) = \mathsf{f}(M)$. Furthermore, $\mathsf{f}(M+2) = \mathsf{f}(M)$ if $P_X(M) + P_X(M+1) < P_X(M-1)$; otherwise $\mathsf{f}(M+2) = \mathsf{f}(M-1)$. It is important to note that by construction, at every intermediate step there is at least one message with accumulated probability strictly less than $\frac{1}{M}$. At the end, this remains true unless all M messages are equiprobable.

Decompressor: The decompressor is defined by $\mathsf{c}(m) = \hat{P}_m$, where

$$\hat{P}_m(a) = \begin{cases} \frac{P_X(a)}{\mathbb{P}[\mathsf{f}(X)=m]}, & \mathsf{f}(a) = m \\ 0, & \text{otherwise} \end{cases} \quad (37)$$

Average distortion analysis: The exact average distortion of (f,c) is given by

$$\mathbb{E}\left[\mathsf{d}(X, \mathsf{c}(\mathsf{f}(X)))\right] = \sum_{a\in\mathcal{X}} P_X(a)\log\frac{\mathbb{P}[\mathsf{f}(X) = \mathsf{f}(a)]}{P_X(a)}. \quad (38)$$

To obtain an upper bound to (38), first suppose that $P_X(a) \geq \frac{1}{M}$; Then a is the only element assigned to the message $\mathsf{f}(a)$ and $\mathsf{d}(a, \mathsf{f}(a)) = 0$. Otherwise, let $b \in \mathcal{X}$ be the last element in $\mathcal{X}$ assigned to $m = \mathsf{f}(a)$. Then,

$$\mathbb{P}[\mathsf{f}(X) = m] < \frac{1}{M} + P_X(b) \leq \frac{1}{M} + P_X(a) < \frac{2}{M} \quad (39)$$

since $\mathbb{P}[X \in \{\mathsf{f}^{-1}(m) \setminus \{b\}\}] < \frac{1}{M}$ and $P_X(b) \leq P_X(a)$. Therefore, for all a such that $P_X(a) < \frac{1}{M}$,

$$\mathsf{d}(a, \mathsf{f}(a)) = \log\frac{1}{\hat{P}_m(a)} \quad (40)$$
$$= \log\frac{\mathbb{P}[\mathsf{f}(X) = m]}{P_X(a)} \quad (41)$$

$$\leq \log\left(1+\exp\left(\imath_X(a)-\log M\right)\right) \tag{42}$$

$$\leq \imath_X(a)-\log\frac{M}{2} \tag{43}$$

and the result follows. □

Example 1 (continued)**.**

$$[2-\log M]^+ \leq d_{\mathrm{f}}^*(M) \leq 2^{-(M-2)}, \tag{44}$$

which implies $d_{\mathrm{f}}^*(2)=1$.

The lower bound in (44) is obtained from (24) and

$$H(X)=\sum_{k=1}^{\infty} k\,2^{-k}=2 \text{ bits}. \tag{45}$$

The upper bound in (44) is obtained from (38) by observing that for this source the proposed greedy algorithm assigns the first $M-1$ masses to the first $M-1$ codewords, while the rest of the masses are assigned to the Mth codeword. Thus,

$$\mathbb{P}[\mathsf{f}(X)=m]=\begin{cases} 2^{-m}, & m<M \\ 2^{-(M-1)}, & m=M \end{cases} \tag{46}$$

and

$$\sum_{a\in\mathcal{X}} P_X(a)\log\frac{\mathbb{P}[\mathsf{f}(X)=\mathsf{f}(a)]}{P_X(a)}=\sum_{k=1}^{\infty}2^{-k}\log\frac{\mathbb{P}[\mathsf{f}(X)=\mathsf{f}(a)]}{2^{-k}} \tag{47}$$

$$=\sum_{k=1}^{M-1}2^{-k}\log\frac{2^{-k}}{2^{-k}}+\sum_{k=M}^{\infty}2^{-k}\log\frac{2^{-(M-1)}}{2^{-k}} \tag{48}$$

$$=\sum_{k=M}^{\infty}2^{-k}(k-(M-1)) \tag{49}$$

$$=2^{-(M-2)} \tag{50}$$

where (50) follows from

$$\sum_{k=1}^{n} k\,2^{-k}=2-2^{-n}(2+n). \tag{51}$$

Example 3. Let X is the number of tails obtained in 100 biased coin flips. That is, X is a Binomial(n,p) source with $n=100$ and $p=0.1$. Figure 1 compares the converse bound (24) with the achievability bounds (33), (34) and (38) for X. As can be seen in Figure 1, (24) and (38) give tight lower and upper bounds on $d_{\mathrm{f}}^*(M)$.

Examples 1 and 3 demonstrate that Theorems 3 and 4 give excellent bounds on $d_{\mathrm{f}}^*(M)$. Finding a code which achieves $d_{\mathrm{f}}^*(M)$ for an arbitrary source X is a difficult combinatorial problem. For example, as [4] points out, when $M=2$ this problem is equivalent to an NP-Complete problem known as Subset Sum. In contrast, as we see next, finding the code which achieves $\epsilon_{\mathrm{f}}^*(M,d)$ is straightforward.

C. Bounds for excess distortion

Theorem 5 (Excess distortion converse)**.** *Assume without loss of generality that $\mathcal{X}=\{1,\ldots,|\mathcal{X}|\}$ and $P_X(a)\geq P_X(b)$ for $a\leq b$. Then,*

$$\epsilon_{\mathrm{f}}^*(M,d)\geq \mathbb{P}\left[X>M\lfloor\exp(d)\rfloor\right]. \tag{52}$$

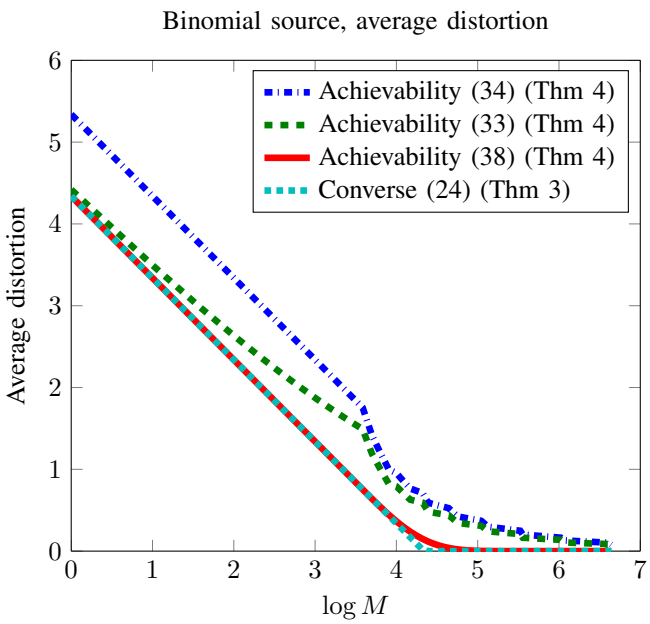

Fig. 1. Bounds on the fundamental trade-off with average distortion $d_{\mathrm{f}}^*(M)$ for the Binomial(n,p) source in Example 3.

Proof. Fix an arbitrary lossy source code (f,c) of size M. For $m\in\mathcal{M}$ let $\mathcal{S}_m$ be the set of all $a\in\mathcal{X}$ d-covered by $\mathsf{c}(m)$. Then,

$$|\cup_{m\in\mathcal{M}}\mathcal{S}_m|\leq M\lfloor\exp(d)\rfloor \tag{53}$$

follows by Lemma 1 and the union bound. Thus, at least $|\mathcal{X}|-M\lfloor\exp(d)\rfloor$ elements must be left uncovered by any code (f,c) of size M. The excess distortion probability is minimized when the uncovered elements are the $|\mathcal{X}|-M\lfloor\exp(d)\rfloor$ elements with lowest probability. That is, any uncovered $a\in\mathcal{X}$ is such that $a>M\lfloor\exp(d)\rfloor$. □

A looser lower bound on $\epsilon_{\mathrm{f}}^*(M,d)$ can be obtained by particularizing the d-tilted converse [5] to the log-loss distortion measure. Indeed, since the d-tilted information for log-loss is given by $\imath_X(x)-d$ we obtain

$$\epsilon_{\mathrm{f}}^*(M,d)\geq\sup_{\gamma>0}\left\{\mathbb{P}\left[\imath_X(X)>d+\log M+\gamma\right]-\exp(-\gamma)\right\}. \tag{54}$$

Definition 2. *Fix P_X on $\mathcal{X}$ and an integer $L\geq 0$. Define a code $(\mathsf{f}_L^*,\mathsf{c}_L^*)$ with M messages by*

$$\mathsf{f}_L^*(a)=\min\left\{M,\left\lceil\frac{a}{L}\right\rceil\right\} \tag{55}$$

and $\mathsf{c}_L^*(m)=\hat{P}_m$ where

$$\hat{P}_m(a)=\begin{cases}\frac{1}{L}, & (m-1)L+1\leq a\leq mL \\ 0, & \textit{otherwise}\end{cases} \tag{56}$$

Note that, $\hat{P}_{\mathsf{f}_L^*(a)}(a)=\frac{1}{L}$ if a is one of the ML most likely elements of $\mathcal{X}$, and zero otherwise.

Theorem 6 (Optimality of $(\mathsf{f}_L^*,\mathsf{c}_L^*)$)**.** *Let $L=\lfloor\exp(d)\rfloor$. Then,*

$$\epsilon_{\mathrm{f}}^*(M,d)=\mathbb{P}\left[\mathsf{d}(X,\mathsf{c}_L^*(\mathsf{f}_L^*(X)))>d\right] \tag{57}$$

$$= \mathbb{P}\left[X > M\lfloor \exp(d)\rfloor\right] \tag{58}$$

$$\le \mathbb{P}\left[\imath_X(X) > \log M + \lfloor d\rfloor\right]. \tag{59}$$

Proof. The code $(\mathsf{f}_L^*, \mathsf{c}_L^*)$ with $L = \lfloor \exp(d)\rfloor$ is exactly the code which covers the $M\lfloor \exp(d)\rfloor$ most likely elements in $\mathcal{X}$, which is the condition for equality in (52). Thus, it achieves the single-shot fundamental limit for excess distortion. To see (59),

$$\mathbb{P}\left[X > M\lfloor \exp(d)\rfloor\right] \le \mathbb{P}\left[\imath_X(X) > \log M + \log\lfloor \exp(d)\rfloor\right] \tag{60}$$

$$\le \mathbb{P}\left[\imath_X(X) > \log M + \lfloor d\rfloor\right] \tag{61}$$

where (61) follows since $\lfloor \exp(d)\rfloor \ge \lfloor \exp(\lfloor d\rfloor)\rfloor = \exp(\lfloor d\rfloor)$. □

Example 1 (continued)**.**

$$\epsilon_{\mathrm{f}}^*(M,d) = \mathbb{P}\left[X > M\lfloor \exp(d)\rfloor\right] \tag{62}$$

$$= 2^{-(M\lfloor \exp(d)\rfloor - 1)} \tag{63}$$

Note that the code $(\mathsf{f}_L^*, \mathsf{g}_L^*)$ can be viewed as a list-decoding version of the conventional lossless compression setting. The single-shot fundamental limit given in Theorem 6 is intimately connected to the single-shot fundamental limits for variable-length strictly lossless source coding, and fixed-length almost lossless source coding, whose fundamental limits are also expressed in terms of $\mathbb{P}[X > k]$ as a function of k. As we will see in Section IV, the same observation holds for lossy variable-length compression with log-loss for probability of excess length.

Specializing Theorem 16 in Section V gives the following pleasing counterpart to the d-tilted converse (54)

$$\epsilon_{\mathrm{f}}^*(M,d) \le \inf_{\gamma>0}\left\{\mathbb{P}\left[\imath_X(X) > d + \log M - \gamma\right] + 2\exp(-\gamma)\right\}. \tag{64}$$

Equation (64) is looser than Theorem 6 and, unlike the d-tilted converse, is not known to hold for general distortion measures.

D. Penalty for random coding

Given a general distortion measure $\mathsf{d} : \mathcal{X}\times\hat{\mathcal{X}} \to [0,\infty)$ we define

$$\mathcal{B}_d(x) = \{\hat{x}\in\hat{\mathcal{X}} : \mathsf{d}(x,\hat{x}) \le d\} \tag{65}$$

to be a d-ball around $x\in\mathcal{X}$. Note that for log-loss distortion $\hat{\mathcal{X}} = \mathcal{P}(\mathcal{X})$ and

$$\mathcal{B}_d(x) = \{\hat{P}\in\mathcal{P} : \hat{P}(x) \ge \exp(-d)\}. \tag{66}$$

The exact performance of random coding for a general distortion measure, derived in [5], is given by

$$\epsilon_{\mathrm{f}}^*(M,d) \le \inf_{P_{\hat{X}}} \mathbb{E}\left[\left(1 - P_{\hat{X}}(\mathcal{B}_d(X))\right)^M\right], \tag{67}$$

while the likelihood encoder bound, derived in [6], is

$$\epsilon_{\mathrm{f}}^*(M,d) \le 1 - M\sup_{P_{\hat{X}|X}} \mathbb{E}\left[\frac{1\{\mathsf{d}(X,\hat{X})\le d\}}{\exp\left(\imath_{X;\hat{X}}(X;\hat{X})\right) + M - 1}\right]. \tag{68}$$

We now specialize (67) and (68) for log-loss by considering a source defined on a finite or countably infinite alphabet $\mathcal{X}$. Note that for log-loss distortion $P_{\hat{X}}(\cdot)$ and $P_{\hat{X}|X}(\cdot|x)$ are probability measures over $\mathcal{P}(\mathcal{X})$. For $M = 1$ equations (67) and (68) recover the results given by Theorem 6. This is not surprising; it is possible to pick an optimizing distribution that selects the single optimal codeword with probability one for both bounds, as we see next.

Lemma 7. *The random coding and likelihood encoder bounds are optimal for $M = 1$. That is, for any $d \ge 0$*

$$\epsilon_{\mathrm{f}}^*(1,d) = 1 - \inf_{P_{\hat{X}}} \mathbb{E}\left[P_{\hat{X}}(\mathcal{B}_d(X))\right] \tag{69}$$

$$= 1 - \sup_{P_{\hat{X}|X}} \mathbb{E}\left[\frac{1\{\mathsf{d}(X,\hat{X})\le d\}}{\exp\left(\imath_{X;\hat{X}}(X;\hat{X})\right)}\right]. \tag{70}$$

Proof. Define

$$\hat{P}(a) = \begin{cases} \frac{1}{\lfloor \exp(d)\rfloor}, & a \le \lfloor \exp(d)\rfloor \\ 0, & \text{otherwise}\end{cases} \tag{71}$$

Then $P_{\hat{X}}(\hat{P}) = 1$ optimizes (69) and $P_{\hat{X}|X}(\hat{P}|x) = 1, \forall x\in\mathcal{X}$ optimizes (70). □

For $M \ge 2$, specializing (67) and (68) to the log-loss distortion measure and optimizing over $P_{\hat{X}}$ and $P_{\hat{X}|X}$, respectively, yields worse achievability bounds than the one given by Theorem 6. We show this by deriving lower bounds on (67) and (68) and showing that these bounds are still worse than the exact value of $\epsilon_{\mathrm{f}}^*(M,d)$ given by Theorem 6.

Optimization Problem 1. Let integers $L \ge 1, M \ge 1, k \in \{2,3,\ldots,\infty\}$ and reals $a_1 \ge a_2 \ge \cdots \ge a_k \in (0,\infty)$ be given. Let

$$\mathsf{g}_1(\mathbf{t}) = \sum_{i=1}^k a_i(1-t_i)^M. \tag{72}$$

Then, the problem is to find $\min \mathsf{g}_1(\mathbf{t})$ subject to

$$\sum_{i=1}^k t_i \le L, \text{ and } 0 \le t_i \le 1. \tag{73}$$

The solution to this problem for $k < \infty$ is given in Lemma 24 in Appendix A. The next lemma, proved in Appendix A, provides lower bounds on the random coding bound for log-loss in terms of Optimization Problem 1.

Lemma 8. *Assume without loss of generality that $\mathcal{X} = \{1,\ldots,|\mathcal{X}|\}$ and $P_X(i) \ge P_X(j)$ for $i \le j$. Let $\mathsf{g}_1(\mathbf{t}^*)$ be the solution to Optimization Problem 1 with $k \leftarrow |\mathcal{X}|$, $a_i \leftarrow P_X(i)$, and $L \leftarrow \lfloor \exp(d)\rfloor$. Then,*

$$\inf_{P_{\hat{X}}} \mathbb{E}\left[\left(1 - P_{\hat{X}}(\mathcal{B}_d(X))\right)^M\right] \ge \mathsf{g}_1(\mathbf{t}^*) \tag{74}$$

with equality if $0 \le d < 1$ or $M = 1$.

For $L = 1$ Optimization Problem 1 matches (67) for the specified regimes. For $L > 1$ Optimization Problem 1 is a relaxation of the problem given by (67).

Optimization Problem 2. Let integers $M \geq 1, k \in \{2,3,\ldots,\infty\}$ and reals $a_1 \geq a_2 \geq \cdots \geq a_k \in (0,\infty)$ be given. Let

$$\mathsf{g}_2(\mathbf{t}) = \sum_{i=1}^{k} \min(a_i, t_i) \frac{1}{\frac{\min(a_i,t_i)}{a_i t_i} + M - 1}. \tag{75}$$

Then the problem is to find $\max \mathsf{g}_2(\mathbf{t})$ subject to

$$\sum_{i=1}^{k} t_i = 1, \quad \text{and } t_i \geq 0. \tag{76}$$

Optimization Problem 3. Let integers $M \geq 1, k \in \{2,3,\ldots,\infty\}$ and reals $a_1 \geq a_2 \geq \cdots \geq a_k \in (0,\infty)$ be given. Let

$$\mathsf{g}_3(\mathbf{t}) = \sum_{i=1}^{k} \frac{a_i t_i}{1 + a_i(M-1)}. \tag{77}$$

Then the problem is to find $\max \mathsf{g}_3(\mathbf{t})$ such that

$$\sum_{i=1}^{k} t_i = 1, \text{ and } t_i \geq 0. \tag{78}$$

Optimization Problem 3 is a relaxation of Optimization Problem 2 and can be immediately solved by inspection. The next lemma provides lower bounds on the likelihood encoder bound for log-loss in terms of Optimization Problems 2 and 3.

Lemma 9. *Assume without loss of generality that $\mathcal{X} = \{1,\ldots,|\mathcal{X}|\}$ and $P_X(i) \geq P_X(j)$ for $i \leq j$. Let $\mathsf{g}_2(\mathbf{t}^*)$ be the solution to Optimization Problem 2 and let $\mathsf{g}_3(\tilde{\mathbf{t}}^*)$ be the solution to Optimization Problem 3 with $k \leftarrow |\mathcal{X}|$, and $a_i \leftarrow P_X(i)$. Then,*

$$1 - M \sup_{P_{\hat{X}|X}} \mathbb{E}\left[\frac{1\{\mathsf{d}(X,\hat{X}) \leq d\}}{\exp\left(\imath_{X;\hat{X}}(X;\hat{X})\right) + M - 1}\right]$$
$$= 1 - M\mathsf{g}_2(\mathbf{t}^*) \tag{79}$$
$$\geq 1 - M\mathsf{g}_3(\tilde{\mathbf{t}}^*) = \frac{1 - P_X(1)}{1 + P_X(1)(M-1)} \tag{80}$$

for $0 \leq d < 1$ and $M \geq 2$.

We can now use Lemmas 8 and 9 to numerically compare the gap between general purpose bounds and special purpose bounds, as shown in Example 3.

Example 3 (continued)**.** Bounds on the fundamental limit with excess distortion $\epsilon_{\mathrm{f}}^*(M,d)$ for a Binomial(n,p) source with $n = 100$ and $p = 0.1$ are given in Figure 2. The converse (54) is compared to the exact fundamental limit given by Theorem 6, general random coding bound (67), and a lower bound on likelihood encoder bound (68) given by (80). While (58) could be computed for large values of $|\mathcal{X}|$, M and d, it is impractical to optimize (67) and (68) for anything beyond toy examples.

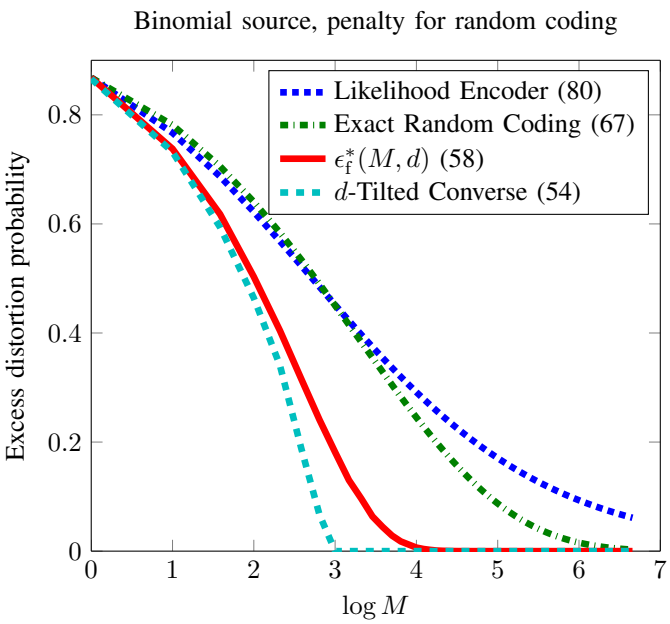

Fig. 2. Bounds on the fundamental limit with excess distortion $\epsilon_{\mathrm{f}}^*(M,d)$ for a Binomial(n,p) source and $0 \leq d < 1$ in Example 3.

IV. Point-to-point: variable-length coding

We begin by establishing an equivalence between $\epsilon_{\mathrm{f}}^*(M,d)$ and the probability of excess length for variable-length codes with and without prefix constraints. We then characterize the average length fundamental limits in terms of the distribution of $\lceil \frac{X}{L} \rceil$. We conclude this section by comparing the average length fundamental limit for log-loss with the general purpose average length bound specialized to log-loss.

A. Fundamental limits

A variable-length lossy source code (f,c) is a pair of mappings,

$$\mathsf{f}\colon \mathcal{X} \to \{0,1\}^*, \quad \mathsf{c}\colon \mathsf{f}(\mathcal{X}) \to \mathcal{P}(\mathcal{X}). \tag{81}$$

If (f,c) is such that no codeword in $\mathsf{f}(\mathcal{X})$ is a prefix of any another codeword in $\mathsf{f}(\mathcal{X})$ we call (f,c) *prefix free*.

Let $\ell(\mathsf{f}(x))$ denote the length of $\mathsf{f}(x)$. A variable-length lossy source code is an (l,d)-code if

$$\mathbb{E}\left[\ell(\mathsf{f}(X))\right] \leq l \text{ and } \mathsf{d}(x,\mathsf{c}(\mathsf{f}(x))) \leq d, \forall x \in \mathcal{X}. \tag{82}$$

A variable-length lossy source code is an (l,d,ϵ)-code if

$$\mathbb{P}\left[\ell(\mathsf{f}(X)) > l\right] \leq \epsilon \text{ and } \mathsf{d}(x,\mathsf{c}(\mathsf{f}(x))) \leq d, \forall x \in \mathcal{X}. \tag{83}$$

The single-shot fundamental limits for variable-length codes are defined as,

$$\ell_{\mathrm{v}}^*(d) = \inf\{l\colon \exists (l,d)\text{-code}\}, \tag{84}$$
$$\epsilon_{\mathrm{v}}^*(l,d) = \inf\{\epsilon\colon \exists (l,d,\epsilon)\text{-code}\}. \tag{85}$$

Analogously, the single-shot fundamental limits for prefix-free variable-length codes are defined as,

$$\ell_{\mathrm{p}}^*(d) = \inf\{l\colon \exists \text{ prefix-free } (l,d)\text{-code}\}, \tag{86}$$

$$\epsilon_{\mathrm{p}}^*(l,d) = \inf\{\epsilon\colon \exists \text{ prefix-free } (l,d,\epsilon)\text{-code}\}. \tag{87}$$

In [7] the authors observe that the non-asymptotic fundamental limits of fixed-length, variable-length, and prefix-free *lossless* source codes are intimately related. We make an analogous observation for lossy source codes in Theorem 10, which holds for *general* distortion measures. Our results are similar in spirit to those in [8, Theorem 3] where the asymptotic fundamental limits of fixed-length and prefix free codes are connected in the lossy setting.

In the special case of log-loss we know the exact fixed-length excess-distortion fundamental limit, $\epsilon_{\mathrm{f}}^*(M,d)$, from Theorem 6. Thus, Theorem 6 together with Theorem 10 yields the excess-length fundamental limits $\epsilon_{\mathrm{v}}^*(l,d)$ and $\epsilon_{\mathrm{p}}^*(l,d)$.

Theorem 10. *Let X be an arbitrary source with an arbitrary distortion measure $\mathsf{d}(\cdot,\cdot)$ and reconstruction alphabet $\hat{\mathcal{X}}$. Assume that X satisfies the following regularity condition (see [9]):*

- *For each $d>0$, there exists a finite or countably infinite subset $\{\hat{x}_i\}$ of $\hat{\mathcal{X}}$ and a 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*

$$\sum_i P_X(E_i)\log\frac{1}{P_X(E_i)} < \infty. \tag{88}$$

Then,

1)
$$\epsilon_{\mathrm{v}}^*(l,d) = \epsilon_{\mathrm{f}}^*(2^{l+1}-1,d), \tag{89}$$

2)
$$\epsilon_{\mathrm{p}}^*(l,d) = \begin{cases} \epsilon_{\mathrm{f}}^*(2^l-1,d), & l<\log M_d \\ 0, & l\ge \log M_d \end{cases} \tag{90}$$

where

$$M_d = \min\{M : \epsilon_{\mathrm{f}}^*(M,d)=0\} \in \{1,\dots,\infty\} \tag{91}$$

is the smallest number of elements in $\hat{\mathcal{X}}$ needed to d-cover $\mathcal{X}$.

The proof of Theorem 10 is given in Appendix B. For the log-loss distortion measure,

$$M_d = \left\lceil \frac{|\mathcal{X}|}{\lfloor \exp(d)\rfloor} \right\rceil \tag{92}$$

which follows from Theorem 6 since $\mathbb{P}\left[X > M\lfloor\exp(d)\rfloor\right]=0$ if and only if $|\mathcal{X}|\le M\lfloor\exp(d)\rfloor$. In light of Theorem 10, the bounds on $\epsilon_{\mathrm{f}}^*(M,d)$ derived in Section III can be leveraged to obtain bounds on $\epsilon_{\mathrm{v}}^*(l,d)$ and $\epsilon_{\mathrm{p}}^*(l,d)$. It is observed in [7] that the prefix condition on variable-length codes incurs a penalty of one bit when the performance metric of interest is probability of excess length. According to Theorem 10, this observation holds in the lossy setting as well.

Example 1 (continued)**.**

$$\epsilon_{\mathrm{v}}^*(l,d) = 2^{-((2^{l+1}-1)\lfloor\exp(d)\rfloor-1)}, \tag{93}$$
$$\epsilon_{\mathrm{p}}^*(l,d) = 2^{-((2^{l}-1)\lfloor\exp(d)\rfloor-1)} \tag{94}$$

B. Converses for average length

Bounds on the average code length for the log-loss distortion measure are provided next. We proceed to give an extension of Kraft's inequality to variable-length lossy source coding under log-loss, as well as a generalization of [7, Theorem 5].

Theorem 11. *For the log-loss distortion measure we have*

1) *For any variable-length code (f,c),*

$$\sum_{x\in\mathcal{X}} \exp\left(-\ell(\mathsf{f}(x)) - \mathsf{d}(x,\mathsf{c}(\mathsf{f}(x)))\right) \le \lfloor\log|\mathcal{X}|\rfloor + 1. \tag{95}$$

2) *Moreover, if (f,c) is prefix free, then*

$$\sum_{x\in\mathcal{X}} \exp\left(-\ell(\mathsf{f}(x)) - \mathsf{d}(x,\mathsf{c}(\mathsf{f}(x)))\right) \le 1. \tag{96}$$

Proof. 1)

$$\begin{aligned}
&\sum_{x\in\mathcal{X}} \exp\left(-\ell(\mathsf{f}(x)) - \mathsf{d}(x,\mathsf{c}(\mathsf{f}(x)))\right) \\
&= \sum_{m\in\mathsf{f}(\mathcal{X})}\ \sum_{x:\mathsf{f}(x)=m} \exp\left(-\ell(m) - \mathsf{d}(x,\mathsf{c}(m))\right) && (97)\\
&= \sum_{m\in\mathsf{f}(\mathcal{X})} \exp\left(-\ell(\mathsf{f}(m))\right) \sum_{x:\mathsf{f}(x)=m} \exp\left(-\mathsf{d}(x,\mathsf{c}(m))\right) && (98)\\
&= \sum_{m\in\mathsf{f}(\mathcal{X})} \exp\left(-\ell(\mathsf{f}(m))\right) \sum_{x:\mathsf{f}(x)=m} \hat{P}_m(x) && (99)\\
&= \sum_{m\in\mathsf{f}(\mathcal{X})} \exp(-\ell(\mathsf{f}(m))) && (100)\\
&\le \lfloor\log|\mathcal{X}|\rfloor + 1 && (101)
\end{aligned}$$

where (100) follows since $\hat{P}_m(\cdot)=\mathsf{c}(m)$ is a probability mass function and (101) follows from [7, Theorem 5].

2) To obtain (96) we follow the same steps up until (100) and then we apply Kraft's inequality since f is a prefix free code.

□

An immediate corollary of Theorem 11 is the following converse bound.

Theorem 12 (Average length converse for log-loss)**.** *For any source X,*

$$\ell_{\mathrm{v}}^*(d) \ge H(X) - d - \log\left(\lfloor\log|\mathcal{X}|\rfloor + 1\right), \tag{102}$$
$$\ell_{\mathrm{p}}^*(d) \ge H(X) - d. \tag{103}$$

Proof. Consider an arbitrary variable-length (l,d)-code. From Theorem 11 we obtain

$$\begin{aligned}
&\exp(-d)\sum_{x\in\mathcal{X}} \exp\left(-\ell(\mathsf{f}(x))\right) \\
&= \sum_{x\in\mathcal{X}} \exp\left(-\ell(\mathsf{f}(x)) - d\right) && (104)\\
&\le \sum_{x\in\mathcal{X}} \exp\left(-\ell(\mathsf{f}(x)) - \mathsf{d}(x,\mathsf{c}(\mathsf{f}(x)))\right) && (105)\\
&\le \left(\lfloor\log|\mathcal{X}|\rfloor + 1\right). && (106)
\end{aligned}$$

Let

$$Q_X(x) = \frac{\exp\left(-\ell(\mathsf{f}(x))\right)}{\sum_{x\in\mathcal{X}} \exp\left(-\ell(\mathsf{f}(x))\right)} \tag{107}$$

be a distribution on $\mathcal{X}$. Consider an arbitrary variable-length lossy source code with prefix constraints, (f,c). Then,

$$\mathbb{E}\left[\ell(\mathsf{f}(X))\right] - H(X) = \sum_{x\in\mathcal{X}} P_X(x) \log \frac{P_X(x)}{\exp(-\ell(\mathsf{f}(X)))} \tag{108}$$

$$= \sum_{x\in\mathcal{X}} P_X(x) \log \frac{P_X(x)}{Q(x)} - \log \sum_{x\in\mathcal{X}} \exp\left(-\ell(\mathsf{f}(x))\right) \tag{109}$$

$$\geq D(P_X\|Q_X) - (\lfloor \log|\mathcal{X}|\rfloor + 1) - d \tag{110}$$

$$\geq -d - \log\left(\lfloor \log|\mathcal{X}|\rfloor + 1\right) \tag{111}$$

where (109) follows from (107) and (110) follows from (106). Rearranging the terms gives (102). Equation (103) follows analogously. □

C. Achievabilities for average length

Next, we define a variable-length counterpart of the code given by Definition 2.

Definition 3. *Fix P_X on $\mathcal{X}$ and an integer $L \geq 0$. Define a variable-length code $(\mathsf{f}_L^\star, \mathsf{c}_L^\star)$ by*

$$\mathsf{f}_L^\star(a) = \begin{cases} \varnothing, & \left\lceil \frac{a}{L} \right\rceil = 1, \\ 0, & \left\lceil \frac{a}{L} \right\rceil = 2, \\ 1, & \left\lceil \frac{a}{L} \right\rceil = 3, \\ 00, & \left\lceil \frac{a}{L} \right\rceil = 4, \\ 01, & \left\lceil \frac{a}{L} \right\rceil = 5, \\ \dots & \end{cases} \tag{112}$$

and

$$\mathsf{c}_L^\star(\mathbf{s}) = \begin{cases} \hat{P}_1, & \mathbf{s} = \varnothing, \\ \hat{P}_2, & \mathbf{s} = 0, \\ \hat{P}_3, & \mathbf{s} = 1, \\ \hat{P}_4, & \mathbf{s} = 00, \\ \hat{P}_5, & \mathbf{s} = 01, \\ \dots & \end{cases} \tag{113}$$

where

$$\hat{P}_i(a) = \begin{cases} \frac{1}{L}, & (i-1)L + 1 \leq a \leq iL \\ 0, & \textit{otherwise} \end{cases} \tag{114}$$

The next result gives an exact expression for the minimum average length $\ell_{\mathrm{v}}^*(d)$.

Theorem 13 (Optimality of $(\mathsf{f}_L^\star, \mathsf{c}_L^\star)$)**.**

$$\ell_{\mathrm{v}}^*(d) = \mathbb{E}\left[\ell\left(\mathsf{f}^\star_{\lfloor\exp(d)\rfloor}(X)\right)\right] \tag{115}$$

$$= \sum_{k=1}^{\infty} \mathbb{P}\left[X \geq 2^k \lfloor\exp(d)\rfloor\right] \tag{116}$$

$$\leq H\left(\left\lceil \frac{X}{\lfloor\exp(d)\rfloor} \right\rceil\right) \tag{117}$$

where the distribution of $\left\lceil \frac{X}{\lfloor\exp(d)\rfloor} \right\rceil$ is given in (12).

Proof. The code $(\mathsf{f}^\star_{\lfloor\exp(d)\rfloor}, \mathsf{c}^\star_{\lfloor\exp(d)\rfloor})$ satisfies the condition

$$\mathsf{d}(x, \mathsf{c}^\star_{\lfloor\exp(d)\rfloor}(\mathsf{f}^\star_{\lfloor\exp(d)\rfloor}(x))) \leq d, \forall x \in \mathcal{X} \tag{118}$$

and so it is an (l,d)-code for some l. We see that $(\mathsf{f}^\star_{\lfloor\exp(d)\rfloor}, \mathsf{c}^\star_{\lfloor\exp(d)\rfloor})$ is optimal since no codeword can d-cover more than $\lfloor\exp(d)\rfloor$ elements and it assigns shortest strings to the most likely elements. We evaluate

$$\mathbb{E}\left[\ell\left(\mathsf{f}^\star_{\lfloor\exp(d)\rfloor}(X)\right)\right] = \sum_{k=1}^{\infty} k\,\mathbb{P}\left[2^{k+1} > \frac{X}{\lfloor\exp(d)\rfloor} \geq 2^k\right] \tag{119}$$

$$= \sum_{k=1}^{\infty} k\left(\mathbb{P}\left[\frac{X}{\lfloor\exp(d)\rfloor} \geq 2^k\right] - \mathbb{P}\left[\frac{X}{\lfloor\exp(d)\rfloor} \geq 2^{k+1}\right]\right) \tag{120}$$

$$= \sum_{k=1}^{\infty} \mathbb{P}\left[X \geq 2^k \lfloor\exp(d)\rfloor\right]. \tag{121}$$

Finally, we can view $(\mathsf{f}^\star_{\lfloor\exp(d)\rfloor}, \mathsf{c}^\star_{\lfloor\exp(d)\rfloor})$ as an optimal lossless variable-length code for the source $\left\lceil \frac{X}{\lfloor\exp(d)\rfloor} \right\rceil$. By [7, Theorem 2]

$$\ell\left(\mathsf{f}^\star_{\lfloor\exp(d)\rfloor}(a)\right) \leq \imath_{\left\lceil \frac{X}{\lfloor\exp(d)\rfloor} \right\rceil}(a). \tag{122}$$

Taking expectation of both sides of (122) yields (117). □

Remark 1. Letting $d = 0$ in (116) recovers the average length fundamental limit for lossless source coding derived in [7, Equation (98)].

Example 1 (continued)**.** In this case, the right hand side of (117) is

$$H\left(\left\lceil \frac{X}{L} \right\rceil\right) = \frac{L2^L}{2^L - 1} - \log\left(2^{L-1}\right). \tag{123}$$

We see, from (116), that $\ell_{\mathrm{v}}^*(d)$ is a piece-wise constant function with jumps at $d = \log_2 n, n \in \{1, 2, \dots\}$, at which arguments it is given by

$$\ell_{\mathrm{v}}^*(d) = \sum_{k=1}^{\infty} \mathbb{P}\left[X \geq 2^k 2^d\right] \tag{124}$$

$$= \sum_{k=1}^{\infty} 2^{-(2^{k+d}-1)} \tag{125}$$

$$= 2\sum_{k=d+1}^{\infty} 2^{-2^k}. \tag{126}$$

This yields $\ell_{\mathrm{v}}^*(0) = 0.632843\dots$.

Example 4. Let X be Poisson with parameter λ. The values of $\ell_{\mathrm{v}}^*(d)$ as a function of λ for $d \in \{0,1,2,3\}$ are given in Figures 3.

In the case of prefix-free codes and finite $\mathcal{X}$ the optimal code can be obtained via the Huffman algorithm.

Theorem 14 (Optimality of Huffman Codes)**.** *If $|\mathcal{X}| \leq \lfloor\exp(d)\rfloor$, then $\ell_{\mathrm{p}}^*(d) = 0$. Otherwise,*

$$\ell_{\mathrm{p}}^*(d) = \mathbb{E}\left[\ell\left(\mathrm{h}\left(\left\lceil \frac{X}{\lfloor\exp(d)\rfloor} \right\rceil\right)\right)\right] \tag{127}$$

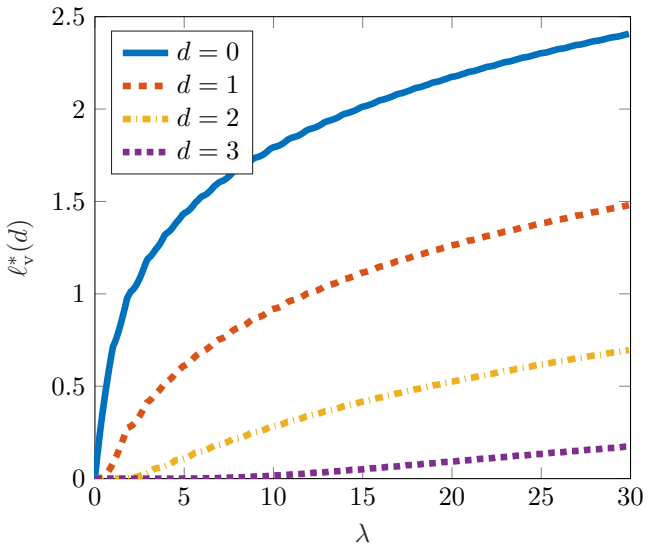

Fig. 3. Bounds on $\ell_{\mathrm{v}}^*(d)$ for the Poisson(λ) source in Example 4.

$$\leq H\left(\left\lceil \frac{X}{\lfloor \exp(d) \rfloor} \right\rceil\right) + 1 \tag{128}$$

where $\left\lceil \frac{X}{\lfloor \exp(d) \rfloor} \right\rceil$ *is given in (12) and* $\mathrm{h}(\cdot)$ *denotes the lossless Huffman code [10] for the source* $\left\lceil \frac{X}{\lfloor \exp(d) \rfloor} \right\rceil$.

Proof. Suppose $|\mathcal{X}| \leq \lfloor \exp(d) \rfloor$. Then the variable-length code $\left(\mathsf{f}^\star_{\lfloor \exp(d) \rfloor}, \mathsf{c}^\star_{\lfloor \exp(d) \rfloor}\right)$ is given by,

$$\mathsf{f}^\star_{\lfloor \exp(d) \rfloor}(x) = \varnothing, \forall x \in \mathcal{X} \tag{129}$$

and $\mathsf{c}(\varnothing) = \hat{P}_1(x)$ where

$$\hat{P}_1(x) = \frac{1}{|\mathcal{X}|}, \forall x \in \mathcal{X}. \tag{130}$$

The code $\left(\mathsf{f}^\star_{\lfloor \exp(d) \rfloor}, \mathsf{c}^\star_{\lfloor \exp(d) \rfloor}\right)$ is a prefix-free $(0, d)$-code and thus $\ell_{\mathrm{p}}^*(d) = 0$.

Suppose $|\mathcal{X}| > \lfloor \exp(d) \rfloor$. An optimal (l, d)-code for X must satisfy the following properties:

1) If $P_X(a) > P_X(b)$ then $\ell(\mathsf{f}(a)) < \ell(\mathsf{f}(b))$,
2) For any $m \in \mathsf{f}(\mathcal{X})$, $|\{a \in \mathcal{X} : \mathsf{f}(a) = m\}| \leq \lfloor \exp(d) \rfloor$,
3) Let $\ell_{\max} = \arg\max_{m \in \mathsf{f}(\mathcal{X})} \ell(m)$ be the maximum length of any string in $\mathsf{f}(\mathcal{X})$. Then

$$|\{a \in \mathcal{X} : \mathsf{f}(a) = m\}| = \lfloor \exp(d) \rfloor \tag{131}$$

holds for every $m \in \mathsf{f}(\mathcal{X})$ such that $\ell(m) < \ell_{\max}$.

Let $\mathsf{f}(\cdot)$ be an encoder satisfying these properties. We can convert $\mathsf{f}(\cdot)$ into a new code $\mathsf{f}'(\cdot)$ without changing the performance of this code by ensuring that the $\lfloor \exp(d) \rfloor$ most likely elements are assigned to the shortest available string, the next $\lfloor \exp(d) \rfloor$ most likely elements are assigned to the next shortest string, and so on.

The new code $\mathsf{f}'(\cdot)$ will correspond to a lossless code for the source $\left\lceil \frac{X}{\lfloor \exp(d) \rfloor} \right\rceil$. A lossless code for $\left\lceil \frac{X}{\lfloor \exp(d) \rfloor} \right\rceil$ has optimal average length if it is a Huffman code which shows (127).

Equation (128) follows since the average length of the Huffman code for $\left\lceil \frac{X}{\lfloor \exp(d) \rfloor} \right\rceil$ can be upper bounded by that of the Shannon code for $\left\lceil \frac{X}{\lfloor \exp(d) \rfloor} \right\rceil$. □

D. Penalty for random coding

The performance of random coding for average length with an arbitrary distortion measure is given by the right hand side of

$$\ell_{\mathrm{v}}^*(d) \leq \inf_{P_{\hat{X}}} \sum_{n=1}^{\infty} \mathbb{E}\left[(1 - P_{\hat{X}}(\mathcal{B}_d(X)))^{2^n - 1}\right] \tag{132}$$

as shown in Theorem 25 (Appendix C). We can assess the gap between (132) and Theorem 13 by appealing to Lemma 8 in Section III.

Corollary 15 (Lemma 8)**.** *Assume without loss of generality that* $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$ *and* $P_X(i) \geq P_X(j)$ *for* $i \leq j$. *Let* $\mathsf{g}_1(\mathbf{t}_n^*)$ *be the solution to Optimization Problem 1 with* $k \leftarrow |\mathcal{X}|$, $a_i \leftarrow P_X(i)$, $M \leftarrow (2^n - 1)$, *and* $L \leftarrow \lfloor \exp(d) \rfloor$. *Then,*

$$\inf_{P_{\hat{X}}} \sum_{n=1}^{\infty} \mathbb{E}\left[(1 - P_{\hat{X}}(\mathcal{B}_d(X)))^{2^n - 1}\right] \geq \sum_{n=1}^{\infty} \mathsf{g}_1(\mathbf{t}_n^*). \tag{133}$$

Proof.

$$\inf_{P_{\hat{X}}} \sum_{n=1}^{\infty} \mathbb{E}\left[(1 - P_{\hat{X}}(\mathcal{B}_d(X)))^{2^n - 1}\right]$$

$$\geq \sum_{n=1}^{\infty} \inf_{P_{\hat{X}}} \mathbb{E}\left[(1 - P_{\hat{X}}(\mathcal{B}_d(X)))^{2^n - 1}\right] \tag{134}$$

$$\geq \sum_{n=1}^{\infty} \mathsf{g}_1(\mathbf{t}_n^*) \tag{135}$$

where (135) follows by Lemma 8. □

Example 3 (continued)**.** Figure 4 illustrates the gap between minimal average length $\ell_{\mathrm{v}}^*(d)$ and a lower bound (133) for the general random coding bound (132). Comparison to the converse bound in Theorem 12 is also provided.

V. Multiterminal single-shot bounds

In this section we state bounds for two multiterminal problems: compression with side information (Wyner-Ziv) and multiple descriptions coding. We focus on excess distortion and fixed-length setting. In both cases a random binning achievability approach gives tight bounds for all but small blocklengths.

A. Compression with side information

A fixed-length lossy source code of size M with side information is a pair of mappings,

$$\mathsf{f} \colon \mathcal{X} \to \mathcal{M}, \quad \mathsf{c} \colon \mathcal{M} \times \mathcal{Y} \to \mathcal{P}(\mathcal{X}). \tag{136}$$

A lossy source code with side information (f, c) is an (M, d, ϵ)-lossy source code if

$$\mathbb{P}\left[\mathsf{d}(X, \mathsf{c}(\mathsf{f}(X), Y)) > d\right] \leq \epsilon \tag{137}$$

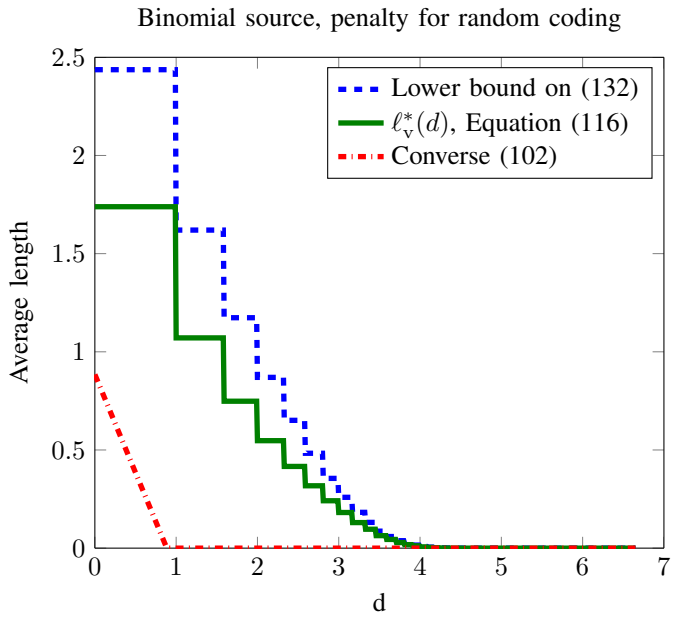

Fig. 4. Bounds on the minimal average length $\ell_{\mathrm{v}}^*(d)$ when X is a Binomial(n,p) source in Example 3.

and an (M, d)-lossy source code if

$$\mathbb{E}\left[\mathsf{d}(X, \mathsf{c}(\mathsf{f}(X), Y))\right] \le d. \tag{138}$$

The single-shot fundamental limits are defined as

$$\epsilon^*_{X|Y}(M, d) = \inf\{\epsilon \colon \exists (M, d, \epsilon)\text{-lossy source code}\} \tag{139}$$

and

$$d^*_{X|Y}(d) = \inf\{d \colon \exists (M, d)\text{-lossy source code}\}. \tag{140}$$

General single-shot achievability bounds for lossy source coding with side information were given in [11], [12]. The structure of the log-loss distortion enables a simple derivation of a special-purpose achievability result.

Theorem 16 (Achievability)**.** *If* $|\mathcal{X}| \le M\lfloor\exp(d)\rfloor$, *then* $\epsilon^*_{X|Y}(M, d) = 0$. *Otherwise,*

$$\epsilon^*_{X|Y}(M, d) \le \inf_{\gamma > 0}\{\mathbb{P}\left[\imath_{X|Y}(X|Y) > d + \log M - \gamma\right] + 2\exp(-\gamma)\}. \tag{141}$$

Proof. Let $L = \lfloor\exp(d)\rfloor$ and suppose $|\mathcal{X}| \le ML$. Then, ignoring the side information and using the code $(\mathsf{f}^*_L, \mathsf{c}^*_L)$ given by Definition 2 gives zero probability of excess distortion. If $|\mathcal{X}| > ML$, we show (141) using a random binning argument.

Compressor: The encoder is constructed by assigning to each $x \in \mathcal{X}$ a bin

$$\mathsf{b}(x) = (i, j) \in \{1, \ldots, M\} \times \{1, \ldots, L\}. \tag{142}$$

The encoder sends $\mathsf{b}_i(x) \in \{1, \ldots, M\}$ corresponding to the row of the bin assigned to x.

Decompressor: Fix $\gamma > 0$. Let the subsets $\mathcal{B}_y(i) \subset \mathcal{X}$ be given by

$$\mathcal{B}_y(i) = \left\{ x \in \bigcup_{j \colon |\mathcal{X}_{ij} \cap \mathcal{L}_y| = 1} \mathcal{X}_{i,j} \bigcap \mathcal{L}_y \right\} \tag{143}$$

where

$$\mathcal{L}_y = \left\{x : \imath_{X|Y}(x|y) \le d + \log M - \gamma\right\} \tag{144}$$

$$\mathcal{X}_{ij} = \mathsf{b}^{-1}(i, j). \tag{145}$$

It is important to note that $0 \le |\mathcal{B}_y(i)| \le L$. The decompressor outputs $\hat{P}_{i,y} = \mathsf{c}(i, y)$ given by

$$\hat{P}_{i,y}(x) = \begin{cases} \frac{1}{|\mathcal{B}_y(i)|}, & x \in \mathcal{B}_y(i) \\ 0, & \text{otherwise} \end{cases} \tag{146}$$

if $\mathcal{B}_y(i) \neq \emptyset$, and $\hat{P}_{i,y} = P_X$ if $\mathcal{B}_y(i) = \emptyset$.

Excess Distortion Analysis: Let $x_0 \in \mathcal{X}$ be the outcome to be compressed, and $y_0 \in \mathcal{Y}$ be the side information received at the decoder. Suppose $\mathsf{b}(x_0) = (i, j)$. If $x_0 \in \mathcal{B}_{y_0}(i)$ then the distortion incurred by (f, c) is

$$\log \frac{1}{\hat{P}_{i,y_0}(x_0)} = \log |\mathcal{B}_{y_0}(i)| \le \log L \le d. \tag{147}$$

If $x_0 \notin \mathcal{B}_{y_0}(i)$, then an excess distortion event occurs. This event is contained in the union of the following two events:

$$\mathcal{E}_1 \colon \imath_{X|Y}(x_0|y_0) > d + \log M - \gamma, \tag{148}$$

$$\mathcal{E}_2 \colon \exists x \in \mathsf{b}^{-1}(\mathsf{b}(x_0)), x \neq x_0 \colon \imath_{X|Y}(x|y_0) \le d + \log M - \gamma. \tag{149}$$

The probability of the first event is given by

$$\mathbb{P}[\mathcal{E}_1] = \mathbb{P}\left[\imath_{X|Y}(X|Y) > d + \log M - \gamma\right] \tag{150}$$

independent of the choice of the code (f, c). To bound $\mathbb{P}[\mathcal{E}_2]$, we fix x_0 and y_0, and average with respect to the choice of bins. If each element in $\mathcal{X}$ is assigned to bin (i, j) with probability $\frac{1}{LM}$ independent of other elements then we denote the random bin assigned to x by $\mathsf{B}(x)$. Then,

$$\mathbb{P}[\mathcal{E}_2] \tag{151}$$

$$\le \sum_{x \neq x_0} \mathbb{P}[\mathsf{B}(x) = \mathsf{B}(x_0)] 1\{\imath_{X|Y}(x|y_0) \le d + \log M - \gamma\}$$

$$\le \frac{1}{ML} \sum_{x \in \mathcal{X}} 1\{\imath_{X|Y}(x|y_0) \le d + \log M - \gamma\} \tag{152}$$

$$\le \frac{M \exp(d) \exp(-\gamma)}{ML} \le 2\exp(-\gamma). \tag{153}$$

Averaging over $\mathcal{X}$ and $\mathcal{Y}$, and optimizing over γ completes the proof. □

The next converse result gives a pleasing counterpart to Theorem 16. It can also be obtained by particularizing the d-tilted converse [13, Theorem 6] to the log-loss distortion measure.

Theorem 17 (Converse)**.**

$$\epsilon^*_{X|Y}(M, d) \tag{154}$$

$$\ge \sup_{\gamma > 0}\{\mathbb{P}\left[\imath_{X|Y}(X|Y) > d + \log M + \gamma\right] - \exp(-\gamma)\}$$

Proof. Fix γ and an arbitrary code (f, c). Define,

$$\mathcal{L} = \left\{(x, y) \colon \imath_{X|Y}(x|y) > d + \log M + \gamma\right\}, \tag{155}$$

$$\mathcal{C} = \left\{(x, y) \colon \mathsf{d}(x, \mathsf{c}(\mathsf{f}(x), y)) \le d\right\}, \tag{156}$$

$$\mathcal{C}_y = \{x \colon (x, y) \in \mathcal{C}\} \tag{157}$$

and observe that $|\mathcal{C}_y| \leq M\exp(d)$ for all $y \in \mathcal{Y}$. Indeed, $\mathsf{c}(\mathsf{f}(x), y)$ must assign at least $\exp(-d)$ probability mass to x in order for (x, y) to belong to $\mathcal{C}$. Thus, a single codeword can d-cover at most $\exp(d)$ elements in $\mathcal{X}$, and there is a total of M codewords. Then,

$$\mathbb{P}[\mathcal{L}] = \mathbb{P}[\mathcal{C}^c \cap \mathcal{L}] + \mathbb{P}[\mathcal{C} \cap \mathcal{L}] \tag{158}$$
$$\leq \mathbb{P}[\mathcal{C}^c] + \mathbb{P}[\mathcal{C} \cap \mathcal{L}] = \mathbb{P}[\mathcal{C}^c] + \mathbb{E}[\mathbb{P}[\mathcal{C} \cap \mathcal{L}]|Y] \tag{159}$$
$$= \mathbb{P}[\mathcal{C}^c] \tag{160}$$
$$+ \mathbb{E}\left[\sum_{x \in \mathcal{C}_Y} P_{X|Y}(x|Y) 1\{\imath_{X|Y}(x|Y) > d + \log M + \gamma\} \middle| Y\right]$$
$$\leq \mathbb{P}[\mathcal{C}^c] + \mathbb{E}\left[\frac{|\mathcal{C}_Y|}{\exp(d+\gamma)M} \middle| Y\right] \tag{161}$$
$$\leq \mathbb{P}[\mathcal{C}^c] + \exp(-\gamma). \tag{162}$$

The derived bound follows by rearranging the terms and optimizing over γ. □

B. Multiple descriptions

A multiple-descriptions source code with two compressors and three decompressors is a set of mappings,

$$\mathsf{f}_i \colon \mathcal{X} \to \mathcal{M}_i, \quad \mathsf{c}_i \colon \mathcal{M}_i \to \mathcal{P}(\mathcal{X}), \quad i \in \{1, 2\},$$
$$\mathsf{c}_0 \colon \mathcal{M}_1 \times \mathcal{M}_2 \to \mathcal{P}(\mathcal{X}). \tag{163}$$

A multiple-descriptions code is an $(M_1, M_2, d_0, d_1, d_2, \epsilon)$-lossy source code if $|\mathcal{M}_i| = M_i$ and $\mathbb{P}\left[\cup_{i=0}^{2} \mathcal{F}_i\right] \leq \epsilon$ where

$$\mathcal{F}_0 = \{\mathsf{d}(X, \mathsf{c}_0(\mathsf{f}_1(X), \mathsf{f}_2(X))) > d_0\} \tag{164}$$
$$\mathcal{F}_i = \{\mathsf{d}(X, \mathsf{c}_i(\mathsf{f}_i(X))) > d_i\}, i \in \{1, 2\}. \tag{165}$$

A multiple-descriptions code is an $(M_1, M_2, d_0, d_1, d_2)$-lossy source code if $|\mathcal{M}_i| = M_i$ and

$$[\mathsf{d}(X, \mathsf{c}_0(\mathsf{f}_1(X), \mathsf{f}_2(X)))] \leq d_0 \tag{166}$$
$$[\mathsf{d}(X, \mathsf{c}_i(\mathsf{f}_i(X)))] \leq d_i, i \in \{1, 2\}. \tag{167}$$

The single-shot fundamental limit for multiple-descriptions coding is defined as

$$\epsilon^*(M_1, M_2, d_0, d_1, d_2) \tag{168}$$
$$= \inf\{\epsilon \colon \exists\, (M_1, M_2, d_0, d_1, d_2, \epsilon)\text{-lossy source code}\}.$$

Theorem 18 (Achievability)**.**

$$\epsilon^*(M_1, M_2, d_0, d_1, d_2) \leq \inf_{\gamma > 0} \{\mathbb{P}[\imath_X(X) > K_\gamma] + 6\exp(-\gamma)\} \tag{169}$$

where

$$K_\gamma = \min\{d_0 + \log M_1 M_2, d_1 + \log M_1, d_2 + \log M_2\} - \gamma \tag{170}$$

Proof. We prove the result by using a random binning argument similar to the one used in Theorem 16.

Compressors: The compressors are constructed by assigning to each x a bin $\mathsf{b}(x) = (i, j) \in \{1, \ldots, M_1\} \times \{1, \ldots, M_2\}$. Then,

$$\mathsf{f}_1(x) = i, \quad \mathsf{f}_2(x) = j. \tag{171}$$

Decompressors: Fix $\gamma > 0$. Define $L_k = \lfloor \exp(d_k) \rfloor, \quad k \in \{0, 1, 2\}$. Suppose $|\mathcal{X}| \leq L_0$. Then, zero probability of excess distortion is achieved by $\hat{P} = \mathsf{c}_0(i, j)$ for all (i, j), where $\hat{P}(x) = \frac{1}{|\mathcal{X}|}$ for all $x \in \mathcal{X}$. Suppose $|\mathcal{X}| > L_0$. To construct decompressor c_0 each x is assigned a secondary bin $\tilde{\mathsf{b}}_0(x) = l_0 \in \{1, \ldots, L_0\}$. Let subsets $\mathcal{B}_0(i, j) \subset \mathcal{X}$ be given by

$$\mathcal{B}_0(i,j) = \left\{ x \in \bigcup_{l_0 \colon |\tilde{\mathcal{X}}_{l_0} \cap \mathcal{X}_{ij} \cap \mathcal{L}_0| = 1} \tilde{\mathcal{X}}_{l_0} \bigcap \mathcal{X}_{ij} \bigcap \mathcal{L}_0 \right\} \tag{172}$$

where

$$\mathcal{L}_0 = \{x : \imath_X(x) \leq d_0 + \log M_1 M_2 - \gamma\}, \tag{173}$$
$$\mathcal{X}_{ij} = \{x : \mathsf{b}(x) = (i, j)\}, \tag{174}$$
$$\tilde{\mathcal{X}}_{l_0} = \left\{x : \tilde{\mathsf{b}}_0(x) = l_0\right\}. \tag{175}$$

The decompressor outputs $\hat{P}_{ij} = \mathsf{c}_0(i, j)$ given by

$$\hat{P}_{ij}(x) = \begin{cases} \frac{1}{|\mathcal{B}_0(i,j)|}, & x \in \mathcal{B}_0(i, j) \\ 0, & \text{otherwise} \end{cases} \tag{176}$$

if $\mathcal{B}_0(i, j)$ is not empty, and $\hat{P}_{i,j} = P_X$ if it is empty.

Suppose $|\mathcal{X}| \leq L_1$. Then, zero probability of excess distortion is achieved by $\hat{P} = \mathsf{c}_1(i)$ for all i, where $\hat{P}(x) = \frac{1}{|\mathcal{X}|}$ for all $x \in \mathcal{X}$. Suppose $|\mathcal{X}| > L_1$. To construct decompressor c_1 each x is assigned a secondary bin $\tilde{\mathsf{b}}_1(x) = l_1 \in \{1, \ldots, L_1\}$. Let subsets $\mathcal{B}_1(i) \subset \mathcal{X}$ be given by

$$\mathcal{B}_1(i) = \left\{ x \in \bigcup_{l_1 \colon |\hat{\mathcal{X}}_{l_1} \cap \mathcal{X}_i \cap \mathcal{L}_1| = 1} \hat{\mathcal{X}}_{l_1} \cap \mathcal{X}_i \bigcap \mathcal{L}_1 \right\} \tag{177}$$

where

$$\mathcal{L}_1 = \{x : \imath_X(x) \leq d_1 + \log M_1 - \gamma\} \tag{178}$$
$$\mathcal{X}_i = \{x : \mathsf{b}_1(x) = i\} \tag{179}$$
$$\hat{\mathcal{X}}_{l_1} = \left\{x : \tilde{\mathsf{b}}_1(x) = l_1\right\}. \tag{180}$$

The decompressor outputs $\hat{P}_i = \mathsf{c}_1(i)$ are given by

$$\hat{P}_i(x) = \begin{cases} \frac{1}{|\mathcal{B}_1(i)|}, & x \in \mathcal{B}_1(i) \\ 0, & \text{otherwise} \end{cases} \tag{181}$$

if $\mathcal{B}_1(i) \neq \emptyset$, and $\hat{P}_{i,j} = P_X$ if $\mathcal{B}_1(i) = \emptyset$. The decompressor c_2 is defined analogously to c_1.

Excess Distortion Analysis: Let x_0 be the outcome to be compressed and suppose $(\mathsf{f}_1(x_0), \mathsf{f}_2(x_0)) = (i, j)$. If $x_0 \in \mathcal{B}_0(i, j)$, then the distortion incurred at decompressor c_0 is

$$\log \frac{1}{\hat{P}_{i,j}(x_0)} = \log |\mathcal{B}_0(i, j)| \leq \log L_0 \leq d_0. \tag{182}$$

If $x_0 \notin \mathcal{B}_0(i, j)$, then an excess distortion event, $\mathcal{F}_0$, occurs. This event is contained in the union of the following two events:

$$\mathcal{E}_{0,1} \colon \imath_X(x_0) > d_0 + \log M_1 M_2 - \gamma, \tag{183}$$
$$\mathcal{E}_{0,2} \colon \exists x \colon x \neq x_0, \mathsf{b}(x) = \mathsf{b}(x_0), \tilde{\mathsf{b}}_0(x) = \tilde{\mathsf{b}}_0(x_0),$$
$$\imath_X(x) \leq d_0 + \log M_1 M_2 - \gamma. \tag{184}$$

If $x_0 \in \mathcal{B}_1(i)$, then the distortion incurred at decompressor c_1 is

$$\log \frac{1}{\hat{P}_i(x_0)} = \log |\mathcal{B}_1(i)| \le \log L_1 \le d_1. \tag{185}$$

If $x_0 \notin \mathcal{B}_1(i)$, then an excess distortion event, $\mathcal{F}_1$, occurs. It is contained in the union of the following two events:

$$\mathcal{E}_{1,1}\colon \imath_X(x_0) > d_1 + \log M_1 - \gamma, \tag{186}$$

$$\mathcal{E}_{1,2}\colon \exists x\colon x \neq x_0, \mathsf{b}_i(x) = \mathsf{b}_i(x_0), \tilde{\mathsf{b}}_1(x) = \tilde{\mathsf{b}}_1(x_0), \quad \imath_X(x) \le d_1 + \log M_1 - \gamma. \tag{187}$$

Likewise, the distortion event $\mathcal{F}_2$ is contained in similarly defined events $\mathcal{E}_{2,1} \cup \mathcal{E}_{2,2}$. Combining these observations with a union bound gives,

$$\mathbb{P}\left[\cup_{i=0}^2 \mathcal{F}_i\right] \le \mathbb{P}\left[\cup_{i=0}^2 \mathcal{E}_{i,1}\right] + \sum_{i=0}^{2} \mathbb{P}\left[\mathcal{E}_{i,2}\right]. \tag{188}$$

The first term on the right side is upper bounded by $\mathbb{P}\left[\imath_X(X) > K_\gamma\right]$. The terms $\mathbb{P}[\mathcal{E}_{0,2}]$, $\mathbb{P}\left[\mathcal{E}_{1,2}\right]$ and $\mathbb{P}\left[\mathcal{E}_{2,2}\right]$ are bounded using steps analogous to (151) and (153). □

The next converse result gives a pleasing counterpart to Theorem 18. It immediately follows by applying (54) separately to each encoder-decoder pair.

Corollary 19 (Converse)**.**

$$\epsilon^*\left(M_1, M_2, d_0, d_1, d_2\right) \ge \sup_{\gamma > 0} \left\{\mathbb{P}\left[\imath_X(X) > K_\gamma\right] - \exp(-\gamma)\right\} \tag{189}$$

where

$$K_\gamma = \max\left\{d_0 + \log M_1 M_2, d_1 + \log M_1, d_2 + \log M_2\right\} + \gamma \tag{190}$$

VI. Asymptotic Theorems for Log-loss

In this section we give asymptotic characterizations of the rate-distortion function and the rate-distortion region for point-to-point coding in Theorem 21, coding with side information in Theorem 22, and multiple descriptions coding in Theorem 23.

Under the assumption that $(\mathbf{X}, \mathbf{Y})$ is stationary, the conventional definitions of unconditional and conditional entropy rates are

$$H(\mathbf{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n), \tag{191}$$

$$H(\mathbf{X}|\mathbf{Y}) = \lim_{n\to\infty} \frac{1}{n} H(X_1 \ldots, X_n | Y_1, \ldots Y_n). \tag{192}$$

Theorems 22 and 23 are a consequence of the single-shot bounds from Section V together with the Shannon-McMillan theorem.

Theorem 20 (Shannon-McMillan)**.** *Let $(\mathbf{X}, \mathbf{Y})$ be jointly stationary and ergodic taking values on discrete alphabets $\mathcal{A}$ and $\mathcal{B}$, with finite entropy rate $H(\mathbf{X}, \mathbf{Y})$. Then, for any $\delta > 0$*

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{1}{n}\imath_{X^n}(X^n) - H(\mathbf{X})\right| > \delta\right] = 0 \tag{193}$$

and

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n) - H(\mathbf{X}|\mathbf{Y})\right| > \delta\right] = 0. \tag{194}$$

Proof. See for example [10] for the proof of (193). Equation (194) follows from (193), as shown in Appendix D. □

A. Point-to-point rate-distortion functions

Let $\mathbf{X} = \{X_1^n\}_{n=1}^\infty$ be an information source. We say that (R, d) is achievable under the maximum distortion criterion (or m-achievable) if there exists a sequence of (M_n, d_n, ϵ_n)-lossy source codes with

$$\limsup_{n\to\infty} \frac{d_n}{n} \le d, \quad \lim_{n\to\infty} \epsilon_n = 0, \text{ and } \limsup_{n\to\infty} \frac{1}{n} \log M_n \le R. \tag{195}$$

The rate distortion function under maximum distortion fidelity criterion is defined as

$$\mathcal{R}_{\mathrm{m}}(d|\mathbf{X}) = \inf\left\{R\colon (R, d) \text{ is m-achievable}\right\}. \tag{196}$$

We say that (R, d) is achievable under the average distortion criterion (or a-achievable) if there exists a sequence of (M_n, d_n)-lossy source codes with

$$\limsup_{n\to\infty} \frac{1}{n} d_n \le d, \text{ and } \limsup_{n\to\infty} \frac{1}{n} \log M_n \le R. \tag{197}$$

The rate distortion function under average distortion fidelity criterion is defined as

$$\mathcal{R}_{\mathrm{a}}(d|\mathbf{X}) = \inf\left\{R\colon (R, d) \text{ is a-achievable}\right\}. \tag{198}$$

The next theorem characterizes the point-to-point rate-distortion function for the log-loss distortion measure.

Theorem 21. *Let $\mathbf{X}$ be stationary ergodic. Then,*

$$\mathcal{R}_{\mathrm{m}}(d|\mathbf{X}) = \mathcal{R}_{\mathrm{a}}(d|\mathbf{X}) = H(\mathbf{X}) - d. \tag{199}$$

Moreover, the strong converse holds. Given any sequence of (M_n, d, ϵ_n)-lossy source codes, if

$$\limsup_{n\to\infty} \frac{1}{n} \log M_n < H(\mathbf{X}) - d, \text{ then } \lim_{n\to\infty} \epsilon_n = 1. \tag{200}$$

Theorem 21 is a special case of Theorem 22.

Example 2 (continued)**.**

$$\mathcal{R}_{\mathrm{m}}(d|\mathbf{X}) = \mathcal{R}_{\mathrm{a}}(d|\mathbf{X}) = \frac{1}{2} - d \tag{201}$$

We can similarly define $\tilde{\mathcal{R}}_{\mathrm{m}}(d|\mathbf{X})$ and $\tilde{\mathcal{R}}_{\mathrm{a}}(d|\mathbf{X})$ to be rate-distortion functions for the case when the output alphabet is restricted to be a set of all product distributions. In that case we can show the following parametric upper bound:

$$\tilde{\mathcal{R}}_{\mathrm{m}}(d|\mathbf{X}) = \tilde{\mathcal{R}}_{\mathrm{m}}(d|\mathbf{X}) \le h(\epsilon_m) \tag{202}$$

$$d = \epsilon_m \sum_{i=1}^{m-1} h\left(\frac{(1-\epsilon_i)\epsilon_{m-i}}{\epsilon_m}\right) + (1-\epsilon_m) \sum_{1}^{m-1} h\left(\frac{(1-\epsilon_i)(1-\epsilon_{m-i})}{1-\epsilon_m}\right)$$

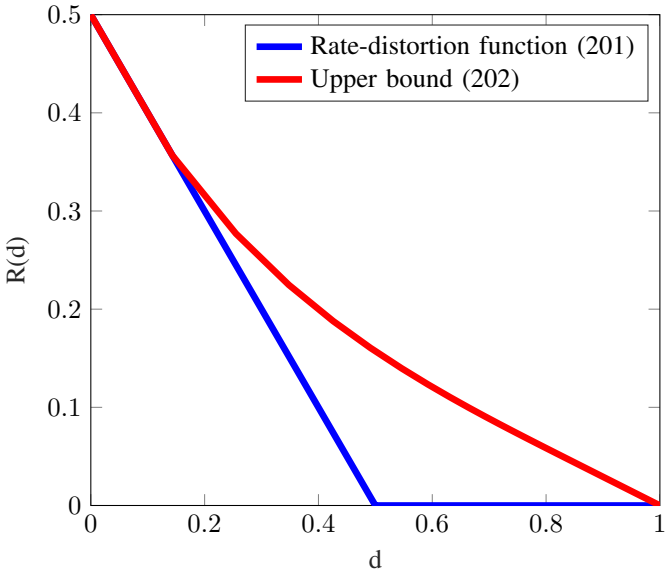

Fig. 5. Bounds on the two kinds of rate-distortion functions in Example 2.

where $m \in \{1, 2, \dots\}$, (202) holds with equality for $m \in \{1, 2\}$, and ϵ_m denotes mth order convolution of ϵ. That is,

$$\epsilon_m = \begin{cases} \epsilon, & m = 1 \\ \epsilon_{m-1}(1-\epsilon) + (1-\epsilon_{m-1})\epsilon, & m > 1. \end{cases} \tag{203}$$

Upper bound (202) is derived by losslessly encoding every mth bit in the Markov chain. The remaining bits are reconstructed with a posterior distribution given the known bits. As can be seen in Figure 5, the upper bound (202) is tight at distortion lower than 0.14.

B. Side information (Wyner-Ziv) rate-distortion function

The general formula for source coding with side information was given in [14], while [12] analyzed dispersion of source coding with side information for general distortion measures and stationary memoryless sources. By focusing on the log-loss distortion measure we are able to obtain a simpler formula than the one derived in [14] for stationary ergodic sources, as well as general sources.

Let $\mathbf{X} = \{X_1^n\}_{n=1}^\infty$ be an information source with side information process $\mathbf{Y} = \{Y_1^n\}_{n=1}^\infty$. We say that (R, d) is achievable under the maximum distortion criterion (or m-achievable) if there exists a sequence of (M_n, d_n, ϵ_n)-lossy source codes with

$$\limsup_{n\to\infty} \frac{d_n}{n} \le d, \quad \lim_{n\to\infty} \epsilon_n = 0, \text{ and } \limsup_{n\to\infty} \frac{1}{n} \log M_n \le R. \tag{204}$$

The rate distortion function under the maximum distortion fidelity criterion is defined as

$$\mathcal{R}_{\mathrm{m}}(d|\mathbf{X}, \mathbf{Y}) = \inf \{R \colon (R, d) \text{ is m-achievable}\}. \tag{205}$$

We say that (R, d) is achievable under the average distortion criterion (or a-achievable) if there exists a sequence of (M_n, d_n)-lossy source codes with

$$\limsup_{n\to\infty} \frac{1}{n} d_n \le d, \text{ and } \limsup_{n\to\infty} \frac{1}{n} \log M_n \le R. \tag{206}$$

The rate distortion function under the average distortion fidelity criterion is defined as

$$\mathcal{R}_{\mathrm{a}}(d|\mathbf{X}, \mathbf{Y}) = \inf \{R \colon (R, d) \text{ is a-achievable}\}. \tag{207}$$

The next result characterizes the Wyner-Ziv rate-distortion function for the log-loss distortion measure.

Theorem 22. *Let $(\mathbf{X}, \mathbf{Y})$ be jointly stationary ergodic. Then,*

$$\mathcal{R}_{\mathrm{m}}(d|\mathbf{X}, \mathbf{Y}) = \mathcal{R}_{\mathrm{a}}(d|\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}|\mathbf{Y}) - d. \tag{208}$$

Moreover, the strong converse holds. Given any sequence of (M_n, d, ϵ_n)-lossy source codes, if

$$\limsup_{n\to\infty} \frac{1}{n} \log M_n < H(\mathbf{X}|\mathbf{Y}) - d, \text{ then } \lim_{n\to\infty} \epsilon_n = 1. \tag{209}$$

Appendix E gives proof of Theorem 22 by applying Theorem 20 to the single-shot bounds in Theorems 16 and 17.

C. Rate-distortion region for multiple descriptions

There have been few works that study multiple descriptions coding with memory [15]–[18], since the rate region is not known even for stationary memoryless sources. Results for Gaussian random processes were obtained in [16], [17], while multi-letter solution for general distortion measures was presented in [18]. The rate-distortion region for log-loss distortion was studied in [19]. The time sharing approach in [19] characterizes the rate-distortion region with two encoders and three decoders for stationary memoryless sources. Our single-shot bounds allow us to obtain a complete single-letter characterization of the rate-distortion region for stationary ergodic sources, as well as general sources. Moreover, our achievability bound, which is based on random binning, can be directly extended to any number of encoders and decoders.

We say that $(R_1, R_2, d_0, d_1, d_2)$ is achievable under the maximum distortion criterion (or m-achievable) if there exists a sequence of $(M_1, M_2, d_0, d_1, d_2, \epsilon)$-lossy source codes with

$$\limsup_{n\to\infty} \frac{d_{n,i}}{n} \le d_i, i \in \{0, 1, 2\}, \tag{210}$$

$$\lim_{n\to\infty} \epsilon_n = 0, \tag{211}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log M_{n,i} \le R_i, i \in \{1, 2\}. \tag{212}$$

The rate distortion region under maximum distortion fidelity criterion is defined as

$$\mathcal{R}_{\mathrm{m}}(d_0, d_1, d_2|\mathbf{X}) = \{(R_1, R_2) \colon (R_1, R_2, d_0, d_1, d_2) \text{ is m-achievable}\}. \tag{213}$$

We say that $(R_1, R_2, d_0, d_1, d_2)$ is achievable under the average distortion criterion (or a-achievable) if there exists a sequence of $\big((M_{n,i})_{i=1}^2, (d_{n,i})_{i=0}^2, \big)$-lossy source codes with

$$\limsup_{n\to\infty} \frac{d_{n,i}}{n} \le d, i \in \{0, 1, 2\}, \tag{214}$$

$$\text{and } \limsup_{n\to\infty} \frac{1}{n} \log M_{n,i} \le R_i, i \in \{1, 2\}. \tag{215}$$

The rate distortion function under average distortion fidelity criterion is defined as

$$\mathcal{R}_{\mathrm{a}}(d_0, d_1, d_2|\mathbf{X}) = \{(R_1, R_2)\colon (R_1, R_2, d_0, d_1, d_2) \text{ is a-achievable}\}. \quad (216)$$

Theorem 23. *Let* $\mathbf{X}$ *be a stationary and ergodic source. Then, the multiple-descriptions rate-distortion regions,* $\mathcal{R}_{\mathrm{m}}(d_0, d_1, d_2|\mathbf{X})$ *and* $\mathcal{R}_{\mathrm{a}}(d_0, d_1, d_2|\mathbf{X})$*, are given by*

$$R_1 \geq H(\mathbf{X}) - d_1, \quad (217)$$

$$R_2 \geq H(\mathbf{X}) - d_2, \quad (218)$$

$$R_1 + R_2 \geq H(\mathbf{X}) - d_0. \quad (219)$$

The result follows from the Shannon-McMillan Theorem applied to Corollary 19 and Theorem 18. See Appendix F for details.

Note that the solution in Theorem 23 reduces to separate compression problems if $d_0 > \min\{d_1, d_2\}$. Otherwise, the corner points of the region are $(H(\mathbf{X}) - d_1, d_1 - d_0)$ and $(d_2 - d_0, H(\mathbf{X}) - d_2)$. Theorem 23 can be extended to any number of compressors and decompressors, making log-loss distortion amenable to multiple descriptions.

VII. Conclusion

The structure of the log-loss distortion measure lets us obtain special purpose bounds which are tighter and simpler than general purpose bounds. Moreover, the single-shot setting exposes connections between lossless source coding with list decoding and lossy compression with log-loss distortion, as well as enables a particularly simple approach to multi-terminal problems, such as lossy source coding with side information, and multiple descriptions coding. The main contributions of this paper are summarized as follows.

A. Single-shot point-to-point bounds

The problem of lossy source coding with log-loss is intimately related to the problem of lossless coding with list decoding. More specifically, Theorem 6 characterizes the single-shot fundamental limit, $\epsilon_{\mathrm{f}}^*(d)$, of excess distortion for fixed-length coding of X which is the same as the minimal probability of error in almost lossless source coding of $\left\lceil \frac{X}{\lfloor \exp(d) \rfloor} \right\rceil$, characterized in [7]. Theorems 13 and 14 expose the same connection between log-loss and lossless coding with list decoding for variable-length codes, with and without prefix constraints. In both cases the optimal average length, $\ell_{\mathrm{v}}^*(d)$ and $\ell_{\mathrm{p}}^*(d)$, for the source X corresponds to the optimal lossless average length of the source $\left\lceil \frac{X}{\lfloor \exp(d) \rfloor} \right\rceil$.

The connection between log-loss lossy compression and lossless coding with list decoding does not hold for the non-asymptotic fundamental limit of average distortion, $d_{\mathrm{f}}^*(M)$. Instead, we provide lower and upper bounds on $d_{\mathrm{f}}^*(M)$ which are tight asymptotically. The lower (converse) bound is given by Theorem 3, while the upper (achievability) bound is given by Theorem 4. The proof of Theorem 4 uses an explicit greedy construction which, as seen in Figure 1, gives excellent performance.

B. Multi-terminal bounds

Two multi-terminal problems are addressed: coding with side information (Wyner-Ziv), and multiple descriptions coding. Theorems 16 and 17 give upper and lower bounds on the non-asymptotic fundamental limit of source coding with side information. A simple application of the Shannon-McMillan Theorem allows us to get the rate-distortion function for stationary ergodic sources in Theorem 22. The single-shot results in Theorems 16 and 17 can also be leveraged to obtain results that do not require stationarity and ergodicity.

Theorem 18 gives an upper bound on the non-asymptotic fundamental limit for multiple descriptions coding relying on the technique of random binning, and can be directly extended to any number of encoders and decoders. Another application of the Shannon-McMillan Theorem to Theorem 18 and Corollary 19 lets us derive the multiple descriptions rate-distortion region in Theorem 23.

C. Results for general distortion measures

Although the focus of this work is on the log-loss distortion measure, two novel results which hold for general distortion measures are also stated. Theorem 10 connects the non-asymptotic fundamental limits for probability of excess distortion in the fixed-length setting and for the probability of excess length in the variable-length setting. Theorem 25 gives a novel random coding achievability bound on the non-asymptotic fundamental limit of average length in the variable length setting without prefix constraints.

D. Penalty for random coding

By focusing on the log-loss distortion measure we can derive exact expressions for the non-asymptotic fundamental limits $\epsilon_{\mathrm{f}}^*(d)$ and $\ell_{\mathrm{v}}^*(d)$. These limits are not known exactly for general distortion measures, and often the best achievability bounds available are based on random coding. Thus, the log-loss distortion measure gives us a unique opportunity to investigate the penalty incurred by random coding in the non-asymptotic setting. This penalty is illustrated in Figures 2 and 4.

E. Future work

Lossy compression with logarithmic loss in the single-shot setting raises the question of how to extend the problem to the n-shot setting. In the case of log-loss there are two natural ways of doing this: the approach adopted here (2) and the standard approach when the output alphabet is restricted to be the set of all product measures over $\mathcal{X}^n$ (i.e. the output soft information is given on a per-symbol basis). Most of the information theoretic literature focuses on rate-distortion functions of memoryless sources, and for this problem both n-shot settings give the same rate-distortion function, as was shown in [2]. As we illustrate in Example 2, this need not hold in the non-asymptotic setting, or when the source has memory. The fundamental limits for the simple Markov chain in Example 2 are very different at low rates; however, at high rates the rate-distortion functions turn out to be the same. In

general, log-loss is a motivating example for why it would be of interest to understand how the size of the output alphabet influences fundamental limits of compression (both asymptotic and non-asymptotic). Finally, while there are good reasons to consider both settings, our setting (2) is more in line with the learning literature, a fact that pays dividends when tackling universal lossy source coding [20].

APPENDIX

A. Comparison of special and general purpose bounds

This appendix provides auxiliary lemmas and proofs for comparison of special and general purpose bounds in Sections III and IV.

Solution to Optimization Problem 1 can be obtained analytically, as shown in the following lemma.

Lemma 24. *For $k < \infty$, the solution to Optimization Problem 1 is given by*

$$t_i = \begin{cases} 0, & i > k_0 \\ 1 - (k_0 - L)\frac{b_i}{\sum_{i=1}^{k_0} b} & \text{otherwise} \end{cases} \tag{220}$$

where

$$b_i = \left(\frac{1}{a_i}\right)^{\frac{1}{M-1}}, \forall i \in \{1, \dots, k\} \tag{221}$$

$$\text{and } k_0 = \max\left\{ i \colon \sum_{j=1}^{i} \frac{b_j}{b_i} \geq i - 1 \right\}. \tag{222}$$

Proof. The characterization of this solution is obtained through the method of Lagrange multipliers. □

The next lemma provides lower bounds on the random coding bound for log-loss in terms of Optimization Problem 1.

Lemma 8. Let $P_{\hat{X}}$ be any probability measure over $\mathcal{P}(\mathcal{X})$ and let $t_i = P_{\hat{X}}\left(\mathcal{B}_d(i)\right)$. Then,

$$\mathbb{E}\left[\left(1 - P_{\hat{X}}(\mathcal{B}_d(X))\right)^M\right] = \mathsf{g}_1(\mathbf{t}) \tag{223}$$

is the function (72) being minimized in Optimization Problem 1. Moreover,

$$\sum_{i\in\mathcal{X}} t_i = \sum_{i\in\mathcal{X}} P_{\hat{X}}\left(\mathcal{B}_d(i)\right) \tag{224}$$

$$= \sum_{i\in\mathcal{X}} \int_{\hat{P}\in\mathcal{P}(\mathcal{X})} \mathrm{d}P_{\hat{X}}(\hat{P}) 1\{\hat{P} \in \mathcal{B}_d(i)\} \tag{225}$$

$$= \int_{\hat{P}\in\mathcal{P}(\mathcal{X})} \mathrm{d}P_{\hat{X}}(\hat{P}) \sum_{i\in\mathcal{X}} 1\{\hat{P} \in \mathcal{B}_d(i)\} \tag{226}$$

$$\leq \int_{\hat{P}\in\mathcal{P}(\mathcal{X})} \mathrm{d}P_{\hat{X}}(\hat{P}) \lfloor \exp(d) \rfloor = \lfloor \exp(d) \rfloor, \tag{227}$$

where (227) follows by Lemma 1. Thus, the lower bound (74) holds since $\mathbf{t}$ is a feasible point for Optimization Problem 1. The lower bound (74) holds with equality when $0 \leq d < 1$ since we have the following valid choice

$$P_{\hat{X}}(\hat{P}) = \begin{cases} t_i^*, & \hat{P} = \delta_i \\ 0, & \text{otherwise} \end{cases} \tag{228}$$

Lower bound (74) holds with equality for $M = 1$ by inspection. □

Lemma 9. Equation (80) follows since

$$\min\left(a_i, t_i\right) \frac{1}{\frac{\min(a_i,t_i)}{a_i t_i} + M - 1}$$

$$= \frac{a_i t_i}{1 + \frac{a_i t_i}{\min(a_i,t_i)}(M-1)} \tag{229}$$

$$\leq \frac{a_i t_i}{1 + a_i(M-1)}. \tag{230}$$

Next, let $d = 0$ and observe that for log-loss $1\{\mathsf{d}(i, P) \leq 0\} = 1$ only if $P = \delta_i$, where, recall, δ_i is a point mass on i. We simplify (68) and obtain

$$\mathbb{E}\left[\frac{1\{\mathsf{d}(X,\hat{X}) \leq d\}}{\exp\left(\imath_{X;\hat{X}}(X;\hat{X})\right) + M - 1}\right]$$

$$= \sum_{i\in\mathcal{X}} P_{X\hat{X}}(i, \delta_i) \frac{1}{\frac{P_{X\hat{X}}(i,\delta_i)}{P_X(i)P_{\hat{X}}(\delta_i)} + M - 1} \tag{231}$$

$$= \sum_{i\in\mathcal{X}} \frac{P_X(i)P_{\hat{X}}(\delta_i)P_{X\hat{X}}(i,\delta_i)}{P_{X\hat{X}}(i,\delta_i) + P_X(i)P_{\hat{X}}(\delta_i)(M-1)}. \tag{232}$$

Now suppose $P_{\hat{X}}(\delta_i)$ is fixed for all $i \in \mathcal{X}$. Equation (231) is maximized whenever $P_{X\hat{X}}(i, \delta_i)$, or equivalently, $P_{\hat{X}|X}(\delta_i|i)$, are maximized for all i. This is achieved by,

$$P_{\hat{X}|X}(\delta_i|i) = \begin{cases} 1, & P_X(i) \leq P_{\hat{X}}(\delta_i) \\ \frac{P_{\hat{X}}(\delta_i)}{P_X(i)}, & \text{otherwise} \end{cases} \tag{233}$$

or equivalently, $P_{X\hat{X}}(i, \delta_i) = \min\left(P_X(i), P_{\hat{X}}(\delta_i)\right)$. By taking $t_i = P_{\hat{X}}(\delta_i)$ we obtain exactly Optimization Problem 2. This shows (79) for $d = 0$.

Next, we argue that it is sufficient to restrict our attention to single-mass distributions in the case $0 \leq d < 1$. This is easy to see if $P_{\hat{X}}$ is assumed to have a discrete support. Indeed, for a fixed $i \in \mathcal{X}$ we have

$$\sum_{\hat{P}\in\mathcal{B}_d(i)} \frac{P_X(i)P_{\hat{X}}(\hat{P})P_{X\hat{X}}(i,\hat{P})}{P_{X\hat{X}}(i,\hat{P}) + P_X(i)P_{\hat{X}}(\hat{P})(M-1)}$$

$$\leq \frac{\sum_{\hat{P}\in\mathcal{B}_d(i)} P_X(i)P_{\hat{X}}(\hat{P}) \sum_{\hat{P}\in\mathcal{B}_d(i)} P_{X\hat{X}}(i,\hat{P})}{\sum_{\hat{P}\in\mathcal{B}_d(i)} P_{X\hat{X}}(i,\hat{P}) + (M-1)\sum_{\hat{P}\in\mathcal{B}_d(i)} P_X(i)P_{\hat{X}}(\hat{P})}. \tag{234}$$

Equation (234) follows by Milne's Inequality [21, Theorem 67],

$$\sum_{k=1}^{n}(a_k + b_k) \sum_{j=1}^{n} \frac{a_j b_j}{a_j + b_j} \leq \sum_{k=1}^{n} a_j \sum_{j=1}^{n} b_j, \tag{235}$$

with $a_k = P_{X\hat{X}}(i,\hat{P})$ and $b_k = (M-1)P_X(i)P_{\hat{X}}(\hat{P})$. If $P_{\hat{X}}$ has continuous support we can use the integral form of Milne's inequality. □

B. Proof of Theorem 10

The regularity condition on X implies that for a fixed $d > 0$ there exists a prefix-free lossy source code $(\mathsf{f}_p, \mathsf{c}_p)$ such that $\mathsf{d}\left(x, \mathsf{c}_p\left(\mathsf{f}_p(x)\right)\right) \le d$ for all $x \in \mathcal{X}$. We will use $(\mathsf{f}_p, \mathsf{c}_p)$ in the proofs for claims 1) and 2).

1) Equation (89) follows since any fixed-length $(2^{l+1} - 1, d, \epsilon)$-code can be used to construct a variable-length (l, d, ϵ)-code, and vice versa. Indeed, fix a variable-length (l, d, ϵ)-code $(\tilde{\mathsf{f}}, \tilde{\mathsf{c}})$ and let $\mathcal{M}_l$ be the subset of $\{0,1\}^*$ consisting of all binary vectors of length at most l (including the empty string). Define

$$\hat{\mathsf{f}}(x) = \begin{cases} \tilde{\mathsf{f}}(x), & \tilde{\mathsf{f}}(x) \in \mathcal{M}_l, \\ \mathbf{s}_0, & \text{otherwise} \end{cases} \tag{236}$$

$$\text{and } \hat{\mathsf{c}}(\mathbf{s}) = \tilde{\mathsf{c}}(\mathbf{s}), \quad \forall \mathbf{s} \in \mathcal{M}_l \tag{237}$$

with $\mathbf{s}_0 \in \mathcal{M}_l$ selected arbitrarily. Then, $(\hat{\mathsf{f}}, \hat{\mathsf{c}})$ is a fixed-length $(2^{l+1} - 1, d, \epsilon)$-code since

$$|\mathcal{M}_l| = 2^{l+1} - 1, \tag{238}$$

$$\mathbb{P}\left[\mathsf{d}\left(X, \hat{\mathsf{c}}\left(\hat{\mathsf{f}}(X)\right)\right) > d\right] \le \mathbb{P}\left[\ell\left(\tilde{\mathsf{f}}(X)\right) > l\right] \le \epsilon. \tag{239}$$

Now, select a fixed-length $(2^{l+1} - 1, d, \epsilon)$-code $(\hat{\mathsf{f}}, \hat{\mathsf{c}})$ and assume without loss of generality that $\hat{\mathsf{f}}(\mathcal{X}) = \mathcal{M}_l$. Define

$$\tilde{\mathsf{f}}(x) = \begin{cases} \hat{\mathsf{f}}(x), & \mathsf{d}\left(x, \hat{\mathsf{c}}\left(\hat{\mathsf{f}}(x)\right)\right) \le d, \\ 0^l \cdot \mathsf{f}_p(x), & \text{otherwise} \end{cases} \tag{240}$$

$$\text{and } \tilde{\mathsf{c}}(\mathbf{s}) = \begin{cases} \hat{\mathsf{c}}(\mathbf{s}), & \mathbf{s} \in \mathcal{M}_l, \\ \mathsf{c}_p(\mathbf{s}'), & \text{if } \mathbf{s} = 0^l \cdot \mathbf{s}' \text{ for some } \mathbf{s}' \in \mathsf{f}_{\mathrm{p}}(\mathcal{X}), \\ \hat{x}_0, & \text{otherwise} \end{cases} \tag{241}$$

where we have chosen an arbitrary $\hat{x}_0 \in \hat{\mathcal{X}}$. It follows by construction that $(\tilde{\mathsf{f}}, \tilde{\mathsf{c}})$ is a variable-length (l, d, ϵ)-code. Finally,

$$\epsilon^*_{\mathrm{v}}(l, d) = \inf\{\epsilon \colon \exists (l, d, \epsilon)\text{-code}\} \tag{242}$$

$$= \inf\{\epsilon \colon \exists (2^{l+1} - 1, d, \epsilon)\text{-code}\} \tag{243}$$

$$= \epsilon^*_{\mathrm{f}}(2^{l+1} - 1, d). \tag{244}$$

2) Suppose $l \ge \log M_d$. Then, let $(\hat{\mathsf{f}}, \hat{\mathsf{c}})$ be a fixed-length (M_d, d, ϵ)-code which achieves $\epsilon^*_{\mathrm{f}}(M_d, d) = 0$. Since $M_d \le 2^l$ we assume without loss of generality that $\hat{\mathsf{f}}(\mathcal{X}) \subset \{0,1\}^l$. Then, we can establish a one-to-one correspondence between this fixed-length code and the prefix-free code $(\tilde{\mathsf{f}}, \tilde{\mathsf{c}})$:

$$\tilde{\mathsf{f}}(x) = \hat{\mathsf{f}}(x), \tag{245}$$

$$\tilde{\mathsf{c}}(\mathbf{s}) = \begin{cases} \hat{\mathsf{c}}(\mathbf{s}), & \mathbf{s} \in \hat{\mathsf{f}}(\mathcal{X}), \\ \hat{x}_0, & \text{otherwise} \end{cases} \tag{246}$$

where, as before, the choice of $\hat{x}_0$ is arbitrary. The prefix free code $(\tilde{\mathsf{f}}, \tilde{\mathsf{c}})$ is an $(l, d, 0)$-code and (90) is proved.

Suppose $l < \log M_d$. Equation (90) follows since any fixed-length $(2^l - 1, d, \epsilon)$-code can be used to construct a prefix-free (l, d, ϵ)-code, and vice versa. Indeed, fix a prefix-free (l, d, ϵ)-code $(\tilde{\mathsf{f}}, \tilde{\mathsf{c}})$ and let

$$\tilde{\mathcal{M}}_l = \mathcal{M}_l \cap \tilde{\mathsf{f}}(\mathcal{X}). \tag{247}$$

Define

$$\hat{\mathsf{f}}(x) = \begin{cases} \tilde{\mathsf{f}}(x), & \tilde{\mathsf{f}}(x) \in \tilde{\mathcal{M}}_l \\ \mathbf{s}_0, & \text{otherwise} \end{cases} \tag{248}$$

$$\hat{\mathsf{c}}(\mathbf{s}) = \tilde{\mathsf{c}}(\mathbf{s}) \tag{249}$$

with $\mathbf{s}_0 \in \tilde{\mathcal{M}}_l$ selected arbitrarily. Then, $(\hat{\mathsf{f}}, \hat{\mathsf{c}})$ is a fixed-length $(2^l - 1, d, \epsilon)$-code since

$$|\tilde{\mathcal{M}}_l| \le 2^l - 1, \tag{250}$$

$$\mathbb{P}\left[\mathsf{d}\left(X, \hat{\mathsf{c}}\left(\hat{\mathsf{f}}(X)\right)\right) > d\right] \le \mathbb{P}\left[\ell\left(\tilde{\mathsf{f}}(X)\right) > l\right] \le \epsilon. \tag{251}$$

Now, select a fixed-length $(2^l - 1, d, \epsilon)$-code $(\hat{\mathsf{f}}, \hat{\mathsf{c}})$ and assume without loss of generality that $\hat{\mathsf{f}}(\mathcal{X}) = \{0,1\}^l \setminus 0^l$. Define

$$\tilde{\mathsf{f}}(x) = \begin{cases} \hat{\mathsf{f}}(x), & \mathsf{d}\left(x, \hat{\mathsf{c}}\left(\hat{\mathsf{f}}(x)\right)\right) \le d, \\ 0^l \cdot \mathsf{f}_p(x), & \text{otherwise} \end{cases} \tag{252}$$

$$\text{and } \tilde{\mathsf{c}}(\mathbf{s}) = \begin{cases} \hat{\mathsf{c}}(\mathbf{s}), & \mathbf{s} \in \hat{\mathsf{f}}(\mathcal{X}), \\ \mathsf{c}_p(\mathbf{s}'), & \text{if } \mathbf{s} = 0^l \cdot \mathbf{s}' \text{ for some } \mathbf{s}' \in \mathsf{f}_{\mathrm{p}}(\mathcal{X}), \\ \hat{x}_0, & \text{otherwise} \end{cases} \tag{253}$$

where, as before, the choice of $\hat{x}_0$ is immaterial. It follows by construction that $(\tilde{\mathsf{f}}, \tilde{\mathsf{c}})$ is a variable-length (l, d, ϵ)-code. Finally

$$\epsilon^*_{\mathrm{p}}(l, d) = \inf\{\epsilon \colon \exists (l, d, \epsilon)\text{-code}\} \tag{254}$$

$$= \inf\{\epsilon \colon \exists (2^{l+1} - 1, d, \epsilon)\text{-code}\} \tag{255}$$

$$= \epsilon^*_{\mathrm{f}}(2^l - 1, d). \tag{256}$$

C. Random coding variable-length bound

The following theorem gives an achievability bound by computing the exact average length achievable by a random code where codewords are drawn independently with distribution $P_{\hat{X}}$.

Theorem 25 (Achievability)**.**

$$\ell^*_{\mathrm{v}}(d) \le \inf_{P_{\hat{X}}} \sum_{k=1}^{\infty} \mathbb{E}\left[\left(1 - P_{\hat{X}}(\mathcal{B}_d(X))\right)^{2^k - 1}\right] \tag{257}$$

$$\le \inf_{P_{\hat{X}}} \sum_{k=1}^{\infty} \mathbb{E}\left[e^{-(2^k - 1) P_{\hat{X}}(\mathcal{B}_d(X))}\right] \tag{258}$$

Proof. Let $\mathbf{a} = a_1, a_2, \ldots$ denote an arbitrary sequence of elements in $\hat{\mathcal{X}}$. Given any such sequence define a variable length code by letting the decompressor assign elements of $\mathbf{a}$ to binary strings in a lexicographical order, that is $\mathsf{c}_{\mathbf{a}}(\emptyset) = a_1$, $\mathsf{c}_{\mathbf{a}}(0) = a_2$, $\mathsf{c}_{\mathbf{a}}(1) = a_3$, $\mathsf{c}_{\mathbf{a}}(00) = a_4$, and so on. The compressor is defined by assigning binary strings to $\mathsf{f}_{\mathbf{a}}(x)$ in lexicographical order based on $m^*(x)$ where

$$m^*(x) = \operatorname*{arg\,min}_{m \in \{1, 2, \ldots\}} \{m \colon a_m \in \mathcal{B}_d(x)\}. \tag{259}$$

That is, $\mathsf{f}_{\mathbf{a}}(x) = \emptyset$ if $m^*(x) = 1$, $\mathsf{f}_{\mathbf{a}}(x) = 0$ if $m^*(x) = 2$, and so on. Fix distribution $P_{\hat{X}}$ and generate a countably infinite i.i.d. sequence $\hat{X}_1, \hat{X}_2, \ldots$. Then, for every $x \in \mathcal{X}$,

$$\mathbb{E}\left[\ell(\mathsf{f}_{\hat{\mathbf{X}}}(x))\right] = \sum_{k=1}^{\infty} k\, \mathbb{P}[\ell(\mathsf{f}_{\hat{\mathbf{X}}}(x)) = k] \tag{260}$$

$$= \sum_{k=1}^{\infty} k \left(1 - (1 - P_{\hat{X}}(\mathcal{B}_d(x)))^{2^k}\right) \prod_{i=0}^{k-1} \left(1 - P_{\hat{X}}(\mathcal{B}_d(x))\right)^{2^i} \tag{261}$$

$$= \sum_{k=1}^{\infty} k \left(\prod_{i=0}^{k-1} \left(1 - P_{\hat{X}}(\mathcal{B}_d(x))\right)^{2^i} - \prod_{i=0}^{k} \left(1 - P_{\hat{X}}(\mathcal{B}_d(x))\right)^{2^i} \right) \tag{262}$$

$$= \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \left(1 - P_{\hat{X}}(\mathcal{B}_d(x))\right)^{2^i} \tag{263}$$

$$= \sum_{k=1}^{\infty} \left(1 - P_{\hat{X}}(\mathcal{B}_d(x))\right)^{2^k - 1} \tag{264}$$

where (263) holds by letting $b_k = \prod_{i=0}^{k-1} \left(1 - P_{\hat{X}}(\mathcal{B}_d(x))\right)^{2^i}$ and noting that $\sum_{k=1}^{\infty} k(b_k - b_{k+1}) = \sum_{k=1}^{\infty} b_k$. Taking the expectation with respect to X yields the expected length where averaging is with respect to both the source and the code selection. Since on average random codes satisfy given bounds there must exist at least one deterministic code which does as well. To obtain (258) we apply $(1-x)^M \leq e^{-MX}$ to each term in the product:

$$\ell_{\mathrm{v}}^*(d) \leq \inf_{P_{\hat{X}}} \sum_{k=1}^{\infty} \mathbb{E}\left[\prod_{i=0}^{k-1} (1 - P_{\hat{X}}(\mathcal{B}_d(X)))^{2^i}\right] \tag{265}$$

$$\leq \inf_{P_{\hat{X}}} \sum_{k=1}^{\infty} \mathbb{E}\left[\prod_{i=0}^{k-1} e^{-2^i P_{\hat{X}}(\mathcal{B}_d(X))}\right] \tag{266}$$

$$= \inf_{P_{\hat{X}}} \sum_{k=1}^{\infty} \mathbb{E}\left[e^{-\sum_{i=0}^{k-1} 2^i P_{\hat{X}}(\mathcal{B}_d(X))}\right] \tag{267}$$

$$= \inf_{P_{\hat{X}}} \sum_{k=1}^{\infty} \mathbb{E}\left[e^{-(2^k-1) P_{\hat{X}}(\mathcal{B}_d(X))}\right]. \tag{268}$$

□

D. Proof of (194)

Define

$$\mathcal{E}_n = \mathbb{P}\left[\left|\frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n) - H(\mathbf{X}|\mathbf{Y})\right| > \delta\right] \tag{269}$$

$$\mathcal{E}_n^1 = \mathbb{P}\left[\left|\frac{1}{n}\imath_{Y^n}(Y^n) - H(\mathbf{Y})\right| > \frac{\delta}{2}\right] \tag{270}$$

$$\mathcal{E}_n^2 = \mathbb{P}\left[\left|\frac{1}{n}\imath_{X^n,Y^n}(X^n,Y^n) - H(\mathbf{X},\mathbf{Y})\right| > \frac{\delta}{2}\right]. \tag{271}$$

Since

$$\left|\frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n) - H(\mathbf{X}|\mathbf{Y})\right|$$

$$= \left|\frac{1}{n}\imath_{X^n,Y^n}(X^n,Y^n) - \frac{1}{n}\imath_{Y^n}(Y^n) - H(\mathbf{X},\mathbf{Y}) + H(\mathbf{Y})\right| \tag{272}$$

$$\leq \left|\frac{1}{n}\imath_{X^n,Y^n}(X^n,Y^n) - H(\mathbf{X},\mathbf{Y})\right| + \left|\frac{1}{n}\imath_{Y^n}(Y^n) - H(\mathbf{Y})\right| \tag{273}$$

it follows that

$$\mathcal{E}_n \subset \mathcal{E}_n^1 \cup \mathcal{E}_n^2. \tag{274}$$

Then

$$\mathbb{P}[\mathcal{E}_n] \leq \mathbb{P}[\mathcal{E}_n^1] + \mathbb{P}[\mathcal{E}_n^2] \tag{275}$$

by the union bound. Finally,

$$\lim_{n\to\infty} \mathbb{P}[\mathcal{E}_n] \leq \lim_{n\to\infty} \mathbb{P}[\mathcal{E}_n^1] + \lim_{n\to\infty} \mathbb{P}[\mathcal{E}_n^2] = 0. \tag{276}$$

follows by applying (193) to each term.

E. Proof of Theorem 22

Consider a sequence of (M_n, nd, ϵ_n)-lossy source codes such that

$$\limsup_{n\to\infty} \frac{1}{n} \log M_n < H(\mathbf{X}|\mathbf{Y}) - d \tag{277}$$

and let

$$\delta = \frac{1}{2}\left(H(\mathbf{X}|\mathbf{Y}) - \limsup_{n} \frac{1}{n} \log M_n - d\right), \tag{278}$$

$\gamma_n = n\delta$. Applying Theorem 17 we obtain

$$\epsilon^*_{X^n|Y^n}(M_n, d)$$

$$\geq \mathbb{P}\left[\imath_{X^n|Y^n}(X^n|Y^n) > nd + \log M_n + \gamma_n\right] - \exp(-\gamma_n) \tag{279}$$

$$= \mathbb{P}\left[\frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n) > d + \frac{1}{n}\log M_n + \delta\right] - \exp(-n\delta) \tag{280}$$

$$\geq \mathbb{P}\left[\frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n) > H(\mathbf{X}|\mathbf{Y}) - \delta\right] - \exp(-n\delta) \tag{281}$$

where (281) holds for infinitely many n. Then, by Theorem 20

$$\limsup_{n\to\infty} \epsilon^*_{X^n|Y^n}(M_n, d)$$

$$\geq \limsup_{n\to\infty} \mathbb{P}\left[\frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n) > H(\mathbf{X}|\mathbf{Y}) - \delta\right]$$

$$- \exp(-n\delta) \tag{282}$$

$$\geq \lim_{n\to\infty} \mathbb{P}\left[\frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n) > H(\mathbf{X}|\mathbf{Y}) - \delta\right] - \exp(-n\delta) \tag{283}$$

$$= 1. \tag{284}$$

This shows (209) and $\mathcal{R}_m(d|\mathbf{X},\mathbf{Y}) \geq H(\mathbf{X}|\mathbf{Y}) - d$.

Fix any $\delta > 0$, $\gamma_n = n\delta$ and $d_n = nd$. Let M_n be such that

$$H(\mathbf{X}|\mathbf{Y}) - d + 2\delta \leq \frac{1}{n}\log M_n \leq H(\mathbf{X}|\mathbf{Y}) - d + 3\delta. \tag{285}$$

Note that an integer M_n satisfying (285) always exists for sufficiently large n. Applying Theorem 16 we see that there exists a sequence of (M_n, d_n, ϵ_n)-lossy source codes such that ϵ_n is upper bounded by

$$\leq \mathbb{P}\left[\imath_{X^n|Y^n}(X^n|Y^n) > nd + \log M_n - \gamma_n\right] + 2\exp(-\gamma_n) \tag{286}$$

$$= \mathbb{P}\left[\frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n) > d + \frac{1}{n}\log M_n - \delta\right] + 2\exp(-n\delta) \tag{287}$$

$$\leq \mathbb{P}\left[\frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n) > H(\mathbf{X}|\mathbf{Y}) + \delta\right] + 2\exp(-n\delta). \tag{288}$$

It follows by Theorem 20 that $\lim \epsilon_n = 0$ and $(H(\mathbf{X}|\mathbf{Y}) - d + 3\delta, d)$ is achievable for any $\delta > 0$. Therefore

$$\mathcal{R}_m(d|\mathbf{X},\mathbf{Y}) \leq H(\mathbf{X}|\mathbf{Y}) - d. \tag{289}$$

To see the result for average distortion consider a sequence of (M_n, d_n)-source codes with

$$\limsup \frac{1}{n}\log M_n < \mathcal{R}_a(d|\mathbf{X},\mathbf{Y}) = H(\mathbf{X}|\mathbf{Y}) - d. \tag{290}$$

Let $d' > d$ be given by

$$d' = d + \frac{H(\mathbf{X}|\mathbf{Y}) - d - \limsup \frac{1}{n}\log M_n}{2}. \tag{291}$$

We can also view this sequence of codes as an (M_n, d', ϵ_n)-source codes for some ϵ_n. By (209) $\limsup \epsilon_n = 1$ and therefore $\limsup d_n \geq \limsup \epsilon_n d' = d' > d$ which shows the lower bound on $\mathcal{R}_a(d|\mathbf{X},\mathbf{Y})$.

Finally,

$$\mathcal{R}_m(d|\mathbf{X},\mathbf{Y}) = \mathcal{R}_m(d|\mathbf{X},\mathbf{Y}) \tag{292}$$

follows from Theorem 16 and the next pair of lemmas.

Lemma 26. *Fix $\Delta > d$ and let (f,c) be an (M,d,ϵ)-lossy source code. Then, there exists an $(M,\tilde{d})$-lossy source code $(\tilde{\mathsf{f}},\tilde{\mathsf{c}})$, with*

$$\tilde{d} \leq d + \Delta\epsilon + \mathbb{E}\left[\imath_{X|Y}(X|Y)1\{\imath_{X|Y}(X|Y) > \Delta\}\right] + 2 + \epsilon. \tag{293}$$

Proof. Let $(\tilde{\mathsf{f}},\tilde{\mathsf{c}})$ be the lossy source code given by

$$\tilde{\mathsf{f}}(x) = \mathsf{f}(x) \tag{294}$$

and

$$\tilde{P}_{m,y}(x) = \frac{1}{2}P_{m,y}(x) + \frac{1}{2}P_{X|Y}(x|y) \tag{295}$$

where $P_{m,y} = \mathsf{c}(m,y)$ and $\tilde{P}_{m,y} = \tilde{\mathsf{c}}(m,y)$. Then

$$\mathsf{d}(x,\tilde{\mathsf{c}}(\tilde{\mathsf{f}}(x),y)) \leq \min\left\{\mathsf{d}(x,\mathsf{c}(\mathsf{f}(x),y)), \imath_{X|Y}(x|y)\right\} + 1. \tag{296}$$

Define

$$Z = \mathsf{d}(x,\mathsf{c}(\mathsf{f}(X),Y)), \tag{297}$$

$$\tilde{Z} = \mathsf{d}(x,\tilde{\mathsf{c}}(\tilde{\mathsf{f}}(X),Y)). \tag{298}$$

The bound follows since

$$\begin{aligned}
&\mathbb{E}\left[\mathsf{d}(X,\tilde{\mathsf{c}}(\tilde{\mathsf{f}}(X),Y))\right] \\
&= \mathbb{E}\left[1\{\tilde{Z} \leq d+1\}\tilde{Z}\right] + \mathbb{E}\left[1\{d+1 < \tilde{Z} \leq \Delta+1\}\tilde{Z}\right] \\
&+ \mathbb{E}\left[1\{\Delta+1 < \tilde{Z}\}\tilde{Z}\right] && (299)\\
&\leq d + \mathbb{E}\left[1\{d < Z\}1\{\tilde{Z} \leq \Delta+1\}\tilde{Z}\right] \\
&+ \mathbb{E}\left[1\{\Delta < \imath_{X|Y}(X|Y)\}\tilde{Z}\right] + 1 && (300)\\
&\leq d + \epsilon(\Delta+1) \\
&+ \mathbb{E}\left[1\{\Delta < \imath_{X|Y}(X|Y)\}(\imath_{X|Y}(X|Y)+1)\right] + 1 && (301)\\
&\leq d + \epsilon\Delta + \mathbb{E}\left[1\{\Delta < \imath_{X|Y}(X|Y)\}\imath_{X|Y}(X|Y)\right] + 2 + \epsilon. && (302)
\end{aligned}$$

□

Lemma 27. *Let $\mathbf{X}$ be a stationary ergodic source with stationary ergodic side information process $\mathbf{Y}$ and suppose $\Delta > H(\mathbf{X}|\mathbf{Y})$. Then,*

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n)1\left\{\frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n) > \Delta\right\}\right] = 0. \tag{303}$$

Proof. Pick any $\gamma > 0$ and let $\Delta_\gamma = H(\mathbf{X}|\mathbf{Y}) - \frac{\gamma}{2}$. Then,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n)1\left\{\frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n) > \Delta\right\}\right] && (304)\\
&= H(\mathbf{X}|\mathbf{Y}) \\
&- \mathbb{E}\left[\frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n)1\left\{\frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n) \leq \Delta\right\}\right] && (305)\\
&\leq H(\mathbf{X}|\mathbf{Y}) \\
&- \mathbb{E}\left[\frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n)1\left\{\Delta_\gamma < \frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n) \leq \Delta\right\}\right] && (306)\\
&\leq H(\mathbf{X}|\mathbf{Y}) \\
&- \Delta_\gamma\mathbb{P}\left[\Delta_\gamma < \frac{1}{n}\imath_{X^n|Y^n}(X^n|Y^n) \leq \Delta\right] && (307)\\
&\leq_n \gamma && (308)
\end{aligned}$$

where the last line holds for n sufficiently large since

$$\lim_{n\to\infty} \mathbb{P}\left[\Delta_\gamma < \imath_{X^n|Y^n}(X^n|Y^n) \leq \Delta\right] = 1 \tag{309}$$

by Theorem 20. □

Combining Lemmas 26 and 27 with Theorem 16 shows the achievability result for average distortion.

F. Proof of Theorem 23

To show achievability fix any $\delta > 0$, $\gamma_n = n\delta$ and $d_{n,i} = nd_i$. Let $M_{n,i}$ be such that

$$H(\mathbf{X}) - d_i + 2\delta \leq \frac{1}{n}\log M_{n,i} \leq H(\mathbf{X}|\mathbf{Y}) - d_i + 3\delta. \tag{310}$$

Note that an integer $M_{n,i}$ satisfying (310) always exists for sufficiently large n. Applying Theorem 18 we see that there exists a sequence of $\left((M_{n,i})_{i=1}^2, (d_{n,i})_{i=0}^2, \epsilon\right)$-lossy source codes with

$$\begin{aligned}
\epsilon_n &\leq \mathbb{P}\left[\frac{1}{n}\imath_{X^n}(X^n) > \frac{1}{n}K_{\gamma_n}\right] + 6\exp(-n\delta) && (311)\\
&\leq \mathbb{P}\left[\frac{1}{n}\imath_{X^n}(X^n) > H(\mathbf{X}) + \delta\right] + 6\exp(-n\delta) && (312)
\end{aligned}$$

It follows by Theorem 20 that $\epsilon_n \to 0$ and thus the claimed region is achievable under the maximum error criterion.

The converse follows trivially by Corollary 19 and applying the same sequence of steps as in Theorem 22 to each term on the right hand side of (190).

The region $\mathcal{R}_{\mathrm{a}}(d_0, d_1, d_2|\mathbf{X})$ follows similarly to Theorem 22 where the average distortion achievability is connected to excess distortion achievability through an analogue of Lemma 26.

REFERENCES

[1] T. Courtade and R. Wesel, "Multiterminal source coding with an entropy-based distortion measure," in *IEEE International Symposium on Information Theory*, St. Petersburg, Russia, July 2011, pp. 2040–2044.

[2] T. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 740–761, Jan 2014.

[3] D. Haussler and M. Opper, "Mutual information, metric entropy and cumulative relative entropy risk," *Ann. Statist.*, vol. 25, no. 6, pp. 2451–2492, 12 1997.

[4] B. Nazer, O. Ordentlich, and Y. Polyanskiy, "Information-distilling quantizers," 2017, preprint.

[5] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3309–3338, June 2012.

[6] E. Song, "A new approach to lossy compression and applications to security," Ph.D. dissertation, Princeton University, 2015.

[7] I. Kontoyiannis and S. Verdú, "Optimal lossless data compression: Non-asymptotics and asymptotics," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 777–795, Feb 2014.

[8] H. Yagi and R. Nomura, "Variable-length coding with epsilon-fidelity criteria for general sources," in *IEEE International Symposium on Information Theory*, Hong-Kong, June 2015, pp. 2181–2185.

[9] 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.

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

[11] S. Verdú, "Non-asymptotic achievability bounds in multiuser information theory," in *50th Annual Allerton Conference on Communication, Control, and Computing*, Oct 2012, pp. 1–8.

[12] S. Watanabe, S. Kuzuoka, and V. Tan, "Nonasymptotic and second-order achievability bounds for coding with side-information," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1574–1605, April 2015.

[13] V. Kostina and S. Verdú, "A new converse in rate-distortion theory," in *46th Annual Conference on Information Sciences and Systems*, Princeton, March 2012, pp. 1–6.

[14] K. Iwata and J. Muramatsu, "An information-spectrum approach to rate-distortion function with side information," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E85-A, no. 6, pp. 1387–1395, 2002.

[15] S. Jalali and T. Weissman, "Multiple description coding of discrete ergodic sources," in *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, Sept 2009, pp. 1256–1261.

[16] J. Chen, C. Tian, and S. Diggavi, "Multiple description coding for stationary and ergodic sources," in *Data Compression Conference, 2007.*, March 2007, pp. 73–82.

[17] H. Wang and P. Viswanath, "Vector gaussian multiple description with two levels of receivers," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 401–410, Jan 2009.

[18] M. Fleming and M. Effros, "The rate distortion region for the multiple description problem," in *IEEE International Symposium on Information Theory*, Sorrento, Italy, 2000, pp. 208–.

[19] T. Courtade, "Two problems in multiterminal information theory," Ph.D. dissertation, University of California Los Angeles, 2012.

[20] Y. Shkel, M. Raginsky, and S. Verdú, "Universal lossy compression under logarithmic loss," in *IEEE International Symposium on Information Theory*, 2017.

[21] G. Hardy, J. Littlewood, and G. Pólya, *Inequalities*, ser. Cambridge Mathematical Library. Cambridge University Press, 1952.