Error exponents for source coding under logarithmic loss

Error exponents for source coding under logarithmic loss

Hamdi Joudeh and Han Wu
ICT Lab, Eindhoven University of Technology, The Netherlands

***Abstract*—In source coding under the logarithmic loss distortion measure, a source is compressed into a message, which is then decompressed into a soft reconstruction (i.e. probability distribution). The distortion is measured by the remaining uncertainty about the source given the message. Shkel and Verdú showed that this lossy source coding setting is intimately related to almost lossless source coding with list decoding, and used this insight to characterize the single-shot excess distortion error probability. In this work, we build upon this connection to list decoding and derive error exponents for source coding under logarithmic loss, without and with side information. The error exponents are closely related to their almost lossless counterparts.**

I. Introduction

In this paper we study the problem of fixed-length lossy source coding of a discrete memoryless source (DMS) under the logarithmic loss (log-loss) distortion measure. While the log-loss is most commonly used in prediction and learning theory, its adoption as a distortion measure in lossy source coding is also natural, specifically in settings where the decoder produces a soft reconstruction (i.e. probability distribution) of the source instead of a point estimate [1]–[3].

The log-loss distortion measure enjoys some mathematical properties that enable elegant characterizations in a number of settings. For instance, under an average distortion criterion, the rate-distortion function is given by [1, Example 2]

$$R(\Delta) = H(X) - \Delta \tag{1}$$

where $H(X)$ is the source entropy and Δ is the average log-loss distortion (assume $0 \leq \Delta \leq H(X)$). The converse for the corresponding coding theorem is obtained by bounding the average log-loss distortion using the conditional entropy of the source given its reconstruction. By building upon this property, Courtade and Weissman [2] derived tight outer bounds in various multi-terminal source coding settings under average log-loss distortion (see also Courtade and Wesel [1]).

More recently, Shkel and Verdú [3] derived single-shot bounds under both excess and average log-loss distortion criteria, without and with decoder side information (see [4], [5] for universal extensions). Key to their approach is a close connection between the log-loss setting and the almost lossless setting with list decoding. As we shall see, this connection to list decoding also plays a central role in our current work.

In this paper, instead of single-shot bounds, we are interested in error exponents under an excess log-loss distortion criterion. We derive error exponents without and with side information, while mainly focusing on universal schemes. We also demonstrate close connections to results in almost lossless settings.

This work was partially supported by the European Research Council (ERC) under the ERC Starting Grant N. 101116550 (IT-JCAS).

Notation: We use standard notation, briefly explained here. $\mathcal{P}(\mathcal{X})$ denotes the probability simplex on a finite alphabet $\mathcal{X}$. For a probability mass function (pmf) $P_X \in \mathcal{P}(\mathcal{X})$, we denote its support by $\mathcal{S}(P_X)$ and its entropy by $H(P_X)$. The relative entropy between two pmfs Q_X and P_X is denoted by $D(Q_X\|P_X)$. For (X, Y) with joint pmf $P_{XY} = P_{X|Y}P_Y$, the conditional entropy of X given Y is denoted by $H(P_{X|Y}|P_Y)$. The type of a sequence $\boldsymbol{x} \in \mathcal{X}^n$ is denoted by $P_{\boldsymbol{x}}$, and $\mathcal{P}_n(\mathcal{X})$ is the set of all types of sequences in $\mathcal{X}^n$. For $Q \in \mathcal{P}_n(\mathcal{X})$, the corresponding type class is denoted by $\mathcal{T}_n(Q)$. Given a second sequence $\boldsymbol{y} \in \mathcal{Y}^n$, $P_{\boldsymbol{xy}}$ and $P_{\boldsymbol{x}|\boldsymbol{y}}$ are the joint and conditional types. $\mathcal{P}_n(\mathcal{X} \times \mathcal{Y})$ and $\mathcal{P}_n(\mathcal{X}|\mathcal{Y})$ are the sets of all joint and conditional types. $\mathcal{T}_n(Q_{X|Y}|\boldsymbol{y})$ is the conditional type class of $Q_{X|Y} \in \mathcal{P}_n(\mathcal{X}|\mathcal{Y})$ given $\boldsymbol{y}$. We will make use of types and type classes and their basic properties, such as cardinality and probability bounds (see, e.g., [6, Ch.2]).

II. Source Coding Under Log-Loss

Consider a DMS with finite alphabet $\mathcal{X}$ that randomly generates i.i.d. source sequences $\boldsymbol{X} \triangleq (X_1, X_2, \ldots, X_n)$ according to a pmf $P_X \in \mathcal{P}(\mathcal{X})$. We use $\boldsymbol{x} \triangleq (x_1, x_2, \ldots, x_n)$ to denote a realization of $\boldsymbol{X}$. A *soft reconstruction* of $\boldsymbol{x}$ is a member of $\mathcal{P}(\mathcal{X}^n)$, i.e. a distribution on $\mathcal{X}^n$, denoted by $\hat{P}_n$. The log-loss distortion between $\boldsymbol{x}$ and $\hat{P}_n$ is defined as

$$\mathsf{d}(\boldsymbol{x}, \hat{P}_n) \triangleq \log \frac{1}{\hat{P}_n(\boldsymbol{x})}. \tag{2}$$

The log-loss, also referred to as the self-information loss, can be understood as the remaining uncertainty about $\boldsymbol{x}$ given its reconstruction $\hat{P}_n$ [1]–[3]. For instance, $\mathsf{d}(\boldsymbol{x}, \hat{P}_n)$ is zero if and only if $\hat{P}_n$ has a single mass point at $\boldsymbol{x}$, i.e. an exact *hard reconstruction*; and infinite whenever $\boldsymbol{x}$ has zero probability under $\hat{P}_n$. For convenience, we work with the normalized (per-symbol) log-loss defined as $\mathsf{d}_n(\boldsymbol{x}, \hat{P}_n) \triangleq \frac{1}{n}\mathsf{d}(\boldsymbol{x}, \hat{P}_n)$.

In the lossy source coding setting considered in this work, the sequence $\boldsymbol{X}$ is encoded into a message index from the finite set $\mathcal{M}_n$, which is then decoded into a soft reconstruction from $\mathcal{P}(\mathcal{X}^n)$. A lossy source code of block-length n is thus a pair of mappings $\phi_n : \mathcal{X}^n \to \mathcal{M}_n$ and $\varphi_n : \mathcal{M}_n \to \mathcal{P}(\mathcal{X}^n)$, referred to as the encoder and decoder respectively.

For a lossy source code (ϕ_n, φ_n), the code *rate* is given by $\frac{1}{n}\log|\mathcal{M}_n|$, while $\mathbb{P}\left[\mathsf{d}_n\big(\boldsymbol{X}, \varphi_n\left(\phi_n(\boldsymbol{X})\right)\big) > \Delta\right]$ is the excess distortion *error probability* for some distortion level $\Delta \geq 0$. We say that (ϕ_n, φ_n) is an (n, R, Δ, ϵ)-code if

$$\frac{1}{n}\log|\mathcal{M}_n| \leq R \quad \text{and} \quad \mathbb{P}\left[\mathsf{d}_n\big(\boldsymbol{X}, \varphi_n\left(\phi_n(\boldsymbol{X})\right)\big) > \Delta\right] \leq \epsilon.$$

The minimal error probability for fixed (n, R, Δ) is defined as

$$\varepsilon(n, R, \Delta) \triangleq \inf \{\epsilon : \text{there exists an } (n, R, \Delta, \epsilon)\text{-code}\}.$$

We are interested in characterizing the asymptotic behaviour of $\varepsilon(n, R, \Delta)$, captured through the error exponent defined as

$$E(R, \Delta) \triangleq \lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\varepsilon(n, R, \Delta)}. \tag{3}$$

Remark 1. The above problem does not fall under the umbrella of standard lossy source coding in discrete memoryless settings [6], [7]. In the standard paradigm, the reconstruction is a sequence drawn from a discrete product alphabet; and the distortion is additive, i.e. a normalized sum of single-letter distortions. In the setting considered here, the reconstruction alphabet $\mathcal{P}(\mathcal{X}^n)$ is not a product alphabet and is not discrete; and the distortion measure is not additive. For $\mathsf{d}_n(\boldsymbol{x}, \hat{P}_n)$ to be additive, the soft reconstruction $\hat{P}_n$ must be a product distribution, as in earlier works on log-loss source coding [1], [2]. This need not be the case in general, and as we shall see, the codes we propose employ non-product soft reconstructions.

A. Connection to list decoding

In [3], Shkel and Verdú established a fundamental connection between lossy source coding under log-loss and almost lossless source coding with list decoding, leading to an exact characterization of $\varepsilon(n, R, \Delta)$. This connection is central to the approach we take here, therefore, we review it in some detail. We start with a key lemma linking the log-loss of a soft reconstruction to the list size in list decoding.

To this end, fix a soft reconstruction $\hat{P}_n \in \mathcal{P}(\mathcal{X}^n)$ and a distortion level $\Delta \geq 0$. We say that a sequence $\boldsymbol{x} \in \mathcal{X}^n$ is Δ-covered by $\hat{P}_n$ if $\mathsf{d}_n(\boldsymbol{x}, \hat{P}_n) \leq \Delta$. If $\hat{P}_n$ Δ-covers every element of a set (or list) $\mathcal{L}_n \subseteq \mathcal{X}^n$, then the set $\mathcal{L}_n$ is also said to be Δ-covered by the soft reconstruction $\hat{P}_n$.

Lemma 1. *Let $\mathcal{L}_n \subseteq \mathcal{X}^n$. There exists a soft-reconstruction $\hat{P}_n \in \mathcal{P}(\mathcal{X}^n)$ that Δ-covers $\mathcal{L}_n$ if and only if*

$$|\mathcal{L}_n| \leq \lfloor \exp(n\Delta) \rfloor. \tag{4}$$

Proof. The direct part holds by taking $\hat{P}_n$ to be uniform on $\mathcal{L}_n$ and zero elsewhere. The converse part follows from [3, Lemma 1], reproduced here for completeness. Let $\mathcal{B}_n(\Delta, \hat{P}_n)$ be the set of all source sequences Δ-covered by $\hat{P}_n$, i.e.

$$\mathcal{B}_n(\Delta, \hat{P}_n) \triangleq \left\{ \boldsymbol{x} \in \mathcal{X}^n : \mathsf{d}_n(\boldsymbol{x}, \hat{P}_n) \leq \Delta \right\}. \tag{5}$$

It is sufficient to show $|\mathcal{B}_n(\Delta, \hat{P}_n)| \leq \lfloor \exp(n\Delta) \rfloor$. Note that $\boldsymbol{x} \in \mathcal{B}_n(\Delta, \hat{P}_n)$ implies $\hat{P}_n(\boldsymbol{x}) \geq \exp(-n\Delta)$, and therefore

$$1 = \sum_{\boldsymbol{x}\in\mathcal{X}^n} \hat{P}_n(\boldsymbol{x}) \geq \sum_{\boldsymbol{x}\in\mathcal{B}_n(\Delta,\hat{P}_n)} \hat{P}_n(\boldsymbol{x}) \geq \left|\mathcal{B}_n(\Delta, \hat{P}_n)\right| \exp(-n\Delta).$$

The bound is tightened by including the floor function. □

Now consider an almost lossless list source code: here a source sequence is encoded into one of $\lfloor \exp(nR) \rfloor$ message indices, and a message index is decoded into a list of $\lfloor \exp(n\Delta) \rfloor$ source sequences. A decoding error occurs if the generated source sequence is not in the decoded list. From Lemma 1, we see that such a code with error probability ϵ can be converted into a (n, R, Δ, ϵ) log-loss source code. Conversely, a log-loss source code can be converted into an almost lossless list source code. This connection leads to the following result.

Theorem. (Shkel-Verdú [3, Theorem 5-6]). *Let $G : \mathcal{X}^n \to \{1, 2, \ldots, |\mathcal{X}^n|\}$ be a a probability rank function that ranks source sequences in decreasing order of their probability. Then*

$$\varepsilon(n, R, \Delta) = \mathbb{P}\left[G(\boldsymbol{X}) > \lfloor \exp(nR) \rfloor \lfloor \exp(n\Delta) \rfloor\right]. \tag{6}$$

III. Error Exponent

We now present the first result of this paper. To this end, we first define the function $F(R)$ on $0 \leq R < \log|\mathcal{S}(P_X)|$ as

$$F(R) \triangleq \min_{Q_X : H(Q_X) \geq R} D(Q_X \| P_X). \tag{7}$$

Note that $F(R) = E(R, 0)$, which is the error exponent in the almost lossless case [6], [8], [9]. $F(R)$ is continuous, convex and increasing on its domain, with $F(R) = 0$ on $0 \leq R \leq H(P_X)$. The expression in (7) is known as a *primal form*. $F(R)$ admits an equivalent *dual form* given as [6, Prob.2.15]

$$F(R) = \sup_{\rho \geq 0} \rho \left(R - H_{\frac{1}{1+\rho}}(X) \right) \tag{8}$$

where $H_{\frac{1}{1+\rho}}(X)$ is the Rényi entropy of order $1/(1+\rho)$.

Recall that the log-loss rate-distortion function is given by $R(\Delta) = H(P_X) - \Delta$, and to ensure that $\varepsilon(n, R, \Delta)$ goes to zero, we must have $R > H(P_X) - \Delta$. On the other hand, for rates satisfying $R \geq \log|\mathcal{S}(P_X)| - \Delta$, the whole support of P_X^n can be covered by lists of size $\mathrm{e}^{n\Delta}$. Therefore, the relevant range is $H(P_X) < R + \Delta < \log|\mathcal{S}(P_X)|$.

Theorem 1. *Let (R, Δ) be a rate-distortion pair such that $H(P_X) < R + \Delta < \log|\mathcal{S}(P_X)|$. Then*

$$E(R, \Delta) = F(R + \Delta). \tag{9}$$

We observe that $E(R, \Delta) = E(R + \Delta, 0)$, i.e. the log-loss error exponent as a function of R is merely a translation of the almost lossless error exponent by Δ bits to the left. This can be understood in light of the optimal source code that achieves (6) as follows. For a log-loss source code of rate R and distortion Δ, an excess distortion error occurs when the source generates a sequence with probability rank greater than $\lfloor \mathrm{e}^{nR} \rfloor \lfloor \mathrm{e}^{n\Delta} \rfloor$, that is the number of sequences covered by $\lfloor \mathrm{e}^{nR} \rfloor$ lists of size $\lfloor \mathrm{e}^{n\Delta} \rfloor$ each. On the other hand, an almost lossless source code of rate $R + \Delta$ makes an error when the source generates a sequence with probability rank greater than $\lfloor \mathrm{e}^{n(R+\Delta)} \rfloor$. Asymptotically, the two error events have almost the same probability, yielding in the same error exponent.

The above argument is sufficient for proving Theorem 1, yet it employs a code that depends on the source pmf (or probability rank function). Further on we present an alternative proof using the method of types, extending the Longo-Sgarro approach [9] (see also Csiszár-Körner [6]) from the almost lossless case to the log-loss case. As is often the case with types-based proofs, a universal scheme emerges.

It is worthwhile mentioning that the log-loss error exponent expression in (9) can be obtained from Marton's error exponent [7] by replacing the general rate-distortion function with its log-loss counterpart. While this is perhaps expected, the result in Theorem 1 does not follow directly from Marton's proof, at least not without modification, as the log-loss setting considered here is not a special case of the classical rate-distortion setting (see Remark 1). We shall see next that Theorem 1 is proved directly using the connection to list decoding.

A. Proof of Theorem 1

1) Achievability: Fix $R > 0$ and $n \in \mathbb{N}$, and let $J_n = |\mathcal{P}_n(\mathcal{X})|$ and $M_n = \left\lfloor (1+n)^{-|\mathcal{X}|} \mathrm{e}^{nR} \right\rfloor$. For every type $Q \in \mathcal{P}_n(\mathcal{X})$, partition $\mathcal{T}_n(Q)$ into M_n lists all roughly of the same size. $\mathcal{L}_n(\boldsymbol{x})$ denotes the list containing $\boldsymbol{x}$. By construction,

$$|\mathcal{L}_n(\boldsymbol{x})| \leq \left\lceil \frac{|\mathcal{T}_n(P_{\boldsymbol{x}})|}{M_n} \right\rceil \tag{10}$$

for every $\boldsymbol{x} \in \mathcal{X}^n$. An emitted source sequence $\boldsymbol{x}$ is encoded into $\phi_n(\boldsymbol{x}) = \big(t_n(\boldsymbol{x}), l_n(\boldsymbol{x})\big)$, where $t_n(\boldsymbol{x}) \in \{1, 2, \ldots, J_n\}$ is the type index and $l_n(\boldsymbol{x}) \in \{1, 2, \ldots, M_n\}$ is the list index. Upon receiving $\phi_n(\boldsymbol{x})$, the decoder reproduces the list $\mathcal{L}_n(\boldsymbol{x})$ containing $\boldsymbol{x}$. The corresponding soft reconstruction $\varphi_n(\phi_n(\boldsymbol{x})) = \hat{P}_n\big(\,\cdot\,|\phi_n(\boldsymbol{x})\big)$ is set as

$$\hat{P}_n\big(\hat{\boldsymbol{x}}|\phi_n(\boldsymbol{x})\big) = \begin{cases} \frac{1}{|\mathcal{L}_n(\boldsymbol{x})|}, & \hat{\boldsymbol{x}} \in \mathcal{L}_n(\boldsymbol{x}) \\ 0, & \text{otherwise.} \end{cases}$$

The rate of this code satisfies $\frac{1}{n}\log\left(J_n M_n\right) \leq R$. For any sequence $\boldsymbol{x} \in \mathcal{X}^n$, the log-loss incurred by the corresponding reconstruction $\varphi_n(\phi_n(\boldsymbol{x}))$ is bounded above as

$$\begin{aligned} \mathsf{d}_n\big(\boldsymbol{x}, \varphi_n(\phi_n(\boldsymbol{x}))\big) &= \frac{1}{n}\log|\mathcal{L}_n(\boldsymbol{x})| \\ &\leq \frac{1}{n}\log\left\lceil \frac{\mathrm{e}^{nH(P_{\boldsymbol{x}})}}{\left\lfloor \mathrm{e}^{nR-|\mathcal{X}|\log(1+n)} \right\rfloor} \right\rceil \qquad (11) \\ &\leq H(P_{\boldsymbol{x}}) - R + \delta_n \qquad (12) \end{aligned}$$

where $\delta_n \geq 0$ goes to zero as n grows large.

We now analyze the error probability. To this end, fix $\Delta \geq 0$ such that $H(P_X) < R + \Delta < \log|\mathcal{S}(P_X)|$. We can see from (12) that all source sequences in the set $\mathcal{B}_n$, defined as

$$\mathcal{B}_n = \bigcup_{Q \in \mathcal{P}_n(\mathcal{X}) : H(Q) \leq R+\Delta-\delta_n} \mathcal{T}_n(Q), \tag{13}$$

are reconstructed with a log-loss not exceeding Δ, and thus the excess distortion error event is included in $\mathcal{X}^n \setminus \mathcal{B}_n$. The error probability under source pmf P_X is bounded above as

$$\begin{aligned} P_X^n(\mathcal{X}^n \setminus \mathcal{B}_n) &= \sum_{Q \in \mathcal{P}_n(\mathcal{X}) : H(Q) > R+\Delta-\delta_n} P_X^n\big(\mathcal{T}_n(Q)\big) \\ &\leq \sum_{Q \in \mathcal{P}_n(\mathcal{X}) : H(Q) > R+\Delta-\delta_n} \mathrm{e}^{-nD(Q\|P_X)} \qquad (14) \\ &\leq (n+1)^{|\mathcal{X}|} \max_{Q \in \mathcal{P}_n(\mathcal{X}) : H(Q) > R+\Delta-\delta_n} \mathrm{e}^{-nD(Q\|P_X)} \qquad (15) \\ &\leq (n+1)^{|\mathcal{X}|} \mathrm{e}^{-nF(R+\Delta-\delta_n)}. \qquad (16) \end{aligned}$$

The above steps are standard and use the properties of types and type classes. Achievability follows from (16) and the continuity of $F(R')$ on $H(P_X) < R' < \log|\mathcal{S}(P_X)|$.

Remark 2. As mentioned earlier, the above source code is universal and does not depend on P_X. The code is also universal with respect to the distortion level Δ, and only depends on the rate R (and block-length n). For fixed R, the same sequence of codes achieves a positive exponent for every P_X and Δ satisfying $R < H(P_X) - \Delta$. The key to the universality with respect to Δ is the *variable list partitioning* of type classes, where list sizes depend on the type and code rate but not on the distortion level (see (10)). The same partitioning is used by Bunte and Lapidoth in [10] in the context of strictly lossless list source coding (also known as task encoding), where the focus is on analyzing list size moments.

Remark 3. In the achievability proof of Marton's error exponent [7], a key ingredient is the type covering lemma which states that for any type $Q \in \mathcal{P}_n(\mathcal{X})$ with a rate-distortion function satisfying $R(Q, \Delta) \leq R - \delta$, the corresponding type class $\mathcal{T}_n(Q)$ can be Δ-covered by e^{nR} reconstruction sequences. The type covering lemma is often proved using random selection (i.e. random coding). In the log-loss setting considered here, type covering is accomplished through simple partitioning, and the proof does not rely on random coding.

2) Converse: Fix a pair (R, Δ) and a source pmf P_X such that $H(P_X) < R + \Delta < \log|\mathcal{S}(P_X)|$. For every block-length n, let $(\phi_n^\star, \varphi_n^\star)$ be an optimal code achieving the minimal error probability $\varepsilon(n, R, \Delta)$. Moreover, define the set $\mathcal{B}_n^\star \triangleq \big\{\boldsymbol{x} \in \mathcal{X}^n : \mathsf{d}_n\left(\boldsymbol{x}, \varphi_n^\star\left(\phi_n^\star(\boldsymbol{x})\right)\right) \leq \Delta\big\}$. An error occurs whenever the source produces a sequence in $\mathcal{X}^n \setminus \mathcal{B}_n^\star$.

Let $M(\Delta, \mathcal{B}_n^\star)$ be the minimum number of soft reconstruction required to Δ-cover $\mathcal{B}_n^\star$. From Lemma 1, we know that any soft reconstruction can Δ-cover at most $\lfloor \mathrm{e}^{n\Delta} \rfloor$ source sequence. Therefore, we must have

$$M(\Delta, \mathcal{B}_n^\star) \geq \left\lceil \frac{|\mathcal{B}_n^\star|}{\lfloor \mathrm{e}^{n\Delta} \rfloor} \right\rceil. \tag{17}$$

It immediately follows that the rate R of $(\phi_n^\star, \varphi_n^\star)$ must satisfy

$$\mathrm{e}^{nR} \geq M(\Delta, \mathcal{B}_n^\star) \geq |\mathcal{B}_n^\star| \mathrm{e}^{-n\Delta}. \tag{18}$$

Now let $\delta_n = \frac{|\mathcal{X}|}{n}\log(1+n) + \frac{1}{n}\log 2$ and let $Q \in \mathcal{P}_n(\mathcal{X})$ be a type such that $H(Q) \geq R + \Delta + \delta_n$. The cardinality of the corresponding type class $\mathcal{T}_n(Q)$ is bounded below as

$$|\mathcal{T}_n(Q)| \geq (1+n)^{-|\mathcal{X}|} \mathrm{e}^{nH(Q)} \geq 2\mathrm{e}^{n(R+\Delta)} \geq 2|\mathcal{B}_n^\star|$$

from which we conclude that at least half of the sequences in $\mathcal{T}_n(Q)$ are not contained in $\mathcal{B}_n^\star$. Therefore

$$\begin{aligned} \varepsilon(n, R, \Delta) &= P_X^n\left(\mathcal{X}^n \setminus \mathcal{B}_n^\star\right) \geq \frac{1}{2} P_X^n\left(\mathcal{T}_n(Q)\right) \\ &\geq \frac{1}{2}(n+1)^{-|\mathcal{X}|} \mathrm{e}^{-nD(Q\|P_X)} = \mathrm{e}^{-nD(Q\|P_X) - n\delta_n} \qquad (19) \end{aligned}$$

obtained from the standard type class probability lower bound. This holds for all types satisfying $H(Q) \geq R + \Delta + \delta_n$, hence

$$-\frac{1}{n}\log\varepsilon(n, R, \Delta) \leq \delta_n + \min_{Q \in \mathcal{P}_n(\mathcal{X}) : H(Q) \geq R+\Delta+\delta_n} D(Q\|P_X)$$

$$= \delta_n + F_n(R + \Delta + \delta_n) \tag{20}$$

where $F_n(R')$ is defined as $F(R')$ in (7) except that the minimization is over types in $\mathcal{P}_n(\mathcal{X})$ instead of all pmfs in $\mathcal{P}(\mathcal{X})$. By definition, we know that $F(R') \leq F_n(R')$. In addition, it can be shown that $F_n(R') \leq F(R') + \delta'_n$ for some $\delta'_n > 0$ that goes to zero as n grows large.[1] By combining this with (20) and taking the limit, the converse result follows.

IV. Side Information

In this section we consider settings with side information. Here we have a pair of DMSs with finite alphabets $\mathcal{X}$ and $\mathcal{Y}$. The sources randomly generate an i.i.d. sequence of pairs $(\boldsymbol{X}, \boldsymbol{Y}) \triangleq \big((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\big)$ according to a joint pmf $P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. The goal remains to compress the sequence $\boldsymbol{X}$ and then decompress it into a soft reconstruction in $\mathcal{P}(\mathcal{X}^n)$, but now $\boldsymbol{Y}$ is available either at both the encoder and decoder sides, or at the decoder side only.

A. Encoder-decoder side information

For the case where the side information sequence is available at both the encoder and decoder sides, a lossy source code of block-length n is given by the pair of mappings $\phi_n : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{M}_n$ and $\varphi_n : \mathcal{M}_n \times \mathcal{Y}^n \to \mathcal{P}(\mathcal{X}^n)$. A (n, R, Δ, ϵ)-code, minimal error probability and error exponent are defined in a standard manner. The latter two are denoted by $\varepsilon_{X|Y}(n, R, \Delta)$ and $E_{X|Y}(R, \Delta)$ respectively.

This problem is very similar to its counterpart with no side information, i.e. given $\boldsymbol{Y} = \boldsymbol{y}$, the problem reduces to encoding and decoding a memoryless source (not necessarily i.i.d.) with distribution $P_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) = \prod_{i=1}^n P_{X|Y}(x_i|y_i)$. Nevertheless, it is still useful to characterize the error exponent in this case, as it provides an upper bound for the more interesting case with decoder side information only. To this end, we define $F_{X|Y}(R)$ on $0 \leq R < \log|\mathcal{S}(P_X)|$ as

$$F_{X|Y}(R) \triangleq \min_{Q_{XY} : H(Q_{X|Y}|Q_Y) \geq R} D(Q_{XY} \| P_{XY}). \tag{21}$$

$F_{X|Y}(R)$ is continuous, convex and increasing on its domain; and is zero for $0 \leq R \leq H(P_{X|Y}|P_Y)$. Moreover, we have $F_{X|Y}(R) = E_{X|Y}(R, 0)$, that is the error exponent in the almost lossless case. $F_{X|Y}(R)$ also admits the following dual form in terms of Arimoto's conditional Rényi entropy [11]

$$F_{X|Y}(R) = \sup_{\rho \geq 0} \rho \left(R - H_{\frac{1}{1+\rho}}(X|Y) \right) \tag{22}$$

obtained using Lagrangian duality techniques (see, e.g., the proof of [10, Equation (32)] by Bunte and Lapidoth).

Theorem 2. *Let (R, Δ) be a rate-distortion pair such that $H(P_{X|Y}|P_Y) < R + \Delta < \log|\mathcal{S}(P_X)|$. Then*

$$E_{X|Y}(R, \Delta) = F_{X|Y}(R + \Delta). \tag{23}$$

The proof of Theorem 2 (omitted for brevity) is very similar to that of Theorem 1, but relies on conditional types.

[1] By continuity and since $\bigcup_{n \in \mathbb{N}} \mathcal{P}_n(\mathcal{X})$ is dense in $\mathcal{P}(\mathcal{X})$ [9, Rem. 2].

B. Decoder side information: Wyner-Ziv

We now turn our attention to the case where the side information sequence $\boldsymbol{Y}$ is only available at the decoder. This is the Wyner-Ziv setting, specialized to the log-loss distortion measure. A lossy source code of block-length n here is given by the pair $\phi_n : \mathcal{X}^n \to \mathcal{M}_n$ and $\varphi_n : \mathcal{M}_n \times \mathcal{Y}^n \to \mathcal{P}(\mathcal{X}^n)$. The minimal error probability and error exponent are denoted by $\varepsilon^{\mathrm{WZ}}_{X|Y}(n, R, \Delta)$ and $E^{\mathrm{WZ}}_{X|Y}(R, \Delta)$ respectively.

Next, we observe that the encoder-decoder side information result in Theorem 2 provides the following converse bound

$$E^{\mathrm{WZ}}_{X|Y}(R, \Delta) \leq E^{\mathrm{sp}}_{X|Y}(R, \Delta) \triangleq F_{X|Y}(R + \Delta). \tag{24}$$

For $\Delta = 0$, the setting reduces to the Slepian-Wolf problem, and we denote the error exponent by $E^{\mathrm{SW}}_{X|Y}(R)$. The bound $E^{\mathrm{SW}}_{X|Y}(R) \leq F_{X|Y}(R)$, a special case of (24), was obtained by Gallager in [12] (see also Csiszár and Körner [13, Theorem 3]). This bound on the Slepian-Wolf error exponent is sometimes referred to as the sphere-packing exponent [14], due to close resemblance to the sphere-packing exponent in channel coding. Similarly, the bound in (24) can be thought of as a sphere-packing exponent for the log-loss Wyner-Ziv setting.

Next, we derive a lower bound for $E^{\mathrm{WZ}}_{X|Y}(R, \Delta)$. Define the function $\tilde{F}_{X|Y}(R)$ on $0 \leq R < \log|\mathcal{S}(P_X)|$ as

$$\tilde{F}_{X|Y}(R) \triangleq \min_{Q_{XY}} \left\{ D(Q_{XY} \| P_{XY}) + \big| R - H(Q_{X|Y}|Q_Y) \big|^+ \right\}$$

where $|a|^+ \triangleq \max\{0, a\}$. Note that $\tilde{F}_{X|Y}(R) \leq F_{X|Y}(R)$. Moreover, $\tilde{F}_{X|Y}(R)$ admits a dual form

$$\tilde{F}_{X|Y}(R) = \max_{\rho \in [0,1]} \rho \left(R - H_{\frac{1}{1+\rho}}(X|Y) \right). \tag{25}$$

$\tilde{F}_{X|Y}(R)$ is an achievable error exponent in the Slepian-Wolf setting [12], [13], referred to as the random-coding error exponent, as it is achieved through random coding (or binning) in close resemblance to the random-coding exponent in channel coding. A corresponding random-coding error exponent for the log-loss Wyner-Ziv problem is presented next.

Theorem 3. *Let (R, Δ) be a rate-distortion pair such that $H(P_{X|Y}|P_Y) < R + \Delta < \log|\mathcal{S}(P_X)|$. Then*

$$E^{\mathrm{WZ}}_{X|Y}(R, \Delta) \geq E^{\mathrm{r}}_{X|Y}(R, \Delta) \triangleq \tilde{F}_{X|Y}(R + \Delta).$$

For fixed Δ, the exponents $E^{\mathrm{r}}_{X|Y}(R, \Delta)$ and $E^{\mathrm{sp}}_{X|Y}(R, \Delta)$ coincide on $H(P_{X|Y}|P_Y) - \Delta < R \leq R_{\mathrm{cr}}$, where R_{cr} is the largest rate at which the convex curve $E^{\mathrm{sp}}_{X|Y}(R, \Delta)$, as a function of R, meets it supporting line of slope 1. Note that R_{cr} is reminiscent of the *critical rate* in channel coding. Above this rate, the two exponents differ in general.

C. Proof of Theorem 3

The proof is based on random binning and a list decoding variant of the universal minimum entropy decoding rule [13], [15]. In the analysis of this scheme, we use $H(\boldsymbol{x}|\boldsymbol{y})$ as a shorthand notation for the conditional entropy $H(P_{\boldsymbol{x}|\boldsymbol{y}}|P_{\boldsymbol{y}})$ calculated from the joint type $P_{\boldsymbol{x}\boldsymbol{y}} = P_{\boldsymbol{x}|\boldsymbol{y}} P_{\boldsymbol{y}}$.

Let $J_n = |\mathcal{P}_n(\mathcal{X})|$ and $M_n = \big\lfloor (1+n)^{-|\mathcal{X}|} \mathrm{e}^{nR} \big\rfloor$ for fixed R and n. A binning function $b_n : \mathcal{X}^n \to \{1, 2, \ldots, M_n\}$ is a mapping that assigns an index $b_n(\boldsymbol{x}) \in \{1, 2, \ldots, M_n\}$ to every source sequence $\boldsymbol{x} \in \mathcal{X}^n$. For a fixed bin assignment, determined by a given binning function b_n, the set of all source sequences with the same bin index as $\boldsymbol{x}$ is denoted by

$$\mathcal{B}_n(\boldsymbol{x}|b_n) \triangleq \{\hat{\boldsymbol{x}} \in \mathcal{X}^n : b_n(\hat{\boldsymbol{x}}) = b_n(\boldsymbol{x})\}.$$

Further on, we will analyze the error probability averaged over an ensemble of binning functions. To that end, we denote a random binning function by B_n, where b_n is a realization of B_n. We use Gallager's ensemble [12]: every source sequence is assigned a bin index uniformly at random; and bin assignments are pairwise independent across sequences.

Encoding: The generated source sequence $\boldsymbol{x}$ is encoded into $\phi_n(\boldsymbol{x}) = \big(t_n(\boldsymbol{x}), b_n(\boldsymbol{x})\big)$, where $t_n(\boldsymbol{x})$ is the type index and $b_n(\boldsymbol{x})$ is the bin index. The rate satisfies $\frac{1}{n}\log(J_n M_n) \le R$.

Decoding: Upon receiving $\phi_n(\boldsymbol{x})$, the decoder knows that $\boldsymbol{x}$ is in the set $\mathcal{T}_n(P_{\boldsymbol{x}}) \cap \mathcal{B}_n(\boldsymbol{x}|b_n)$. Now suppose that the side information sequence is equal to $\boldsymbol{y}$. For every sequence $\hat{\boldsymbol{x}} \in \mathcal{T}_n(P_{\boldsymbol{x}}) \cap \mathcal{B}_n(\boldsymbol{x}|b_n)$, the decoder computes the conditional entropy $H(\hat{\boldsymbol{x}}|\boldsymbol{y})$ and produces a list $\mathcal{L}_n(\phi_n(\boldsymbol{x}), \boldsymbol{y})$ of size

$$|\mathcal{L}_n(\phi_n(\boldsymbol{x}), \boldsymbol{y})| = \min\left\{ \big\lfloor \mathrm{e}^{n\Delta} \big\rfloor, |\mathcal{T}_n(P_{\boldsymbol{x}}) \cap \mathcal{B}_n(\boldsymbol{x}|b_n)| \right\}$$

comprising source sequences with the lowest conditional entropy. The soft reconstruction is taken to be uniformly supported on $\mathcal{L}_n(\phi_n(\boldsymbol{x}), \boldsymbol{y})$. It is clear that an excess distortion error occurs if $\mathcal{L}_n(\phi_n(\boldsymbol{x}), \boldsymbol{y})$ does not include the encoded source sequence $\boldsymbol{x}$. Moreover, by setting $\Delta = 0$, we recover the classical minimum entropy decoder.

Error probability: Let $\mathcal{E}_n(\boldsymbol{x}|\boldsymbol{y}, b_n)$ denote the set of all source sequences other than $\boldsymbol{x}$, but with the same type and bin as $\boldsymbol{x}$, and a conditional entropy smaller than or equal to that of $\boldsymbol{x}$ given $\boldsymbol{y}$. For an excess distortion error to occur, we must have $|\mathcal{E}(\boldsymbol{x}|\boldsymbol{y}, b_n)| \ge \big\lfloor \mathrm{e}^{n\Delta} \big\rfloor = \mathrm{e}^{n(\Delta - \delta_n)}$, where $\delta_n \ge 0$ goes to zero as n grows large. The excess distortion error probability, averaged over B_n, is hence bounded above by

$$\mathbb{P}\left[|\mathcal{E}_n(\boldsymbol{X}|\boldsymbol{Y}, B_n)| \ge \mathrm{e}^{n(\Delta-\delta_n)}\right] \le \sum_{\boldsymbol{x},\boldsymbol{y}} P_{XY}^n(\boldsymbol{x}, \boldsymbol{y}) \min\left\{1, \mathrm{e}^{-n(\Delta-\delta_n)} \mathbb{E}\big[|\mathcal{E}_n(\boldsymbol{x}|\boldsymbol{y}, B_n)|\big]\right\} \quad (26)$$

which follows from Markov's inequality combined with the trivial upper bound of 1. We now take a small detour to bound $\mathbb{E}\left[|\mathcal{E}_n(\boldsymbol{x}|\boldsymbol{y}, B_n)|\right]$. To this end, define the set $\mathcal{E}_n'(\boldsymbol{x}|\boldsymbol{y}) \triangleq \big\{\hat{\boldsymbol{x}} \in \mathcal{T}_n(P_{\boldsymbol{x}}) : \hat{\boldsymbol{x}} \ne \boldsymbol{x}, H\big(\hat{\boldsymbol{x}}|\boldsymbol{y}\big) \le H\big(\boldsymbol{x}|\boldsymbol{y}\big)\big\}$, and observe that

$$\mathbb{E}\left[|\mathcal{E}_n(\boldsymbol{x}|\boldsymbol{y}, B_n)|\right] = \sum_{\hat{\boldsymbol{x}} \in \mathcal{E}_n'(\boldsymbol{x}|\boldsymbol{y})} \mathbb{P}\left[B_n(\hat{\boldsymbol{x}}) = B_n(\boldsymbol{x})\right] = \frac{|\mathcal{E}'(\boldsymbol{x}|\boldsymbol{y})|}{M_n}$$

which follows from uniform pairwise independent bin assignment. Next, we note that

$$|\mathcal{E}_n'(\boldsymbol{x}|\boldsymbol{y})| \le \sum_{Q_{X|Y} \in \mathcal{P}_n(\mathcal{X}|\mathcal{Y}) : H(Q_{X|Y}|P_{\boldsymbol{y}}) \le H(\boldsymbol{x}|\boldsymbol{y})} |\mathcal{T}_n(Q_{X|Y}|\boldsymbol{y})| \le (n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|} \mathrm{e}^{nH(\boldsymbol{x}|\boldsymbol{y})}. \quad (27)$$

From the above and $M_n = \big\lfloor (1+n)^{-|\mathcal{X}|} \mathrm{e}^{nR} \big\rfloor$, we obtain

$$\mathbb{E}\left[|\mathcal{E}_n(\boldsymbol{x}|\boldsymbol{y}, B_n)|\right] \le \mathrm{e}^{nH(\boldsymbol{x}|\boldsymbol{y}) - nR + n\delta_n'} \quad (28)$$

from some $\delta_n' \ge 0$ which goes to zero as n grows large. Defining $\delta_n'' \triangleq \delta_n + \delta_n'$, and using (28), it follows that

$$\min\left\{1, \mathrm{e}^{-n(\Delta-\delta_n)} \mathbb{E}\big[|\mathcal{E}_n(\boldsymbol{x}|\boldsymbol{y}, B_n)|\big]\right\} \le \mathrm{e}^{-n|R+\Delta-H(\boldsymbol{x}|\boldsymbol{y})|^+ + n\delta_n''}.$$

Plugging this back into (26), and invoking the usual random coding argument of the existence of a code, we obtain

$$\begin{aligned}
\varepsilon_{X|Y}^{\mathrm{WZ}}(n, R, \Delta) &\le \sum_{\boldsymbol{y} \in \mathcal{Y}^n} \sum_{\boldsymbol{x} \in \mathcal{X}^n} P_{XY}^n(\boldsymbol{x}, \boldsymbol{y}) \mathrm{e}^{-n|R+\Delta-H(\boldsymbol{x}|\boldsymbol{y})|^+ + n\delta_n''} \\
&\le \sum_{Q_{XY} \in \mathcal{P}_n(\mathcal{X}\times\mathcal{Y})} \mathrm{e}^{-nD(Q_{XY}\|P_{XY})} \mathrm{e}^{-n|R+\Delta-H(Q_{X|Y}|Q_Y)|^+ + n\delta_n''} \\
&\le (n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|} \mathrm{e}^{-n\tilde{F}_{X|Y}(R+\Delta) + n\delta_n''}. \quad (29)
\end{aligned}$$

The result in Theorem 3 follows.

V. Concluding Remarks

In the high rate regime where $E_{X|Y}^{\mathrm{r}}(R, \Delta)$ and $E_{X|Y}^{\mathrm{sp}}(R, \Delta)$ diverge, it is possible to derive a tighter achievable exponent, which is a log-loss counterpart of the expurgated exponent in source coding [15]. We skip this due to lack of space. As an extension, it is of interest to derive error exponents and universal schemes for the multi-terminal settings in [1], [2].

References

[1] T. A. Courtade and R. D. Wesel, "Multiterminal source coding with an entropy-based distortion measure," in *Proc. IEEE Int. Symp. Inf. Theory*, 2011, pp. 2040–2044.

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

[3] Y. Y. Shkel and S. Verdú, "A single-shot approach to lossy source coding under logarithmic loss," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 129–147, 2018.

[4] Y. Shkel, M. Raginsky, and S. Verdú, "Universal lossy compression under logarithmic loss," in *Proc. IEEE Int. Symp. Inf. Theory*, 2017, pp. 1157–1161.

[5] Y. Shkel, M. Raginsky, and S. Verdu, "Universal compression, list decoding, and logarithmic loss," in *Proc. IEEE Int. Symp. Inf. Theory*, 2018, pp. 206–210.

[6] I. Csiszár and J. Körner, *Information theory: Coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[7] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inf. Theory*, vol. 20, no. 2, pp. 197–199, 1974.

[8] F. Jelinek, *Probabilistic information theory: Discrete and memoryless models*. McGraw-Hill, 1968.

[9] G. Longo and A. Sgarro, "The source coding theorem revisited: A combinatorial approach," *IEEE Trans. Inf. Theory*, vol. 25, no. 5, pp. 544–548, 1979.

[10] C. Bunte and A. Lapidoth, "Source coding, lists, and Rényi entropy," in *Proc. IEEE Inf. Theory Workshop*, 2013, pp. 1–5.

[11] S. Arimoto, "Information measures and capacity of order α for discrete memoryless channels," in *Topics in Information Theory*, I. Csiszár and P. Elias, Eds. Amsterdam: North-Holland Publ. Co, 1977, pp. 41–52.

[12] R. G. Gallager, "Source coding with side information and universal coding," M.I.T. LIDS, Tech. Rep., 1976.

[13] I. Csiszár and J. Körner, "Towards a general theory of source networks," *IEEE Trans. Inf. Theory*, vol. 26, no. 2, pp. 155–165, 1980.

[14] B. G. Kelly and A. B. Wagner, "Improved source coding exponents via Witsenhausen's rate," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 5615–5633, 2011.

[15] I. Csiszár and J. Körner, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 5–12, 1981.