Mismatched Rate-Distortion Theory: Ensembles, Bounds, and General Alphabets

Millen Kanabar, Graduate Student Member, IEEE, and Jonathan Scarlett, Member, IEEE

Abstract—In this paper, we consider the mismatched rate-distortion problem, in which the encoding is done using a codebook, and the encoder chooses the minimum-distortion codeword according to a mismatched distortion function that differs from the true one. For the case of discrete memoryless sources, we establish achievable rate-distortion bounds using multi-user coding techniques, namely, superposition coding and expurgated parallel coding. We study examples where these attain the matched rate-distortion trade-off but a standard ensemble with independent codewords fails to do so. On the other hand, in contrast with the channel coding counterpart, we show that there are cases where structured random codebooks can perform worse than their unstructured counterparts. In addition, in view of the difficulties in adapting the existing and above-mentioned results to general alphabets, we consider a simpler i.i.d. random coding ensemble, and establish its achievable rate-distortion bounds for general alphabets.

Index Terms—Mismatched encoding, rate-distortion theory, random coding.

I. INTRODUCTION

RATE-DISTORTION theory is one of the most fundamental topics in information theory, and since its introduction [1], an extensive and diverse range of rate-distortion settings have been considered. In this paper, we are interested in a setting studied by Lapidoth [2], in which the encoding is mismatched, i.e., it is done with respect to an “incorrect” distortion measure $d_0$ that may be different from the true measure $d_1$. As discussed in [2] and [3], this problem can arise when the true distortion measure is not known when designing the encoder/decoder, or when the optimal encoder is ruled out due to implementation constraints (e.g., finite-precision arithmetic). More broadly, the problem is fundamental in nature, and may provide useful tools for other problems involving compression and/or imperfect encoding rules, e.g. estimation from compressed data [4], [5].

Manuscript received 29 April 2022; revised 20 December 2022; accepted 14 January 2024. Date of publication 19 January 2024; date of current version 16 February 2024. This work was supported by the National University of Singapore (NUS) Early Career Research Award. An earlier version of this paper was presented in part at the 2022 IEEE International Symposium on Information Theory (ISIT) [DOI: 10.1109/ISIT50566.2022.9834423].

Corresponding author: Jonathan Scarlett.

Millen Kanabar is with the School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland (e-mail: millen.kanabar@epfl.ch).

Jonathan Scarlett is with the Department of Computer Science, the Department of Mathematics, and the Institute of Data Science, National University of Singapore (NUS), Singapore 117417 (e-mail: scarlett@comp.nus.edu.sg).

Communicated by A. Somekh-Baruch, Associate Editor for Shannon Theory and Information Measures.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2024.3356477.

Digital Object Identifier 10.1109/TIT.2024.3356477

0018-9448 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

The mismatched rate-distortion problem serves as a natural counterpart to mismatched decoding in channel coding [3]. However, different aspects turn out to be easier in one problem than the other. For instance, an important conjecture on the tightness of a multi-letter achievable rate [6], [7] remains open in channel coding, but its analog was resolved in the rate-distortion setting [2]. On the other hand, general alphabets have been studied in detail in channel coding [8], [9], but appear to be more challenging in the rate-distortion problem.

In this paper, we study the following lesser-understood aspects of the mismatched rate-distortion problem: (i) multi-user coding techniques in the case of discrete memoryless sources, and (ii) regular random coding techniques in the case of general alphabets. These aspects have been studied extensively in the mismatched channel coding problem [9], [10], [11], and their counterparts in mismatched rate-distortion theory were included in a list of open problems in [3, Ch. 10].

Regarding multi-user coding techniques, a simple example consisting of parallel sources in [2] demonstrated that such techniques can be beneficial, but general results have remained unexplored. Regarding general alphabets, we are not aware of any existing results for the setup we consider, but this direction has been explored in related rate-distortion settings with only a single distortion function (e.g., see [12], [13]).

Briefly, our main contributions are as follows:

- For discrete memoryless sources, we derive two new achievability results on the mismatched distortion-rate function based on superposition coding and expurgated parallel coding.
- We demonstrate that, in contrast to the channel coding problem, introducing structure into the codebook (i.e., statistical dependencies between the random codewords) can be detrimental to the rate-distortion trade-off.
- On the positive side, we explore examples where multi-user techniques provide a strict improvement compared to the ensemble with independent codewords, one of which is conceptually very different from the known parallel source example.
- In view of the apparent difficulties in generalizing the existing and above-mentioned achievability results to general alphabets, we study the simpler i.i.d. random coding ensemble, and establish an achievable rate that extends to general alphabets.

A. Problem Setup

Consider a discrete memoryless source (DMS) where each symbol is drawn independently from a source distribution.
$\Pi_1(x)$ over a finite alphabet $X$. The distribution of the resulting $n$-letter sequence $X$ is then given by $\Pi^n_1(x) = \prod_{i=1}^n \Pi_1(x_i)$. The encoder maps $x$ to an index $m \in \{1, 2, \ldots, M\}$ for some integer $M$, and the rate of the encoder is given by $R = \frac{1}{n} \log M$.

The choice of $m$ is dictated by a codebook $\mathcal{C} = \{\tilde{x}^{(1)}, \ldots, \tilde{x}^{(M)}\}$, with codewords in $\tilde{X}^n$ for some reconstruction alphabet $\tilde{X}$. We assume that the encoder chooses, for some non-negative encoding metric $d_0 : X \times \tilde{X} \rightarrow \mathbb{R}_{\geq 0}$,

$$m = \arg \min_{1 \leq j \leq M} d_0^n(x, \tilde{x}^{(j)}), \quad d_0^n(x, \tilde{x}) = \sum_{i=1}^n d_0(x_i, \tilde{x}_i). \quad (1)$$

The distortion incurred is given by $d_1^n(x, \tilde{x}) = \sum_{i=1}^n d_1(x_i, \tilde{x}_i)$ for some true non-negative distortion measure $d_1 : X \times \tilde{X} \rightarrow \mathbb{R}_{\geq 0}$. If $d_0 = d_1$ then this is a standard rate-distortion problem, but otherwise, the encoding is said to be mismatched.

The choice of tie-breaking strategy in (1) can sometimes be significant (see our extended technical report [14, Appendix C] for an example), with choices including the following:

- Following [2], the pessimistic strategy assumes that any ties in (1) always lead to the worst-case value of $d_1^n$ with respect to those tied codewords.
- Alternatively, ties may be broken uniformly at random.
- For convenience, in one of our results, we will assume that ties are always broken by choosing the tied codeword with the lowest index in $\{1, \ldots, M\}$.

Except where stated otherwise, the achievability results that we state hold regardless of the tie-breaking strategy, i.e., even under pessimistic tie-breaking. The focus on the pessimistic rule in [2] was primarily for the purpose of proving a multi-letter converse.

**Definition 1:** (Mismatched Distortion-Rate and Rate-Distortion Functions): A rate-distortion pair $(R, D_1)$ is said to be achievable if, for any $\delta > 0$, there exists a sequence of codebooks indexed by $n$ with $M \leq e^{n(R+\delta)}$ codewords such that

$$\mathbb{P}\left[ \frac{1}{n} d_1^n(X, \hat{X}) \geq D_1 + \delta \right] \leq \delta \quad (2)$$

for sufficiently large $n$, where $X \sim \Pi_1^n$ and $\hat{X}$ is the resulting estimate via (1). The rate-distortion function $R^*(D_1)$ is defined as the smallest rate such that $(R, D_1)$ is achievable. The distortion-rate function $D_1^*(R)$ is defined as the smallest distortion $D_1$ such that $(R, D_1)$ is achievable.

**B. Related Work**

Our work is most closely related to that of Lapidoth [2], whose main achievable result for constant-composition (or near-constant composition) coding is stated below in Lemma 2.\(^1\) It was also shown in [2] that this result can be improved by applying it to the product source $(\pi_1^n, \pi_2^n)$, and so on) and that the resulting achievable distortion has a matching converse in the limit of an increasing product. We will also build on two example sources studied in [2].

To our knowledge, follow-up works of [2] in settings with multiple distortion measures have remained fairly limited, at least in aspects that are of direct relevance to our work. Related problems were studied for joint source-channel coding and strategic communication in [15] and [16]. Another interesting related setup was recently considered in [17], in which the distortion measure is unknown when designing the codebook, but is revealed to the encoder at runtime.

Another prominent form of mismatch in the literature is that in which there is only a single distortion measure, but the codebook is designed for the wrong source or is required to work for multiple different sources [2], [12], [13], [18], [19], [20], [21], [22], [23]. We will make use of several useful auxiliary results from [12] and [13] that fall in this category, even though we consider a different setup with two distortion measures.

When it comes to continuous sources, perhaps the most well-known result concerns the performance of Gaussian codes applied to non-Gaussian sources with the standard squared-distance distortion measure [2], [18], [19]. To our knowledge, the setup we consider has not been considered previously with continuous (or more general) alphabets.

Another line of works has studied universal codes for rate-distortion theory [12], [13], [18], [19], [24], [25], with the goal of finding codes that work well simultaneously for a large class of sources. In this setting, the encoder and decoder may be designed specifically to achieve this goal, in contrast with our study of mismatched encoding.

Finally, our problem setup serves as a natural counterpart to mismatched decoding in channel coding, for which multi-user ensembles were studied in [9], [10], and [11], and general alphabets were studied in [8], [9], [26], and [27], among others. See [3] for a survey of this topic.

**C. Notation**

Our analysis will be based on the method of types [28]. For a given type $P_U$ (i.e., a possible empirical distribution with block length $n$), we let $T^n_u(P_U)$ denote the set of sequences $u$ whose empirical distribution is $P_U$, and similarly for joint types. For a fixed sequence $\mathbf{u}$ whose empirical distribution matches the marginal of some joint type $P_{UX}$, we let $T^n_{\mathbf{u}}(P_{UX})$ denote the set of all $\mathbf{x}$ sequences such that $(\mathbf{u}, \mathbf{x})$ has joint empirical distribution $P_{UX}$. The marginals of a joint distribution $P_{UX}$ are written as $P_U$ and $P_X$.

The set of distributions on a given alphabet is denoted by $\Pi(\cdot)$, and the set of types is denoted by $\Pi_n(\cdot)$. Logarithms and information quantities have base $e$ except where stated otherwise. We make use of the binary entropy function $H_2(a)$ (i.e., the entropy with probabilities $(a, 1-a)$) and the ternary entropy function $H_3(a, b)$ (i.e., the entropy with probabilities $(a, b, 1-a-b)$).

**II. RANDOM CODING WITH INDEPENDENT CODEWORDS**

As a precursor to our main results, we overview the achievable distortion-rate function derived in [2] based on
constant-composition random coding, and present a closely-related result for i.i.d. random coding.

A. Constant-Composition Ensemble

Consider an auxiliary distribution \( Q_{\bar{X}} \in \mathcal{P}(\bar{X}) \), and let \( Q_{\bar{X},n} \in \mathcal{P}_n(\bar{X}) \) be a type with the same support as \( Q_{\bar{X}} \) such that \( ||Q_{\bar{X},n} - Q_{\bar{X}}||_\infty \leq \frac{1}{n} \). In the constant-composition ensemble, we draw each codeword independently from

\[
P_{\bar{X}}(\bar{x}) = \frac{1}{|T^n(Q_{\bar{X},n})|} \mathbb{1}(\bar{x} \in T^n(Q_{\bar{X},n})),
\]

(3)

where \( T^n(Q_{\bar{X},n}) \) is the type class corresponding to \( Q_{\bar{X},n} \). The following lemma is stated in [3, Lemma 4.3] based on the analysis of [2].

**Lemma 1 (Occurrences of Joint Types [2], [3]):** Fix \( \Pi_X \), \( Q_{\bar{X}} \), and \( R \), and consider a random codebook \( C_n = \{ X^{(1)}, \ldots, X^{(M)} \} \) with \( M = e^{nR} \) codewords of length \( n \) drawn independently from \( P_{\bar{X}} \) in (3). For any \( \delta > 0 \), conditioned on any \( X = x \) with \( x \in T(P_{X,n}) \) for some type \( P_{X,n} \), the following statements hold with probability approaching one as \( n \to \infty \):

1) For all \( \bar{x} \in C_n \), if \( (x, \bar{x}) \in T^n(\bar{P}_{X,\bar{X}}) \) for some \( \bar{P}_{X,\bar{X}} \in \mathcal{P}_n(X \times \bar{X}) \), then it must hold that \( P_{\bar{X}} = P_{X,n} \) and \( \bar{P}_{\bar{X}} = Q_{\bar{X},n} \).

2) Defining the set

\[
S^Q_{n,\delta} = \left\{ \bar{P}_{X,\bar{X}} \in \mathcal{P}_n(X \times \bar{X}) : \bar{P}_{\bar{X}} = Q_{\bar{X},n}, I_{\bar{P}}(X, \bar{X}) \leq R + \delta \right\},
\]

(4)

all joint types \( \bar{P}_{X,\bar{X}} \) induced by the codewords (i.e., \( (x, \bar{x}) \in T^n(\bar{P}_{X,\bar{X}}) \) for some codeword \( \bar{x} \)) must satisfy \( P_{\bar{X},n} \in S^Q_{n,\delta}(P_{X,n}) \).

3) For any joint type \( \bar{P}_{X,\bar{X}} \) in the set

\[
S^e_{n,\delta} = \left\{ \bar{P}_{X,\bar{X}} \in \mathcal{P}_n(X \times \bar{X}) : \bar{P}_{\bar{X}} = Q_{\bar{X},n}, I_{\bar{P}}(X, \bar{X}) \leq R - \delta \right\},
\]

(5)

there exists at least one codeword \( \bar{x} \in C_n \) for which \( (x, \bar{x}) \in T^n(\bar{P}_{X,\bar{X}}) \).

Once this lemma is established, the intuition is that the encoder chooses some codeword inducing a joint type \( \bar{P}_{X,\bar{X}} \) that minimizes the \( d_0 \)-distortion \( \mathbb{E}_{\bar{P}}[d_0(X,X)] \) subject to the above constraints, and these constraints asymptotically simplify to \( P_{\bar{X}} = \Pi_X \), \( \bar{P}_{\bar{X}} = Q_{\bar{X}} \), and \( I_{\bar{P}}(X; \bar{X}) \leq R \) as \( n \to \infty \) and \( \delta \to 0 \). Then, the \( d_1 \)-distortion is upper bounded by the maximum \( d_1 \)-distortion among all such minimizers, leading to the following.

**Lemma 2: (Achievability via the Constant-Composition Ensemble [2]):** Under the mismatched rate-distortion setup with a discrete memoryless source \( \Pi_X \) and distortion measures \( (d_0, d_1) \), the following distortion is achievable at rate \( R \) via constant-composition random coding with an auxiliary distribution \( Q_{\bar{X}} \in \mathcal{P}(\bar{X}) \):

\[
\bar{D}_1(Q_{\bar{X}}, R) = \max_{\bar{P}_{X,\bar{X}} \in \bar{P}} \mathbb{E}_{\bar{P}}[d_1(X, \bar{X})],
\]

(6)

where

\[
\bar{P} = \left\{ \bar{P}_{X,\bar{X}} : \bar{P}_{X,\bar{X}} \in \mathcal{P}, \arg\min_{\bar{P}_{X,\bar{X}} : \bar{P}_{X,\bar{X}} = Q_{\bar{X}}} \mathbb{E}_{\bar{P}}[d_0(X, \bar{X})] \right\}.
\]

(7)

Consequently, we have \( D^*_1(R) \leq \min_{Q_{\bar{X}}} \bar{D}_1(Q_{\bar{X}}, R) \).

Formalizing the above intuition requires a rather technical continuity argument to transfer from joint types to general joint distributions and replace \( \delta \) by zero. Since these details were omitted in [3] and are slightly different from [2] due to the different ensemble used (constant-composition vs. nearly-constant composition), we provide an overview in an extended technical report [14, Appendix D], along with a discussion of how to adapt it to the other ensembles we consider.

B. i.i.d. Ensemble

While Lapidoth’s results are based on the constant-composition ensemble (or very closely related ensembles), an even more common approach in general information theory problems is the i.i.d. random coding ensemble, in which each codeword is independently drawn from

\[
P_{\bar{X}}(\bar{x}) = \prod_{i=1}^n Q_{\bar{X}}(\bar{x}_i)
\]

(8)

for some auxiliary distribution \( Q_{\bar{X}} \). Note that unlike constant-composition coding, this ensemble naturally applies to continuous (or more general) alphabets. Accordingly, it will be our main focus when considering general alphabets in Section IV.

For discrete memoryless sources, the following result follows via an almost identical argument to that of Lemma 2 (see [2] and [3, Ch. 4]), so the details are omitted.

**Lemma 3 (Achievability via the i.i.d. Ensemble):** Under the mismatched rate-distortion setup with a discrete memoryless source \( \Pi_X \) and distortion measures \( (d_0, d_1) \), the following distortion is achievable at rate \( R \) via i.i.d. random coding with an auxiliary distribution \( Q_{\bar{X}} \in \mathcal{P}(\bar{X}) \):

\[
\bar{D}_{1}^{\text{iid}}(Q_{\bar{X}}, R) = \max_{\bar{P}_{X,\bar{X}} \in \bar{P}_{\text{iid}}} \mathbb{E}_{\bar{P}}[d_1(X, \bar{X})],
\]

(9)

where

\[
\bar{P}_{\text{iid}} = \left\{ \bar{P}_{X,\bar{X}} : \bar{P}_{X,\bar{X}} \in \mathcal{P}, \arg\min_{\bar{P}_{X,\bar{X}} : \bar{P}_{X,\bar{X}} = Q_{\bar{X}}} \mathbb{E}_{\bar{P}}[d_0(X, \bar{X})] \right\}.
\]

(10)

Consequently, we have \( D^*_1(R) \leq \min_{Q_{\bar{X}}} \bar{D}_{1}^{\text{iid}}(Q_{\bar{X}}, R) \).

Intuitively, the constraint \( \bar{P}_{X,\bar{X}} = Q_{\bar{X}} \) is absent in (10) because the codewords no longer necessarily have an empirical distribution equal to (or close to) \( Q_{\bar{X}} \). In addition, the mutual information \( I_{\bar{P}}(X; \bar{X}) \) is replaced by \( D(\bar{P}_{X,\bar{X}} || \Pi_X \times Q_{\bar{X}}) \), since the two are no longer equivalent when \( \bar{P}_{X,\bar{X}} \) may differ from \( Q_{\bar{X}} \).

As we will see via a more general discussion in Section III-C, neither the i.i.d. nor constant-composition ensembles are guaranteed to outperform one another in general.
III. MULTI-USER RANDOM CODING TECHNIQUES

In this section, we present two achievability results based on multi-user coding techniques that were previously used for mismatched channel coding [9], [10], [11], namely, superposition coding and parallel coding. We also briefly mention that connections and differences between these ensembles (and other closely related ensembles) have been explored in detail in matched multi-user problems, e.g., see [29], [30], and [31] and the references therein.

A. Superposition Coding

1) Codebook Construction: The ensemble is defined in terms of an auxiliary alphabet $U$, an auxiliary codeword distribution $P_U \in \mathcal{P}(U^n)$ and a conditional codeword distribution $P_{X|U} \in \mathcal{P}(X^n|U^n)$. We fix $(R_0, R_1)$ and generate the codebook in the following manner:

- First, $M_0 = \{e^{nR_0}\}$ codewords are drawn uniformly and independently from the $n$-letter distribution $P_U$, generating $\{U^{(i)}\}_{1 \leq i \leq M_0}$.

- For each $U^{(i)} \in \{U^{(i)}\}_{1 \leq i \leq M_0}$, $M_1 = \{e^{nR_1}\}$ codewords are drawn uniformly and conditionally independently from the $n$-letter conditional distribution $P_{X|U}^{(i)}$, generating $\{X^{(i,j)}\}_{1 \leq j \leq M_0}$.

For a given (joint) input distribution $Q_{UX,n} \in \mathcal{P}(U \times X, n)$, let $Q_{UX,n} \in \mathcal{P}(U \times X)$ be a joint type such that $\|Q_{UX,n} - Q_{UX}\|_{\infty} \leq \frac{1}{n}$. We then consider the following (constant-composition) choices of $P_U$ and $P_{X|U}$:

$$P_U(u) = \frac{1}{|T^n(Q_{U,n})|} \mathbf{1}\{u \in T^n(Q_{U,n})\}$$

$$P_{X|U}(x|u) = \frac{1}{|T^n(Q_{UX,n})|} \mathbf{1}\{(u, x) \in T^n(Q_{UX,n})\}.$$  \hspace{1cm} (11) (12)

The encoder maps a sequence $x \in X^n$ to $(m_0, m_1)$ such that $(m_0, m_1) = \arg\min_{i,j} d_0^n(x, X^{(i,j)})$, and the corresponding reconstruction sequence $\hat{x}$ is given by $\hat{x} = X^{(m_0, m_1)}$. The rate is given by

$$R = \frac{1}{n} \log(M_0M_1) = R_0 + R_1.$$ \hspace{1cm} (13)

Under this ensemble, we have the following achievability result, which we compare to Lemma 2 in Sections III-C–III-D.

Theorem 1 (Achievability for Superposition Coding):

Under the superposition coding ensemble described above with input distribution $Q_{UX} \in \mathcal{P}(U \times X)$ and rates $R_0$ and $R_1$, the following distortion is achievable at rate $R = R_0 + R_1$:

$$D_1(Q_{UX}, R_0, R_1) = \max_{P_{X|UX} \in \mathcal{P}} \mathbb{E}_P[d_1(X, \hat{X})],$$ \hspace{1cm} (14)

where

$$\mathcal{P} = \left\{ P_{X|UX} : P_{X|UX} \in \mathcal{P}_{X|UX} : P_X = P_X, \right\}$$

Consequently, we have the inequality $D_1^*(R) \leq \min_{R_0 + R_1 = R} D_1(Q_{UX}, R_0, R_1)$.

B. Expurgated Parallel Coding

We consider a codebook generated from two auxiliary codebooks of size $M_1$ and $M_2$ over auxiliary alphabets $X_1$ and $X_2$, along with a function $\psi : X_1 \times X_2 \rightarrow X$ mapping to the reconstruction alphabet. Specifically, following analogous ideas used in mismatched channel coding [10], the codebook is generated as follows:

- For $\nu = 1, 2$, auxiliary constant-composition codebooks $\{X^{(i)}\}_{i=1}^{M_\nu}$ with input distributions $Q_{X_\nu}$ and with codewords each are generated, each independently drawn uniformly over the type class $T^n(Q_{X_\nu,n})$ with $\|Q_{X_\nu,n} - Q_{X_\nu}\|_{\infty} \leq \frac{1}{n}$.

- An initial codebook of size $M_1M_2$ is constructed as $\hat{X}^{(i,j)} = \psi^n(X^{(i)}, X^{(j)})$ for $1 \leq i \leq M_1, 1 \leq j \leq M_2$, where $\psi^n(\cdot, \cdot)$ applies the function $\psi$ entry-by-entry.

- The codebook is expurgated, keeping only the codewords $\hat{X}^{(i,j)}$ whose corresponding pairs $(X^{(i)}, X^{(j)})$ have empirical distributions within $\delta > 0$ of $Q_{X_1} \times Q_{X_2}$ in $\ell_\infty$-norm. Let $I$ denote the set of $(i, j)$ pairs that are kept.

An input sequence $x$ is mapped to $(m_1, m_2)$ such that

$$\arg\min_{(i,j) \in I} d_0^n(x, \psi^n(X^{(i)}, X^{(j)})) = (m_1, m_2).$$ \hspace{1cm} (16)

The reconstruction sequence is given as $\hat{X} = \psi^n(\hat{X}_1^{(m_1)}, \hat{X}_2^{(m_2)})$, and the rate is given by $R = \frac{1}{n} \log |I|$. While this technically means that the rate is random, it will turn out that $R$ approaches $R_1 + R_2$ with probability approaching one, where $R_\nu = \frac{1}{n} \log M_\nu$.

Under this ensemble, we have the following achievability result, which we compare to Lemma 2 in Sections III-C–III-D.
Theorem 2 (Achievability for Expurgated Parallel Coding): For given auxiliary distributions $Q_{X_1}, Q_{X_2}$, function $\psi(\cdot)$ and rates $R_1, R_2$, using expurgated parallel coding, the following distortion is achievable at rate $R = R_1 + R_2$:

$$
\mathcal{D}_1(Q_{X_1}, Q_{X_2}, \psi, R_1, R_2) = \max_{\tilde{P}_{X_1, \hat{X}_1} \tilde{P}_{X_2, \hat{X}_2} \in \tilde{P}} \mathbb{E}_P[d_1(X, \psi(\hat{X}_1, \hat{X}_2))],
$$

where

$$
\tilde{P} = \left\{ \tilde{P}_{X_1, \hat{X}_1}, \tilde{P}_{X_2, \hat{X}_2} : \tilde{P}_{X_1, \hat{X}_1} \in \arg \min \mathbb{E}_P[d_0(X, \psi(\hat{X}_1, \hat{X}_2))] \right\}.
$$

Consequently, we have the inequality $D_1^*(R) \leq \min_{(Q_{X_1}, Q_{X_2}, \psi, R_1, R_2), R_1 + R_2 = R} \mathcal{D}_1(Q_{X_1}, Q_{X_2}, \psi, R_1, R_2)$.

Once again, in view of the similarities to the channel coding counterpart [9], [11], we only provide a short proof outline here, and defer the full details to an extended technical report [14, Appendix E]. The main steps are analogous to those of Theorem 1, but in this case, the independence of the codewords $\hat{X}^{(i,j)}$ and $\hat{X}^{(i',j')}$ is dependent whenever $i = i'$ or $j = j'$. We can still use the truncated union bound to upper bound the relevant existence probabilities, whereas the lower bound instead uses de Caen’s inequality [32]. This is done in the same way as [9], and while the precise bound used in [9] is slightly too loose due to a factor of $\frac{1}{2}$, we can easily sharpen their analysis by avoiding a step of the form $a + b + c + d \leq 4 \max\{a, b, c, d\}$. By doing so and again using standard properties of types followed by a continuity argument, we obtain the desired result.

C. Structured vs. Unstructured Random Codebooks

In the channel coding problem with mismatched decoding, it is known (analogos of) the above ensembles never provide an achievable rate worse than that of independent constant-composition codewords with the same marginal code distribution. It is therefore natural to ask whether the same holds for mismatched-rate-distortion theory.

In fact, it is straightforward to see that for fixed choices of parameters $(Q_{X, X'}, R_0, R_1)$ or $(Q_{X_1, X_2}, R_1, R_2, \psi)$, the answer is negative. To see this, fix the encoding function $d_0$, and consider any setup in which (7) has a unique minimizer $\tilde{P}^*_{X\hat{X}}$, and (15) or (18) has a unique minimizer with a joint marginal $\tilde{P}^{**}_{X\hat{X}}$ that differs from $\tilde{P}^*_{X\hat{X}}$. The existence of examples satisfying these requirements can be inferred from [2] or from Section III-D to follow.

Then, consider the following possibilities for the true distortion function:

$$
d_1^*(x, \hat{x}) = \log \tilde{P}^*_{X\hat{X}}(x, \hat{x}), \quad d_1^{**}(x, \hat{x}) = \log \tilde{P}^{**}_{X\hat{X}}(x, \hat{x}).
$$

The non-negativity of KL divergence implies that any maximization problem of the form $\max_P \mathbb{E}_P[\log Q(X)]$ is uniquely maximized by $P = Q$, and hence, we conclude that $\tilde{P}^*_{X\hat{X}}$ yields strictly higher distortion under $d_1^*$, whereas $\tilde{P}^{**}_{X\hat{X}}$ yields strictly higher distortion under $d_1^{**}$. Hence, there is no general inequality between the achievable mismatched distortion functions.

On the other hand, when the parameters $(Q_{X, X'}, R_0, R_1)$ or $(Q_{X_1, X_2}, R_1, R_2, \psi)$ are optimized, we can easily guarantee being at least as good as the existing bound in Lemma 2. This is because the superposition coding ensemble becomes equivalent to having independent codewords when $R_0 = 0$ (or $R_1 = 0$), and similarly for expurgated parallel coding with $R_2 = 0$ and $\psi(x_1, x_2) = x_1$ (or $R_0 = 0$ and $\psi(x_1, x_2) = x_2$).

More importantly, the benefit of multi-user coding techniques is seen by constructing examples where they provably meet the matched rate distortion function but the existing bound fails to do so. In the next subsection, we revisit one such example from [2] for expurgated parallel coding, and provide a new example of this kind for superposition coding.

We note that the preceding discussion not only applies to independent codewords vs. multi-user coding techniques, but also to i.i.d. vs. constant-composition random coding. For fixed $Q_{X, X'}$ and general choices of $d_0$ and $d_1$, it was already noted in [3] that neither ensemble is guaranteed to outperform the other, and the above arguments provide another way of seeing this. On the other hand, when $d_0 = d_1$, the i.i.d. ensemble is always as good or better, because the minimum in (10) is less constrained compared to (7), meaning there are more joint distributions to choose from for making $\mathbb{E}_P[d_0(X, \hat{X})]$ small. Intuitively, the added diversity in the i.i.d. random codebook can give better performance when the correct distortion measure is used for encoding, whereas when $d_0 \neq d_1$, the situation is more subtle depending on how the encoder may be led astray by the mismatch. Our extended technical report [14, Appendix C] gives an example with $d_0 \neq d_1$ where the constant-composition ensemble is strictly better.

D. Examples

In this subsection, we give examples that illustrate improvements in the distortion-rate function when using the codebook generation techniques described in the previous section.

1) Parallel Binary Source: Consider a source with $\mathcal{X} = \{0, 1\}$ and reconstruction alphabet, with $\Pi_X(x) = \frac{1}{2}, \forall x$. As in [2], we consider the encoding metric

$$
d_0(x, \hat{x}) = \lambda \mathbb{I}\{x_1 \neq \hat{x}_1\} + (1 - \lambda) \mathbb{I}\{x_2 \neq \hat{x}_2\},
$$

and a distortion metric of

$$
d_1(x, \hat{x}) = \frac{1}{2}(\mathbb{I}\{x_1 \neq \hat{x}_1\} + \mathbb{I}\{x_2 \neq \hat{x}_2\}).
$$

For convenience, work in bits (rather than nats) in this example. It was shown in [2] that the achievable distortion with independent codewords (Lemma 2) simplifies to

$$
\mathcal{D}_1(Q_{X, X'}, R) = \frac{1}{2}(\delta_1^* + \delta_2^*)
$$

and

$$
\begin{align*}
(\tilde{\delta}_1^*, \tilde{\delta}_2^*) &= \arg \min_{(\delta_1, \delta_2) : (1 - H_2(\delta_1)) + (1 - H_2(\delta_2)) \leq R} \lambda \tilde{\delta}_1 + (1 - \lambda) \tilde{\delta}_2,
\end{align*}
$$

\(\text{Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.} \)
whereas the matched distortion-rate function is given by $D_1^*$ such that
\[ 2(1 - H_2(D_1^*)) = R. \] (24)

The fact that parallel coding outperforms this result and attains the matched performance was already noted in [2] using direct arguments, namely, the fact that the two encoding metrics are equivalent when the codebook has a product structure. Here we outline how the same (with expurgation) can be established via the analytical expression in Theorem 2, thus serving as a sanity check for this general result.

Consider taking $\psi(x_1, x_2) = (x_1, x_2)$ and $Q_{\bar{X}} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ ($\nu = 1, 2$) in Theorem 2. As was noted in [2], the independence of $X_1$ and $X_2$ is sufficient to lower bound $I_1(X; \bar{X}_1, \bar{X}_2)$ by $\tilde{I}_1(X_1; \bar{X}_1) + \tilde{I}_1(X_2; \bar{X}_2)$. In addition, we trivially have $I_1(X; \bar{X}_\nu) \geq \tilde{I}_1(X_\nu; \bar{X}_\nu)$ for $\nu = 1, 2$. Since both $X$ and $\bar{X}$ follow the uniform distribution on $\{0, 1\}^2$, the relevant margins $\tilde{P}_{X, \bar{X}}$ must follow a binary symmetric channel law. Denoting the associated crossover probabilities by $\tilde{\delta}_1$ and $\tilde{\delta}_2$, it follows that
\[ I_1(X; \bar{X}_1, \bar{X}_2) \geq (1 - H_2(\tilde{\delta}_1)) + (1 - H_2(\tilde{\delta}_2)), \] (25)
\[ I_1(X; \bar{X}_\nu) \geq (1 - H_2(\tilde{\delta}_\nu)), \quad \nu = 1, 2. \] (26)

Again following [2], we observe that these bounds hold with equality when the underlying BSCs are independent, and that such independence must hold under the optimal $\tilde{P}_{X, \bar{X}}$. This is because if any dependence were present, moving to the independent version would allow us to decrease $R_1$ and/or $R_2$ without increasing the distortion.

In view of (25)–(26) holding with equality, we are now left with the following analog of (22)–(23):
\[ \bar{D}_1(Q_{\bar{X}}, Q_{\bar{X}}; \psi, R_1, R_2) = \frac{1}{2}(\tilde{\delta}_1 + \tilde{\delta}_2) \] (27)
\[ (\tilde{\delta}_1, \tilde{\delta}_2) = \arg \min_{(\delta_1, \delta_2): 1 - H_2(\delta_1) \leq R_1, 1 - H_2(\delta_2) \leq R_2} \lambda \tilde{\delta}_1 + (1 - \lambda) \tilde{\delta}_2, \] (28)

where the constraint $(1 - H_2(\tilde{\delta}_1)) + (1 - H_2(\tilde{\delta}_2)) \leq R_1 + R_2$ is omitted because it is automatically guaranteed by the other two. Finally, we set $R_1 = R_2 = \frac{R}{2}$, and observe that the minimum in (28) amounts to separate minimizations over the two parameters, each giving an optimal value of $\tilde{\delta}_\nu$ that satisfies $1 - H_2(\tilde{\delta}_\nu) = R/2$. In view of (24), the resulting $d_1$-distortion coincides with that attained in the matched case.

2) Uniform Ternary Source: Consider the uniform ternary source with $\Pi_X(x) = \frac{1}{3}$ for $x \in \{0, 1, 2\}$. We consider an encoding metric $d_0(x, \bar{x}) = I\{x \neq \bar{x}\}$ and a true distortion measure $d_1(x, \bar{x}) = I\{x \neq \bar{x}, x \neq \bar{x}\}$. Moreover, we consider the case that both $\Pi_X$ and $Q_X$ are uniform:
\[ \Pi_X = Q_X = \frac{1}{3}, \frac{1}{3}, \frac{1}{3}. \] (29)

We first outline the results for the matched setting, and for unstructured random coding with mismatch (Lemma 2), and then turn to our own achievable result (Theorem 1).

Matched distortion-rate function. In the matched setting with distortion $d_1$ only, the achievable distortion-rate function with a fixed codeword distribution $Q_{\bar{X}}$ is given as follows (e.g., see [12], or Lemma 2 with $d_0 = d_1$):
\[ \bar{D}_1^*(Q_{\bar{X}}, R) = \min_{\tilde{P}_{X, \bar{X}}: I(\tilde{P}_{X, \bar{X}}; \bar{P}_X) \leq R} \mathbb{E}_{\tilde{P}_X}[d_1(X, \bar{X})]. \] (30)

Standard constant-composition coding. The achievable mismatched distortion-rate function achieved using standard constant-composition coding (i.e., independent codewords) was characterized by Lapidoth [2], who showed that that there is a unique $\tilde{P}_{X^*}$ that minimizes $E[d_0(X, \bar{X})]$ in (6), with $\tilde{P}_{X^*}$ being a ternary symmetric channel with some cross-symbol transition probability $\delta^*$ such that $I(\tilde{P}_{X}; \bar{X}) = R$. The resulting distortion-rate function is given by $\bar{D}_1^*(R) = \frac{2\delta^*}{3}$, where $\delta^*$ implicitly depends on $R$. It is shown in [3, Fig. 4.2] that the resulting rate-distortion trade-off is strictly worse than the matched case (for uniform ternary $Q_{\bar{X}}$).

Superposition coding. We take the distribution $Q_{U, \bar{X}}$ as
\[ Q_{U, \bar{X}} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} \end{bmatrix}, \] (31)

with the rows indexed by $U$ and columns by $\bar{X}$. We then have the following result for superposition coding.

Lemma 4: In the ternary source example described above, for any rate $R$, there exists $0 \leq R_0 \leq R$ such that the distortion-rate function for mismatched encoding using superposition coding, with $Q_{U, \bar{X}}$ as described in (31) and $R_1 = R - R_0$, equals the distortion-rate function of the matched case (using $d_1$ for encoding) with $Q_{\bar{X}}$ being the $\bar{X}$-marginal of (31).

The proof is given in Appendix B, and utilizes an almost-sure convergence result from [13] regarding the empirical distribution of $(X, \bar{X})$ when $X$ is the first codeword in an infinite-length codebook to attain a pre-specified distortion level $D_0$. As noted in [13], their result generalizes and strengthens an earlier “favorite type” theorem for the finite-alphabet setting [33].

In our analysis, we slightly adapt the result of [13] to characterize the empirical $d_1$-distortion induced by $(X, \bar{X})$ (whereas [13] focuses on probabilities of events). However, even after doing so, a notable difficulty is that in our setting $D_0$ is not pre-specified, as we consider the minimum-distortion encoder with a fixed rate $R$. To address this, we consider an intersection of events (each corresponding the adapted result of [13] with a different $D_0$ value) over a rational subset of $D_0$ values, noting that a countable intersection of almost-sure events also holds almost surely. From there, the proof consists of arguing that (i) with a fixed rate $R$, the $d_0$-distortion of the selected codeword approaches $D_0^{\text{id}}(Q_{\bar{X}}, R)$ in (32), and (ii) via the preceding almost-sure event and a somewhat technical continuity argument, the corresponding $d_1$-distortion converges to $\bar{D}_1^{\text{id}}(Q_{\bar{X}}, R)$.

IV. GENERAL ALPHABETS

In this section, we turn to general alphabets (in particular, possibly countably infinite or continuous), which have been extensively studied in the mismatched channel coding
problem [8], [9], [26], [27], [34], but to our knowledge, not for the mismatched rate-distortion problem that we consider. The other hand, this direction has been considered for a distinct notion of mismatched rate-distortion in [12] and [13]; specifically, while these works consider \( d_0 = d_1 \), they use “mismatched codebooks” in the sense of having a suboptimal choice of \( Q \). Our analysis directly builds on the results from these works.

For the purpose of capturing the examples of practical interest, it is useful to think of \( \Pi_X \) and \( Q_X \) as being mass functions for finite or countably infinite alphabets, and density functions for continuous alphabets. Formally, however, since we build on the tools in [13], the alphabets may be as general as those therein, where the only restriction is that the source and reconstruction alphabets are Polish spaces such that all singletons are measurable.

A. Alternative Expression for the i.i.d. Ensemble

Extending the constant-composition result from Lemma 2 to general alphabets appears to be difficult (and Theorems 1 and 2 even more so). One might consider using cost-constrained random coding [8], [9], but the difficulty that is compared to channel coding, the techniques needed for the achievability result and ensemble tightness are reversed. The ensemble tightness of cost-constrained random coding for channel coding currently remains open, and establishing an achievability result in the present setting is similarly challenging.

In view of this difficulty, in the remainder of this section, we consider the simpler i.i.d. random coding ensemble (see Section II-B), for which we present a useful reformulation of Lemma 3. Here we make use of the matched \( d_0 \)-distortion function with fixed \( Q_X \), defined as

\[
\mathcal{D}^\text{id}_0(Q_X, R) = \min_{\bar{P}_{X|\hat{X}}: P_X = \Pi_X, \quad D(\bar{P}_{X|\hat{X}}\|\Pi_X \times Q_X) \leq R} \mathbb{E}_P[d_0(X, \hat{X})].
\]  

(32)

Moreover, as in [12] and [13], we restrict our attention to rates (and their associated distortion levels) in a restricted range. Specifically, we define the following extreme values of the \( d_0 \)-distortion:

\[
D_{0,\text{min}} = \mathbb{E}_P[\text{essinf}_{\hat{X} \sim Q} d_0(X, \hat{X})],
\]

(33)

\[
D_{0,\text{prod}} = \mathbb{E}_P d_0(X, \hat{X}),
\]

(34)

Intuitively, \( D_{0,\text{min}} \) is the average \( d_0 \)-distortion that would be attained by an infinite-length codebook, and \( D_{0,\text{prod}} \) is the average distortion that would be attained by a codebook with just a single random codeword. Accordingly, we consider rates in the interval \( R \in (0, R_{\text{max}}) \), where

\[
R_{\text{max}} = \lim_{d_0 \to D_{0,\text{min}}} \mathcal{R}^\text{id}(Q_X, D_0).
\]

(35)

with the limit taken from above, and where \( \mathcal{R}^\text{id} \) is the matched rate-distortion function defined as [12]

\[
\mathcal{R}^\text{id}(Q_X, D_0) = \min_{\bar{P}_{X|\hat{X}}: P_X = \Pi_X, \quad \mathbb{E}_P[d_0(X, \hat{X})] \leq D_0} D(\bar{P}_{X|\hat{X}}\|\Pi_X \times Q_X).
\]

(36)

Working in the range \( R \in (0, R_{\text{max}}) \) (corresponding to \( d_0 \)-distortion in \( (D_{0,\text{min}}, D_{0,\text{prod}}) \)) turns out to give a unique minimizer in (10), whereas for higher rates this is not necessarily the case (see our extended technical report [14, Appendix C] for an example).

Although Lemma 3 is stated for discrete memoryless sources, the expressions in (9)-(10) can still be taken as valid definitions for general memoryless sources. We proceed by stating an equivalent form for \( \mathcal{D}^\text{id}_0 \) (9), and then turn to proving a counterpart to Lemma 3 for general alphabets.

Lemma 5: Consider the mismatched rate-distortion problem with a given source \( \Pi_X \) and auxiliary distribution \( Q_X \), and suppose that \( D_{0,\text{prod}} < \infty \) and \( R \in (0, R_{\text{max}}) \) with \( R_{\text{max}} \) given in (35). Then, the quantity \( \mathcal{D}^\text{id}_0(Q_X, R) \) in (9) can equivalently be expressed as

\[
\mathcal{D}^\text{id}_0(Q_X, R) = \mathbb{E}_{\tilde{P}_X}[d_1(X, \hat{X})],
\]

(37)

where \( \tilde{P}_{X|\hat{X}} \) is defined according to the Radon-Nikodym derivative

\[
\frac{d\tilde{P}_{X|\hat{X}}}{d(\Pi_X \times Q_X)}(x, \hat{x}) = \frac{e^{\lambda^* d_0(x, \hat{x})}}{\mathbb{E}_Q[e^{\lambda^* d_0(x, \hat{x})}]},
\]

(38)

and where \( \lambda^* \leq 0 \) is the unique value such that the function \( \Lambda(\lambda) = \mathbb{E}_Q[\log \mathbb{E}_Q[e^{\lambda d_0(X, \hat{X})}]] \) satisfies \( \Lambda(\lambda) = D_0 \), with \( D_0 = \mathcal{D}^\text{id}_0(Q_X, R) \).

The proof is given in Appendix B, and makes use of various findings from [12], in which the matched setting is considered (i.e., \( d_0 = d_1 \)), but we will use \( d_0 \) when discussing this setting, and the joint distribution \( \tilde{P}_{X|\hat{X}} \) is utilized. The authors of [12] show that the rate-distortion function is given in (36), and provide the following equivalent “dual” expression:

\[
\mathcal{R}^\text{id}(Q_X, D_0) = \sup_{\lambda \leq 0} \{D_0 - \Lambda(\lambda)\}.
\]

(39)

Thus, in our setting, the definition of \( \lambda^* \) in Lemma 5 follows naturally from solving the maximum (39); the existence and uniqueness of \( \lambda^* \) (as per its definition) when \( D_0 \in (D_{0,\text{min}}, D_{0,\text{prod}}) \) is established in [12]. Once \( \lambda^* \) is specified, the joint distribution in (38) is also specified. With the preceding findings and definitions in place, the proof of Lemma 5 amounts to showing that \( \tilde{P}_{X|\hat{X}} \) is the unique minimizer in (10).

B. Achievability Result for General Alphabets

Analogous to \( D_{0,\text{prod}} \) in (34), we define

\[
D_{1,\text{prod}} = \mathbb{E}_{\Pi \times Q}[d_1(X, \hat{X})],
\]

(40)

and we impose the assumption that \( D_{1,\text{prod}} < \infty \), along with \( D_{0,\text{prod}} < \infty \) (as hinted above, this is also assumed for the single distortion measure considered in [12] and [13]).

Our main result for general alphabets is stated as follows.

Theorem 3 (Achievable Distortion With General Alphabets): Consider the general-alphabet mismatched rate-distortion setup with fixed \( \Pi_X \) and \( Q_X \) satisfying \( D_{0,\text{prod}} < \infty \) and
$D_{1,\text{prod}} < \infty$, let $R \in (0, R_{\max})$ with $R_{\max}$ defined in (35), and suppose that ties are broken in (1) by choosing the tied codeword with the smallest index. Then, the distortion $\overline{D}_1(Q_{\tilde{X}}, R)$ is achievable under i.i.d. random coding, where $\overline{D}_1$ can equivalently be expressed as (9)–(10) or (37)–(38).

The proof is given in Appendix B, and utilizes an almost-sure convergence result from [13] regarding the empirical distribution of $(X, \tilde{X})$ when $\tilde{X}$ is the first codeword in an infinite-length codebook to attain a pre-specified distortion level $D_0$. As noted in [13], their result generalizes and strengthens an earlier “favorite type” theorem for the finite-alphabet setting [33].

In our analysis, we slightly adapt the result of [13] to characterize the empirical $d_1$-distortion induced by $(X, \tilde{X})$ (whereas [13] focuses on probabilities of events). However, even after doing so, a notable difficulty is that in our setting $D_0$ is not pre-specified, but is instead induced by the minimum-distortion rule with a given rate $R$. To address this, we consider an intersection of events (each corresponding the result from [13] with a different $D_0$ value) over a rational subset of $D_0$ values, noting that a countable intersection of almost-sure events also holds almost surely. From there, the proof consists of arguing that (i) with a fixed rate $R$, the $d_0$-distortion of the selected codeword approaches $\overline{D}_0(\tilde{Q}_{\tilde{X}}, R)$ in (32), and (ii) via the preceding almost-sure event and a somewhat technical continuity argument, the corresponding $d_1$-distortion converges to $\overline{D}_1(Q_{\tilde{X}}, R)$.

C. Example: Gaussian Source With a Mismatched Distortion Function

As an example of Theorem 3, we consider the case that $X \sim N(0, \sigma^2)$, $\tilde{X} \sim N(0, \tau^2)$, and $d_0(x, \tilde{x}) = (x - \tilde{x})^2$, for some $\sigma^2, \tau^2 > 0$. We will consider both the matched case (corresponding to $d_1 = d_0$), and the mismatched case with $d_1(x, \tilde{x}) = 1\{\text{sign}(x) \neq \text{sign}(\tilde{x})\}$ (i.e., the true distortion only seeks that the signs be recovered correctly). This can be viewed as a toy example of a scenario in which the compression is performed assuming the goal of accurate estimation, but it is then used only for the purpose of binary classification.

In the matched case where $d_1$ is also used for encoding (i.e., $d_0$ plays no role), the problem becomes equivalent to compressing an equiprobable binary source, and the rate-distortion function is $R^*(D_1) = \log 2 - H_2(D_1)$ for $D_1 \in [0, 1/2]$ [35, Sec. 10.3.1].

Moreover, when $d_0$ is used for both encoding and for measuring the final distortion (i.e., $d_1$ plays no role), the rate-distortion trade-off for i.i.d. random coding was characterized in [12, Ex. 1]: We have $D_{0,\min} = 0$, $D_{0,\text{prod}} = \sigma^2 + \tau^2$, and the following for $D_0 \in (D_{0,\min}, D_{0,\text{prod}})$ when $R_{\text{prod}}$ is measured in nats:

$$\overline{R}_{\text{prod}}(Q_{\tilde{X}}, D_0) = \frac{1}{2} \log \frac{v}{D_0} - \frac{(v - D_0)(v - \sigma^2)}{2v\tau^2},$$

$$v = \frac{1}{2} \left( \tau^2 + \sqrt{\tau^4 + 4D_0\sigma^2} \right).$$

Using this finding as a starting point, in the mismatched setting with the above choices of $d_0$ and $d_1$, we can evaluate the rate-distortion curve as follows:

- Sweep over a range of values $\lambda \leq 0$ and consider the corresponding joint distributions in $P_{X\tilde{X}}^*$ in (38) (implicitly depending on $\lambda$).
- Observe that due to the assumptions $X \sim N(0, \sigma^2)$, $\tilde{X} \sim N(0, \tau^2)$, and $d_0(x, \tilde{x}) = (x - \tilde{x})^2$, all of the terms in (38) give an expression whose exponent is quadratic with respect to $x$ and $\tilde{x}$. Thus, $P_{X\tilde{X}}^*$ is a bivariate Gaussian distribution whose parameters $(\mu_X^*, \mu_{\tilde{X}}^*, \sigma_X^*, \sigma_{\tilde{X}}^*, \rho^*)$ (with $\rho^* \in [-1, 1]$ being the correlation coefficient) can be computed as a function of $(\sigma, \tau, \lambda^*)$. Omitting the tedious calculations, we state the resulting parameters as follows. The coefficients to $-x^2$, $-\tilde{x}^2$, and $-x\tilde{x}$ in the exponent in (38) are easily computed to be

$$a = \frac{1}{\lambda^2 - \lambda}, \quad b = \frac{1}{\lambda^2}, \quad c = 2\lambda$$

(with the remaining coefficients to $x$ and $\tilde{x}$ being zero), and from these values, the desired parameters can be shown to be as follows:

$$\mu_X^* = 0, \quad \mu_{\tilde{X}}^* = 0, \quad (\sigma_X^*)^2 = \sqrt{\frac{2b}{4ab - c^2}},$$

$$(\sigma_{\tilde{X}}^*)^2 = \sqrt{\frac{2a}{4ab - c^2}}, \quad \rho^* = \frac{-c}{2\sqrt{ab}}.$$
seen to be highly suboptimal here. This is particularly the case at low distortion levels, where the rate becomes unbounded in the mismatched case, despite a rate of one being trivial. Intuitively, this is because typical $X$ sequences contain many low-valued entries, and using the nearest-neighbor encoder is likely to flip their signs unless there are a huge number of codewords to choose from (i.e., a high rate).

V. Conclusion

We have studied multi-user coding techniques and general alphabets for the mismatched rate-distortion problem, and established several new results that serve as natural counterparts to those that were previously known for mismatched channel coding. Several possible directions remain for future work, including (i) studying the refined superposition coding ensemble considered for channel coding in [9]; (ii) establishing conditions under which the various ensembles provably outperform one another; and (iii) attaining general-alphabet counterparts of the rate-distortion curves achieved by the constant constant-composition and/or multi-user ensembles.

APPENDIX A

ANALYSIS OF THE UNIFORM TERNARY SOURCE EXAMPLE

In this section, we study the ternary source example of Example III-D.2, and prove Lemma 4 therein. Throughout this appendix, all information measures are in units of nats.

A. Alternative Expressions

Since both the $X$-marginal and $	ilde{X}$-marginal of $\tilde{P}_{X\tilde{X}}$ are uniform, we can express $\tilde{P}_{X\tilde{X}}$ in the following form:

$$
\tilde{P}_{X\tilde{X}} = \begin{bmatrix}
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3}
\end{bmatrix}
$$

(46)

where we write the joint distribution in matrix form with rows indexed by $x$ and columns indexed by $\tilde{x}$, and $(p_{01}, p_{10}, p_{02}, p_{12})$ are constrained to take values such that all elements of the matrix are non-negative. In this notation, the matched distortion in (30) can be rewritten as

$$
\overline{D}_1^* (Q_X, R) = \min_{\tilde{P}_{X\tilde{X}} : I_{\tilde{P}}(X; \tilde{X}) \leq R} p_{01} + p_{10} + p_{02} + p_{12}. 
$$

(47)

To make this more explicit, we write $I_{\tilde{P}}(X; \tilde{X})$ in terms of the ternary entropy function $H_3(a, b)$ (i.e., the entropy with probabilities $(a, b, 1-a-b)$) as follows:

$$
I_{\tilde{P}}(X; \tilde{X}) = H_3(X) - H_{\tilde{P}}(X|\tilde{X})
$$

$$
= \log 3 - H_{\tilde{P}}(X|\tilde{X})
$$

$$
= \log 3 - \sum_x \Pi_X(x) H_{\tilde{P}}(X|\tilde{X} = x)
$$

$$
= \log 3 - \frac{1}{3} \left( H_3(3p_{01}, 3p_{02}) + H_3(3p_{10}, 3p_{12}) + H_3(3p_{02} + (3p_{01} - 3p_{10}), 3p_{12} - (3p_{01} - 3p_{10})) \right)
$$

(51)

where (51) follows since each $H(\tilde{X}|X = x)$ is the ternary entropy associated with multiplying the relevant row of (46) by 3 (i.e., dividing by $\Pi_X(x) = \frac{1}{3}$), and (52) follows by applying Jensen’s inequality to the first two terms, and using $H_3(a, b) \leq H_3(\frac{a+b}{3}, \frac{a+b}{3})$ (again by concavity of entropy) in the third term. Moreover, since the ternary entropy function is strictly concave, equality holds if and only if $p_{01} = p_{10}$ and $p_{02} = p_{12}$.

B. Analysis of the Matched Performance

As a step towards proving Lemma 4, which states that the matched performance is attainable even in the presence of mismatch, it is useful to first study the minimization problem (30) associated with the matched setting. We present several useful lemmas regarding this.

Lemma 6: In the present ternary source example, for $R \in [0, \log 3)$, any distribution achieving the minimum in (30) satisfies both $p_{01} = p_{10}$ and $p_{02} = p_{12}$ when expressed as given in (46).

Proof: This result trivially holds for $R = \log 3$, since the constraint $I_{\tilde{P}}(X; \tilde{X}) \leq R$ is then satisfied by all valid distributions of the form (46). Selecting $\tilde{X} = X$ (i.e., $p_{01} = p_{02} = p_{10} = p_{12} = 0$) satisfies this with zero distortion. Conversely, if $D^*_1 = 0$ then the choice of distortion function implies that $p_{01} = p_{02} = p_{10} = p_{12} = 0$ (see (46)), which in turn implies $\tilde{X} = X$ and $R = \log 3$. Thus, whenever $R < \log 3$, it must be the case that $D^*_1 (Q_X, R) > 0$, and we focus on this scenario in the rest of the analysis.

Fix $R < \log 3$, and assume for contradiction that some distribution $\tilde{P}_a$ is asymmetric (in the sense that $p_{01,a} \neq p_{10,a}$ and/or $p_{02,a} \neq p_{12,a}$) but achieves the minimum in (30). We define the symmetrized version $\tilde{P}_a' = \tilde{P}_a + \tilde{P}_a^\dagger$, where $\tilde{P}_a^\dagger$ is the distribution with $p_{01,a}' = p_{10,a}', p_{02,a}' = p_{12,a}'$ (i.e., the suitable pairs are interchanged). Observe that $\tilde{P}_a'$ is symmetric in the sense that $p_{01,a}' = p_{10,a}$ and $p_{02,a}' = p_{12,a}$.

We now have the following:

$$
I_{\tilde{P}}(X; \tilde{X}) < I_{\tilde{P}_a}(X; \tilde{X}) = I_{\tilde{P}_a'}(X; \tilde{X}),
$$

(53)

where the inequality follows from (52) (and the condition for strict inequality stated just after), and the equality follows since the mutual information is invariant to interchanging the pairs as described above (see (51)). Moreover, since the distortion is given by $p_{01} + p_{10} + p_{02} + p_{12}$ (see (47)), we have $\mathbb{E}_{\tilde{P}_a} [d_1(X; \tilde{X})] = \mathbb{E}_{\tilde{P}_a'} [d_1(X; \tilde{X})]$. Thus, the fact that $I_{\tilde{P}}(X; \tilde{X}) \leq R$ implies that $I_{\tilde{P}_a}(X; \tilde{X}) < R$, which implies that $\tilde{P}_a$ also achieves the minimum in (30).

If $p_{01,a} > 0$, then from the continuity of mutual information, there exists some $p_{01,a}' < p_{01,a}$ such that the corresponding mutual information still satisfies $I_{\tilde{P}_a'}(X; \tilde{X}) \leq R$ (with $p_{02,a}' = p_{02,a}$). The corresponding value of $\mathbb{E}_{\tilde{P}_a'} [d_1(X; \tilde{X})]$ is
then smaller than that of $\tilde{P}_a$, thus contradicting the assumption that $\tilde{P}_a$ achieves the minimum.

If $p_{01,s} = 0$, then we must have $p_{02,s}$, since we have assumed that we are not in the zero-distortion region. Then, by the same reasoning as the case where $p_{01,s} > 0$, we find that there exists $p_{02,s} < p_{02,s}$ such that $I_{\tilde{P}_s}(X;\tilde{X}) \leq R$. This leads to a distortion smaller than the minimizing distribution, arriving at a contradiction and completing the proof. □

**Lemma 7:** In the present ternary source example, for $R \in [0, \log 3]$, any distribution achieving the minimum in (30) satisfies $I_{\tilde{P}}(X;\tilde{X}) = R$.

**Proof:** Suppose for contradiction that there exists a distribution $\tilde{P}_s$ (assumed symmetric according to Lemma 6) achieving the minimum with $I_{\tilde{P}_s}(X;\tilde{X}) < R$. Then, from the analysis in the proof of Lemma 6, there exists $\tilde{P}_s'$ with $E_{\tilde{P}_s'}[d_1(X,\tilde{X})] < E_{\tilde{P}_s}[d_1(X,\tilde{X})]$ and $I_{\tilde{P}_s}(X;\tilde{X}) \leq R$. This is a contradiction, since we assumed $P_s$ achieves the minimum.

In view of Lemma 6, we can reduce the four parameters $(p_{01}, p_{10}, p_{02}, p_{12})$ in (46) to just two parameters $(p_{01}, p_{02})$. To satisfy the constraint of non-negative matrix entries, we require the following:

$$0 \leq p_{01} + p_{02} \leq \frac{1}{3}$$  \hspace{1cm} (54)

$$0 \leq p_{02} \leq \frac{1}{6}.$$  \hspace{1cm} (55)

Moreover, the symmetry from Lemma 6 implies that (52) holds with equality, and the mutual information becomes

$$I_{\tilde{P}}(X;\tilde{X}) = \log 3 - \frac{1}{3} \left(2H(3p_{01}, 3p_{02}) + H(3p_{02}, 3p_{02})\right).$$

Hence, (47) reduces to minimizing $2(p_{01} + p_{02})$ subject to (54)–(55), as well as (56) being at most $R$.

From the two-parameter expression (56), we can readily evaluate the partial derivatives with respect to $p_{01}$ and $p_{02}$ (the details amount to standard calculus and are omitted):

$$\frac{\partial}{\partial p_{01}} I_{\tilde{P}}(X;\tilde{X}) = 2 \log \left(\frac{p_{01}}{\frac{1}{3} - p_{01} - p_{02}}\right),$$

$$\frac{\partial}{\partial p_{02}} I_{\tilde{P}}(X;\tilde{X}) = 2 \log \left(\frac{p_{02}}{(\frac{1}{3} - p_{01} - p_{02})(\frac{1}{3} - 2p_{02})}\right).$$

A final useful lemma regarding the matched case is as follows.

**Lemma 8:** In the present ternary source example with $R \in [0, \log 3]$, the optimal value of $p_{02}$ for a distribution achieving the matched distortion-rate function, when $d_1$ is used as the distortion metric, is at most $\frac{1}{3}$.

**Proof:** We first note that the constraint $p_{01} + p_{02} \leq \frac{1}{3}$ in (54) is inactive, because the choice $p_{01} = p_{02} = \frac{1}{9}$ gives a smaller value of distortion (namely, $2p_{01} + 2p_{02}$) compared to any pair $(p_{01}, p_{02})$ such that $p_{01} + p_{02} = \frac{1}{3}$, and is guaranteed to be feasible since it gives $I_{\tilde{P}}(X;X) = 0$.

Moreover, the constraint $p_{02} \leq \frac{1}{6}$ in (55) is inactive, since for any fixed value of $p_{01}$, (58) implies that $I_{\tilde{P}}(X;\tilde{X})$ increases to its value at $p_{02} = \frac{1}{6}$ infinitely sharply. Hence, when $p_{02} = \frac{1}{6}$, we can hold $p_{01}$ constant and slightly decrease $p_{02}$, thereby decreasing both $I_{\tilde{P}}(X;\tilde{X})$ and the distortion.

We also claim that in the positive-distortion regime (i.e., $p_{01} + p_{02} > 0$), neither of $p_{01}$ and $p_{02}$ can be 0. This is seen by fixing $p_{01}$ and $p_{02}$ to a positive value $p_{\text{sum}}$ and observing from (57)–(58) that

$$\lim_{p_{02} \to 0} \frac{\partial I_{\tilde{P}}(X;\tilde{X})}{\partial p_{02}} \to -\infty, \quad \nu = 1, 2; \quad (59)$$

whereas the partial derivative at $p_{02} = p_{\text{sum}} (\nu = 1, 2)$ is bounded and finite. If only one of the two is zero (say $p_{01}$), we see that increasing $p_{01}$ and decreasing $p_{02}$ by the same small quantity leads to a net decrease in $I_{\tilde{P}}(X;\tilde{X})$ while keeping the distortion constant, contradicting the fact that $I_{\tilde{P}}(X;\tilde{X}) = R$ (from Lemma 7).

Combining the preceding paragraphs, we have shown that the solution to the symmetrized version of (47) is unchanged when the constraints in (54)–(55) are dropped altogether. Hence, we are left with studying the simplified problem

$$\min_{(p_{01}, p_{02})} \log 3 - \frac{1}{3} \left(2H(3p_{01}, 3p_{02}) + H(3p_{02}, 3p_{02})\right) \leq R$$

where the expression in the constraint comes from (56).

Applying the KKT optimality conditions to (60), we find that there exists $\lambda > 0$ such that

$$\nabla I_{\tilde{P}}(X;\tilde{X}) = -\lambda \left[\frac{2}{3}\right],$$

from which it follows that

$$\frac{\partial}{\partial p_{01}} I_{\tilde{P}}(X;\tilde{X}) \Big|_{p_{01}} = \frac{\partial}{\partial p_{02}} I_{\tilde{P}}(X;\tilde{X}) \Big|_{p_{02}}$$

$$\Rightarrow p_{01} = \left(\frac{p_{02}}{3} - 2p_{02}\right),$$

where (63) follows by equating (57) with (58) and simplifying. A simple analysis of the derivative of (63) reveals that $p_{01}$ is a strictly increasing function of $p_{02}$ in the admissible range $p_{02} \in [0, \frac{1}{3}]$, and therefore, so is the distortion. On the other hand, we know that $p_{01}^* = p_{02}^* = \frac{1}{3}$ corresponds to a rate of zero. By the non-increasing nature of the distortion-rate function, we conclude that distortion cannot exceed its zero-rate value, and therefore, $p_{02}$ (and in fact, also $p_{01}$) cannot exceed $\frac{1}{3}$.

**C. Analysis of Superposition Coding**

Since the $X$-marginal and $\tilde{X}$-marginal are still uniform here, we can again consider $P_{XU \tilde{X}}$ in the form given in (46). Moreover, the following analog of Lemma 6 holds.

**Lemma 9:** Consider the minimization problem with respect to $P_{XU \tilde{X}}$ in (15). In the present ternary source example, $P_{XU \tilde{X}}$ is uniquely determined by $P_{X \tilde{X}}$, and any distribution achieving the minimum satisfies both $p_{01} = p_{10}$ and $p_{02} = p_{12}$ when $P_{X \tilde{X}}$ is expressed as given in (46).

**Proof:** The choice of $Q_{XU \tilde{X}}$ in (31) implies that $U$ is deterministic given $\tilde{X}$, which means that once $P_{X \tilde{X}}$ is
specified, so is $\tilde{P}_{X\hat{X}}$ (recalling that we constrain $\tilde{P}_{U\hat{X}} = Q_{U\hat{X}}$). This proves the first claim.

For the second claim, the proof is similar to that of Lemma 6, so we only outline the differences. In the symmetrization step, interchanging $p_{01} \leftrightarrow p_{10}$ and/or $p_{02} \leftrightarrow p_{12}$ still does not impact the distortion (this time $d_0$ instead of $d_1$). Since $U$ is deterministic given $\hat{X}$, the mutual information $I(X;U;\hat{X})$ reduces to $I(X;\hat{X})$, so can be handled similarly to the matched case. Moreover, using the fact that $U = 1\{\hat{X} = 2\}$, we find that the marginal $P_{XU}$ can be obtained by combining the first two columns of (46) to obtain

$$P_{XU} = \begin{bmatrix} \frac{1}{3} - p_{02} & p_{02} \\ \frac{1}{3} - p_{12} & p_{12} \\ p_{02} + p_{12} & \frac{1}{3} - p_{02} - p_{12} \end{bmatrix},$$

which has no dependence on $(p_{01}, p_{10})$. Then, following similar steps to (48)–(52) (see also (73) below), we obtain that $I_{\tilde{P}}(X;U)$ is invariant to interchanging $p_{02} \leftrightarrow p_{12}$, and can always be made smaller by symmetrizing with respect to this interchanging (or it remains the same if such symmetry already held).

With these findings in place, the symmetrization argument follows in the same way as the proof of Lemma 6, considering both mutual information terms when handling $(p_{02}, p_{12})$, but only considering $I_{\tilde{P}}(X;U)$ when handling $(p_{01}, p_{10})$ (since $I_{\tilde{P}}(X;U)$ does not depend on these values). Using Lemma 9 and recalling the choice of $Q_{U\hat{X}}$ in (31), Theorem 1 simplifies as follows:

$$D_1(Q_{U\hat{X}}, R_0, R_1) = \max_{(p_{01}, p_{10})} 2(p_{01} + 2p_{02}),$$

where

$$\tilde{P} = \left\{ \frac{\text{arg min}}{(p_{01}, p_{10}) : I_{\tilde{P}}(X;U) \leq R_0, I_{\tilde{P}}(X;U;\hat{X}) \leq R} (2p_{01} + 2p_{02}) \right\},$$

and where the correspondence between $(p_{01}, p_{10})$ and $\tilde{P}_{XU}$ is as follows:

$$P_U(0) = \frac{2}{3}, P_U(1) = \frac{1}{3}.$$

$$\tilde{P}_{XU|U=0} = \frac{3}{2} \begin{bmatrix} \frac{1}{3} - p_{01} - p_{02} & 0 \\ p_{01} & \frac{1}{3} - p_{01} - p_{02} \\ p_{02} & 0 \end{bmatrix},$$

$$\tilde{P}_{XU|U=1} = \begin{bmatrix} 0 & 0 & p_{02} \\ 0 & p_{01} & 0 \\ 0 & 0 & \frac{1}{3} - 2p_{02} \end{bmatrix}.$$

Once again, $p_{01}$ and $p_{02}$ are constrained such that the matrix entries are non-negative.

With this symmetrization in place, the marginal distribution $P_{XU}$ shown in (64) simplifies to

$$P_{XU} = \begin{bmatrix} \frac{1}{3} - p_{02} & p_{02} \\ \frac{1}{3} - p_{12} & p_{12} \\ 2p_{02} & \frac{1}{3} - 2p_{02} \end{bmatrix},$$

which yields the following (with $H_2(\cdot)$ being the binary entropy function):

$$I_{\tilde{P}}(X;U) = H(U) - H(U|X)$$

$$= \log 3 - \frac{2}{3} \log 2 - \sum_x H(U|X = x)\Pi_X(x)$$

$$= \log 3 - \frac{2}{3} \log 2 - \frac{1}{3}(2H_2(3p_{02}) + H_2(6p_{02})),$$

since each $H(U|X = x)$ is the binary entropy associated with multiplying the relevant row of (70) by 3 (i.e., dividing by $\Pi_X(x) = \frac{1}{3}$). Moreover, we have already established that $I_{\tilde{P}}(X;U, \hat{X}) = I_{\tilde{P}}(X;\hat{X})$. Using these results, we are now in a position to prove our main finding.

**Proof of Lemma 4:** Recall that we simplified the $d_0$-distortion to $2(p_{01} + 2p_{02})$ using (46) with $p_{01} = p_{10}$ and $p_{02} = p_{12}$. For the $d_0$-distortion, the first two entries in the final row also contribute, giving a value of $2(p_{01} + 2p_{02})$. As with Lemma 6, since $X = \hat{X}$ trivially satisfies all constraints when $R = \log 3$, we focus on the cases where $R < \log 3$, and consequentially, $p_{01} + p_{02} > 0$.

We claim that the simplified matched problem (60) is uniquely minimized by some values $p_{01,m}^*$ and $p_{02,m}^*$. To see this, assume by contradiction that there are multiple minimizers. The average of any two of these will have the same distortion as the minimum, while $I_{\tilde{P}}(X;\hat{X}) = \log 3 - \frac{1}{3}(2H_2(3p_{01} + 3p_{02}) + H_2(3p_{02}, 3p_{02}))$ will be strictly less than $R$ (due to strict concavity), contradicting Lemma 7.

We note from Lemma 8 that $p_{02,m}^* \leq \frac{1}{3}$, and we consider the following choice of the parameter $R_0 = R - R_1$ in superposition coding:

$$R_0 = I_{\tilde{P}}(X;U),$$

where $I_{\tilde{P}}(X;U)$ is the value obtained on substituting $p_{02,m}^*$ in (73). For all $(p_{01}, p_{02})$ distinct from $(p_{01,m}^*, p_{02,m}^*)$ such that $I_{\tilde{P}}(X;U) \leq R_0$ and $I_{\tilde{P}}(X;\hat{X}) \leq R$, we have

$$p_{01,m}^* + p_{02,m}^* < p_{01} + p_{02},$$

since the minimizer of $d_1$ for the matched case is unique. Moreover, we claim that

$$p_{02,m}^* \leq p_{02}.$$

This is seen by combining the fact that $p_{02,m}^* \leq \frac{1}{3}$ (established above) with the fact that $I_{\tilde{P}}(X;U)$ is monotonically decreasing for $p_{02} \leq \frac{1}{3}$ (see Figure 2: an analytical proof is also straightforward). Thus, under the constraint $I_{\tilde{P}}(X;U) \leq R_0$, it is not feasible to go below $p_{02,m}^*$.

Since the $d_0$-distortion equals $2(p_{01} + 2p_{02})$, we conclude from (75)–(76) that every feasible pair $(p_{01}, p_{02}) \neq (p_{01,m}^*, p_{02,m}^*)$ attains a strictly higher $d_0$-distortion than $(p_{01,m}^*, p_{02,m}^*)$. This means $(p_{01,m}^*, p_{02,m}^*)$ is also uniquely optimal for the $d_0$-minimization problem in (66), and thus, the same $d_1$-distortion is attained under mismatched encoding as that of matched encoding. □
APPENDIX B

PROOFS FOR THE GENERAL ALPHABET SETTING

We will be interested in events that hold asymptotically almost surely with respect to \( X \), i.e., if we view \( X \) as the first \( n \) entries of an infinitely long i.i.d. sequence \( X_1^\infty = (X_1, X_2, \ldots) \), then there exists a subset \( \mathcal{X}_{\text{YP}}^\infty \) of such sequences such that \( \mathbb{P}[X_1^\infty \in \mathcal{X}_{\text{YP}}^\infty] = 1 \), and such that the desired event holds for all sequences in \( \mathcal{X}_{\text{YP}}^\infty \). When this is the case, we say that the event holds \( \Pi_X^\infty \)-almost surely (\( \Pi_X^\infty \)-a.s.), or for \( \Pi_X^\infty \)-almost all \( x \).

A. Auxiliary Results

We first provide some useful auxiliary results from the works \([12], [13]\), which consider “mismatched codebooks” in a setting with only a single distortion function \( d_0 \). We start with two simple results from \([12]\).

**Lemma 10 (Asymptotic Distortion Ball Probability \([12, \text{Thm. 1}]\):** Consider any i.i.d. source\(^4\) \( X \sim \Pi_X^\infty \) such that \( D_{0,\text{prod}} < \infty \), and for some auxiliary distribution \( Q_{\hat{X}} \), let \( \hat{X} \sim Q_{\hat{X}} \) be independent of \( X \). Then, for any \( D \in (D_{0,\text{min}}, D_{0,\text{prod}}) \), it holds for \( \Pi_X^\infty \)-almost all \( x \) that

\[
-\frac{1}{n} \log \mathbb{P}[d_0(x, \hat{X}) \leq nD_0 | X = x] \rightarrow R_{\text{ind}}(Q_{\hat{X}}, D_0)
\]

as \( n \rightarrow \infty \), where \( R_{\text{ind}} \) is defined in (36).

**Lemma 11 (Primal-Dual Equivalence \([12, \text{Thm. 2}]\):** For any \( \Pi_X \) and \( Q_{\hat{X}} \) such that \( D_{0,\text{prod}} < \infty \), the expressions (36) and (39) are equal for all \( D_0 \in (D_{0,\text{min}}, D_{0,\text{prod}}) \).

Noting that \( R_{\text{ind}}(Q_{\hat{X}}, \cdot) \) and \( D_{\text{ind}}(Q_{\hat{X}}, \cdot) \) (see (32)) are functional inverses of each other, the following result readily follows from Lemma 10; recall that \( R_{\text{max}} \) is defined in (35).

**Corollary 1:** (Achievable Matched Distortion for a Fixed Rate): For any source \( X \sim \Pi_X^\infty \) and auxiliary distribution \( Q_{\hat{X}} \) such that \( D_{0,\text{prod}} < \infty \), under the i.i.d. ensemble with rate \( R \in (0, R_{\text{max}}) \), it holds with probability approaching one that the codeword \( \hat{X} \) with the smallest \( d_0 \)-distortion satisfies

\[
\frac{1}{n} d_0(X, \hat{X}, \hat{X}^*) \rightarrow D_{\text{ind}}(Q_{\hat{X}}, R) \quad \text{as} \quad n \rightarrow \infty.
\]

We now state a less elementary result concerning the first codeword in an infinitely long codebook whose normalized distortion is no higher than a pre-specified threshold \( D_0 \); note that for \( D_0 \in (D_{0,\text{min}}, D_{0,\text{prod}}) \), an infinitely long codebook will always contain such a codeword.

**Lemma 12:** (Empirical Distribution for the First Matching Codeword \([13, \text{Thm. 2}]\): Suppose that \( D_{0,\text{prod}} < \infty \), and fix \( D_0 \in (D_{0,\text{min}}, D_{0,\text{prod}}) \). Consider an infinite codebook \( \mathcal{C} = (\hat{X}_1, \hat{X}_2, \ldots) \) independently drawn from \( Q_{\hat{X}}^n \), and let \( \hat{P}_n \) be the empirical distribution of \((X, \hat{X}^*)\) with \( \hat{X}^* \) being the first codeword in \( \mathcal{C} \) satisfying \( d_0^*(X, \hat{X}) \leq nD_0 \). Then, we have the following:

(i) For any measurable set \( E \) (not depending on \( n \)) and any \( \delta > 0 \), it holds for \( \Pi_X^\infty \)-almost all \( x \) that

\[
\mathbb{P}[|\hat{P}_n(E) - P_{X,\hat{X}}^*(E)| > \delta | X = x] \rightarrow 0,
\]

where the convergence to zero as \( n \rightarrow \infty \) is exponentially fast in \( n \), and \( P_{X,\hat{X}}^* \) is defined in (38) with \( \lambda^* \) defined according to \( D_0 \).

(ii) It holds with probability one that \( \hat{P}_n \) converges in distribution to \( P_{X,\hat{X}}^* \) as \( n \rightarrow \infty \).

The formal statement in \([13]\) actually concerns the empirical distribution of \( \hat{X} \) alone, instead of the joint empirical distribution of \((X, \hat{X})\). However, the proof therein readily gives the form stated above. Specifically, the proof of \([13, \text{Thm. 1}] \) (a conditional limit theorem from which \([13, \text{Thm. 2}] \) follows almost immediately) shows that a certain asymptotic distribution on \( \hat{X} \) must equal \( P_{\hat{X}}^* \) by first showing the stronger result that the asymptotic distribution on \((X, \hat{X})\) must equal \( P_{X,\hat{X}}^* \). We obtain Lemma 12 by skipping this step and directly using this stronger result.

In our setting, we are not directly interested in probabilities with respect to \( \hat{P}_n \), but instead, the average of \( d_1 \) with respect to \( \hat{P}_n \) (i.e., \( \frac{1}{n} d_1^*(x, \hat{X}) \)). The convergence result in the second part of Lemma 12 implies an analogous convergence result for \( \frac{1}{n} d_1^*(x, \hat{X}) \) in the case that \( d_1 \) is bounded and continuous, but these requirements rule out standard distortion functions such as \( d_1(x, \hat{X}) = |x - \hat{X}|^2 \). Fortunately, Lemma 12 admits a natural counterpart for averages instead of probabilities without such requirements, which we state as follows.

**Lemma 13:** (Empirical Averages for the First Matching Codeword \([13, \text{Thm. 2}]\): Suppose that \( D_{0,\text{prod}} < \infty \) and \( D_{1,\text{prod}} < \infty \), and fix \( D_0 \in (D_{0,\text{min}}, D_{0,\text{prod}}) \). Consider an infinite codebook \( \mathcal{C} = (\hat{X}_1, \hat{X}_2, \ldots) \) independently drawn from \( Q_{\hat{X}}^n \), and let \( \hat{P}_n \) be the empirical distribution of \((X, \hat{X}^*)\) with \( \hat{X}^* \) being the first codeword in \( \mathcal{C} \) satisfying \( d_0^*(X, \hat{X}) \leq nD_0 \). Then, we have the following:

(i) The empirical \( d_1 \)-distortion satisfies for any \( \delta > 0 \) and \( \Pi_X^\infty \)-almost all \( x \) that

\[
\mathbb{P}[|d_1(\hat{P}_n) - d_1(\hat{P}_{X,\hat{X}}^*)| > \delta | X = x] \rightarrow 0,
\]

where we use the shorthand \( d_1(P) = \mathbb{E}_P[d_1(X, \hat{X})] \), \( \hat{P}_{X,\hat{X}}^* \) is defined in (38) with \( \lambda^* \) defined according to \( D_0 \), and the convergence to zero as \( n \rightarrow \infty \) is exponentially fast in \( n \).

(ii) It holds with probability one that \( d_1(\hat{P}_n) \rightarrow d_1(\hat{P}_{X,\hat{X}}^*) \) as \( n \rightarrow \infty \).
The proof of this result follows the exact same steps as those of Lemma 12 (given in [13]) based on large deviations theory. Accordingly, we do not repeat the analysis here, but we provide the main ideas and differences in Appendix B-D.

B. Proof of Lemma 5 (Reformulation for the i.i.d. Ensemble)

Recall the two equivalent expressions for $\overline{R}^\text{id}(Q^\times, D_0)$ in (36) and (39). As stated following (39), it is known from [12] that the supremum in (39) is uniquely attained by some $\lambda^*$ satisfying $\Lambda' = D_0$ when $D_0 \in (D_{0,\text{min}}, D_{0,\text{prod}})$, where $\Lambda(\lambda) = \mathbb{E}_{Q} [\log \mathbb{E}[e^{\lambda d_0(X,X^*)}]]$. Moreover, it is shown in [12] that the unique corresponding minimizer in (36) is $P_{X^*}^*$, defined in (38). In view of these known results, it is then natural that the formulation in Lemma 3 reduces to that in Lemma 5. We now proceed by showing this formally.

While we build on [12], a key issue that we need to address is that their results concern the rate after specifying $D_0$, whereas in (10) we have specified $\epsilon$. Moreover, it is shown in [12] that the unique corresponding minimizer in (36) is $P_{X^*}^*$, defined in (38). In view of these known results, it is then natural that the formulation in Lemma 3 reduces to that in Lemma 5. We now proceed by showing this formally.

As outlined above, we know from [12] that $P_{X^*}^*$ is the unique minimizer of the problem

$$\min_{P_{X^*}^* : P_{X^*} \equiv \Pi_{X^*}, \mathbb{E}_P[d_0(X,X^*)] \leq R} D(P_{X^*}^* \| \Pi_{X^*} \times Q^\times).$$

(81)

Moreover, the minimum value is $R$, since the assumption $R \in (0, R_{\text{max}})$ implies that we are in the regime where the inequality constraint in (80) is active (this was also noted in [12]). To show that $P_{X^*}^*$ also minimizes (80), suppose for contradiction that there were to exist a feasible distribution $P'_{X^*}$ in (80) with value $\mathbb{E}_{P'}[d_0(X,X^*)] < R$. Then, for $\epsilon$ sufficiently small, it must hold that the distribution $P'_{X^*} = (1 - \epsilon)\overline{P}^*_{X^*} + \epsilon(\Pi_{X^*} \times Q^\times)$ satisfies $\mathbb{E}_{P'}[d_0(X,X^*)] \leq R$. But then we have

$$D(P'_{X^*} \| \Pi_{X^*} \times Q^\times) \leq (1 - \epsilon)D(\overline{P}^*_{X^*} \| \Pi_{X^*} \times Q^\times) \leq (1 - \epsilon)R,$$

(82)

where the first step substitutes $P'_{X^*} = (1 - \epsilon)\overline{P}^*_{X^*} + \epsilon(\Pi_{X^*} \times Q^\times)$ and $\Pi_{X^*} \times Q^\times = (1 - \epsilon)(\Pi_{X^*} \times Q^\times) + \epsilon(\Pi_{X^*} \times Q^\times)$ and then applies the convexity of KL divergence and $\mathbb{E}_P[D_0(X,X^*)] = 0$. This shows that $D(P'_{X^*} \| \Pi_{X^*} \times Q^\times)$ is feasible in (81) but yields a smaller objective than the optimal value of $R$, which is a contradiction, and establishes the desired claim.

For uniqueness, recall that we are in the regime where the inequality constraint in (80) is active. If there is another optimal distribution $\overline{P}_{X^*}^*$ distinct from $\overline{P}^*_{X^*}$, then they must both yield the same values of $D(\overline{P}^*_{X^*} \| \Pi_{X^*} \times Q^\times)$ and $\mathbb{E}_{P'_0}[d_0(X,X^*)]$. However, this would mean that $\overline{P}^*_{X^*}$ is also optimal in (81), contradicting the uniqueness known from [12].

C. Proof of Theorem 3

We would like to use Lemma 13 with $D_0 = D_{0,\text{prod}}^\text{id}(Q^\times, R)$ to conclude that its behavior similarly to $\overline{P}_{X^*}^*$, implying that the $d_0$-distortion incurred is $\mathbb{E}_{P_0}[d_1(X,X^*)]$. The difficulty is that we are in the fixed-rate setting, whereas Lemma 13 holds for the fixed-distortion setting. Since there is a continuum of possible $D_0$ values, we cannot directly deduce that Lemma 13 holds for the actual $d_0$-distortion value incurred by the source and the chosen codeword. To circumvent this, we extend the almost-sure event to a countably infinite set, as well as exploiting our tie-breaking strategy that chooses the first index.

We first state the following corollary of Lemma 13.

Corollary 2 (Uniform Guarantee Over Rational $D_0$ Values):

Fix any $D_0^\infty \in (D_{0,\text{min}}, D_{0,\text{prod}})$, and let $\epsilon > 0$ be sufficiently small such that $[D_0^\infty - \delta, D_0^\infty + \delta] \subseteq (D_{0,\text{min}}, D_{0,\text{prod}})$. Consider an infinite codebook $C = (X_1, X_2, \ldots)$ with codewords independently drawn from $Q^\times_0$. Let $\mathbb{P}^n_{X^*}$ be as defined in (38) with an explicit dependence on $D_0$, and let $\mathbb{P}_{D_0}(X)$ be the empirical distribution of $(X_1, X_2, \ldots)$ with $\mathbb{P}_{D_0}(X)$ being the first codeword in $C$ satisfying $d_0^n(X, X^*) \leq n D_0$. Then, with probability one, it holds simultaneously for all rational values of $D_0 \in [D_0^\infty - \delta, D_0^\infty + \delta]$ that $d_1(\mathbb{P}_{D_0^n} - d_1(\mathbb{P}_{D_0^n}) \rightarrow d_1(\mathbb{P}_{D_0^n})$ as $n \rightarrow \infty$.

Proof: Observe that we restrict $D_0$ to ensure that $D_0 \in (D_{0,\text{min}}, D_{0,\text{prod}})$. For any such value of $D_0$, Lemma 13 gives the desired convergence of $d_1(\mathbb{P}_{D_0^n})$ to $d_1(\mathbb{P}_{D_0^n})$ almost surely. We can then take the intersection over countably many values of $D_0$ (since we are restricting to rational numbers), and the corollary follows since a countable intersection of almost-sure events still holds almost surely.

Let $D_{0,\text{max}} = \frac{1}{2}d_0^\infty(X, X^*)$ be the $d_0$-distortion incurred at block length $n$. We know from Corollary 1 that with probability $1 - o(1)$, it holds that $D_{0,\text{max}} = D_0^\infty \cdot (1 + o(1))$, where $D_0^\infty = D_0^\text{id}(Q^\times, R)$. Hence, for any $\delta > 0$, we can assume that $D_{0,\text{max}} \in [D_0^\infty - \delta, D_0^\infty + \delta]$ for sufficiently large $n$. In addition, we note that this choice of $D_0^\infty$ satisfies the condition $D_0^\infty \in (D_{0,\text{min}}, D_{0,\text{prod}})$ in Corollary 2, due to the fact that $R \in (0, R_{\text{max}})$. Thus, Corollary 2 reveals that $d_1(\mathbb{P}_{D_0^n}) \rightarrow d_1(\mathbb{P}_{D_0^n})$ for all rational values of $D_0 \in [D_0^\infty - \delta, D_0^\infty + \delta]$, with probability one.

With $D_{0,n}$ being the minimum $d_0$-distortion induced by the codebook, we let $D_{0,\text{max}}$ be a rational number arbitrarily close to $D_{0,n}^*$, and we require it to be sufficiently close such that $D_{0,\text{max}}^*$ is below all the other distortion levels (not equal to $D_{0,n}^*$) induced by the codewords. Since the rationals are dense, this is always possible. There could be multiple codewords with distortion $D_{0,n}^*$, but even if this is the case, our assumed tie-breaking strategy (namely, choosing the smallest index) implies that $X^*$ must be the first codeword having distortion at most $D_{0,n}^*$. By the conclusion of the previous paragraph,
it follows that the $d_1$-distortion incurred is $d_1(\tilde{P}_X^{(\alpha_0,n)}) \cdot (1 + o(1))$.

To deduce the distortion level in (37) (with $D_0$ replaced by $D_0^n$), it remains to argue that $d_1(\tilde{P}_X^{(D_0^n,n)}) \cdot (1 + o(1))$ reduces to $d_1(\tilde{P}_X^{*X}) \cdot (1 + o(1))$, where in the latter expression, $\tilde{P}_X^{*X}$ is defined with respect to $D_0^n$. To see this, recall that we already established $D_0,n = D_0^n \cdot (1 + o(1))$ (with probability approaching one), which in turn implies that the rational rounded version can also be chosen to satisfy $D_0,n = D_∞ \cdot (1 + o(1))$. It is shown in [12, Sec. II-B] that the function $\Lambda(\cdot)$ defining $\lambda∗$ in (38) satisfies $\Lambda′′(\lambda) > 0$ for all $\lambda < 0$. Thus, with $\lambda∗$ being the solution to $\Lambda(\lambda) = D_0$ with $D_0 ∈ (D_0, min, D_0, prod)$, we have that $\lambda∗$ is a continuous function of $D_0$. Thus, with $D_0,n → D_∞$, the corresponding $\lambda∗$ value similarly approaches the value corresponding to $D_0^n$.

We now argue that the right-hand side of (38) is continuous with respect to $\lambda∗$ for $Π_X$-almost all $x$ and any $\tilde{x}$. Observe that this ratio is non-negative and is the reciprocal of $\mathbb{E}_Q[e^{\lambda∗(d_0(x,X) − d_0(x,\tilde{x}))}]$, which is a convex function of $\lambda∗$ and is therefore continuous in the region where it is finite. Thus, since the numerator in (38) is a fixed positive constant, it only remains to handle the possibility of dividing by zero or infinity. However, we claim that the denominator $\mathbb{E}_Q[e^{\lambda∗(d_0(x,X))}]$ is at most one, and is bounded away from zero for $Π_X$-almost all $x$. The former property follows from $\lambda∗ < 0$ and $d_0(\cdot, \cdot) ≥ 0$, and the latter follows from $\mathbb{E}_Q[e^{\lambda∗(d_0(x,X))}] ≥ e^{\lambda∗D_0, prod}$ via Jensen's inequality. The desired continuity with respect to $\lambda∗$ readily follows.

This pointwise convergence for the Radon-Nikodym derivative implies that the distortion $d_1(\tilde{P}_X^{(D_0^n,n)}) \cdot (1 + o(1))$ reduces to $d_1(\tilde{P}_X^{*X}) \cdot (1 + o(1)) = \mathbb{E}_Q[|d_1(X,X)|] \cdot (1 + o(1))$, where $\tilde{P}_X^{*X}$ is defined with respect to $D_0^n$ as stated above. This completes the proof of Theorem 3; we derived the form given in (37)–(38), but by Lemma 5, the form (9)–(10) is equivalent.

D. Discussion on Lemma 13 (Empirical Averages for the First Matching Codeword)

The proof of Lemma 12 in [13] is based on applying a conditional large-deviations analysis to the two-dimensional vector $(\frac{1}{n}d_0^0(\hat{X}, \hat{X}), \tilde{P}_X(E))$, where $\hat{X} \sim Q_{\hat{X}}$, and the conditioning is on $d_0^0(\hat{X}, \hat{X}) ≤ nD_0$ (and $X = x$). This conditioning comes from $\hat{X}$ being defined in Lemma 12 to be the first codeword having this property.

Here we outline how a similar approach can be taken for Lemma 13, considering the two-dimensional vector $(\frac{1}{n}d_0^0(\hat{X}, \hat{X}), \frac{1}{n}d_0^0(\hat{X}, \hat{X}))$, written as $(d_0(\tilde{P}_n), d_1(\tilde{P}_n))$ for short. According to the conditioning event $d_0(\tilde{P}_n) ≤ D_0$ and the definition of conditional probability, we are interested in the ratio

$$
\frac{\mathbb{P}[d_0(\tilde{P}_n) \leq D_0 \text{ and } d_1(\tilde{P}_n) \leq d_1(\tilde{P}_X^{*X}) - \delta]}{\mathbb{P}[d_0(\tilde{P}_n) \leq D_0]}.
$$

as well as the analog with the second event replaced by $d_1(\tilde{P}_n) ≥ d_1(\tilde{P}_X^{*X}) + \delta$. The two are handled similarly, so we focus only on (84). The denominator is a standard quantity characterized in [12], so the main challenge is in upper bounding the numerator.

The relevant log-moment generating function is

$$
\tilde{\Lambda}_n(\lambda) = \log \mathbb{E}_Q[e^{\lambda d_0(\tilde{P}_n)} + \lambda_1 d_1(\tilde{P}_n)],
$$

where $\lambda = (\lambda_0, \lambda_1) ∈ (-∞, 0]^2$, and by the ergodic theorem, $\frac{1}{n}\Lambda_n(nλ)$ converges almost surely to the following for $Π_X$ almost all $x$ (analogous to [13, Eq. (22)):

$$
\tilde{\Lambda}(\lambda) = \mathbb{E}_Q[\log \mathbb{E}_Q[e^{\lambda d_0(\tilde{P}_X,X) + \lambda_1 d_1(\tilde{P}_X,X)}]].
$$

Since $d_0$ and $d_1$ are non-negative and $\lambda ∈ (-∞, 0]^2$, we have $\tilde{\Lambda}(\lambda) ≤ 0$. On the other hand, by Jensen's inequality, $\tilde{\Lambda}(λ) ≥ λ_0 D_0, prod + λ_1 D_1, prod$; thus, the assumptions $D_0, prod < ∞$ and $D_1, prod < ∞$ are sufficient to ensure finite $\tilde{\Lambda}(\lambda)$. This is the only assumption made in Lemma 13 but not Lemma 12.

Rather than giving the technical details of the large-deviations analysis, we provide the main intuition. It is known that it weren't for the event $d_1(\tilde{P}_n) ≤ d_1(\tilde{P}_X^{*X}) - \delta$ in the numerator of (84), the limiting distribution of $\tilde{P}_n$ would be $\tilde{P}_X^{*X}$ given in (38) [12, 13]. The idea is that the event $d_1(\tilde{P}_n) ≤ d_1(\tilde{P}_X^{*X}) - \delta$ makes this distribution infeasible, and accordingly, $\tilde{P}_n$ is forced towards a slightly different distribution. This makes the exponent of the numerator in (84) strictly larger than that of the denominator, giving the desired exponential decay to zero in the first part of Lemma 13.

By the same argument as that used for Lemma 12 (see [13]), the second part of Lemma 13 follows easily from the first: Summing the left-hand side of (79) over all $n$ gives a finite value due to the exponential decay, and then the Borel-Cantelli lemma implies that with probability one, the event $|d_1(\tilde{P}_n) - d_1(\tilde{P}_X^{*X})| > δ$ occurs for at most finitely many $n$. Since $δ$ is arbitrarily small, it follows that $d_1(\tilde{P}_n) → d_1(\tilde{P}_X^{*X})$ almost surely, as desired.

ACKNOWLEDGMENT

The authors are very grateful to Ioannis Kontoyiannis for bringing their attention to the article [13] and providing them with helpful comments.

REFERENCES

[1] C. E. Shannon, “Coding theorems for a discrete source with a fidelity criterion,” *IRE Nat. Conv. Rec.*, vol. 4, pp. 325–350, Mar. 1959.

[2] A. Lapidoth, “On the role of mismatch in rate distortion theory,” *IEEE Trans. Inf. Theory*, vol. 43, no. 1, pp. 38–47, Jan. 1997.

[3] J. Scarlett, A. G. Fábregas, A. Somekh-Baruch, and A. Martinez, “Information-theoretic foundations of mismatched decoding,” *Found. Trends Commun. Inf. Theory*, vol. 17, pp. 149–401, Aug. 2020.

[4] Z. Zhang and T. Berger, “Estimation via compressed information,” *IEEE Trans. Inf. Theory*, vol. 34, no. 2, pp. 198–211, Mar. 1988.

[5] A. Kipnis and G. Reeves, “Gaussian approximation of quantization error for estimation from compressed data,” *IEEE Trans. Inf. Theory*, vol. 67, no. 8, pp. 5562–5579, Aug. 2021.

[6] I. Csiszár and P. Narayan, “Channel capacity for a given decoding metric,” *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 53–43, Jan. 1995.

[7] A. Somekh-Baruch, “Multi-letter converse bounds for the mismatched discrete memoryless channel with an additive metric,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 531–535.

[8] I. E. Telatar, A. Lapidoth, and A. Ganti, “Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit,” *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2315–2328, Nov. 2000.
[9] J. Scarlett, A. Martínez, and A. Guillén i Fabregas, “Multiuser random coding techniques for mismatched decoding,” *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 3950–3970, Jul. 2016.

[10] A. Lapidoth, “Mismatched decoding and the multiple-access channel,” *IEEE Trans. Inf. Theory*, vol. 42, no. 5, pp. 1439–1452, Sep. 1996.

[11] A. Somekh-Baruch, “On achievable rates and error exponents for channels with mismatched decoding,” *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 205–218, Feb. 2015.

[12] A. Dembo and L. Kontoyiannis, “Source coding, large deviations, and approximate pattern matching,” *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1590–1615, Jun. 2002.

[13] I. Kontoyiannis and R. Zamir, “Mismatched codebooks and the role of entropy coding in lossy data compression,” *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 1922–1938, May 2006.

[14] M. Kanabar and J. Scarlett, “Mismatched-rate-distortion theory: Ensembles, bounds, and general alphabets,” 2022, arXiv:2203.15193.

[15] E. Akyl, C. Langbort, and T. Basar, “Information-theoretic approach to strategic communication as a hierarchical game,” *Proc. IEEE*, vol. 105, no. 2, pp. 205–218, Feb. 2017.

[16] M. Le Treust and T. Tomala, “Point-to-point strategic communication,” in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2021, pp. 1–5.

[17] A. Mahmood and A. B. Wagner, “Lossy compression with universal strategic communication as a hierarchical game,” *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 3950–3970, Jul. 2016.

[18] D. J. Sakrison, “The rate of a class of random processes,” *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2377–2382, Nov. 1998.

[19] D. J. Sakrison, “The rate distortion function for a class of sources,” *IEEE Trans. Inf. Theory*, vol. 44, no. 1, pp. 47–65, Jan. 1998.

[20] E.-H. Yang and J. C. Kieffer, “On the performance of data compression algorithm Fédérale de Lausanne (EPFL), Switzerland. Since January 2018, he has been Assistant Professor with the Department of Computer Science and the Department of Mathematics, National University of Singapore. His research interests include information theory, machine learning, signal processing, and high-dimensional statistics. He received the Singapore National Research Foundation (NRF) Fellowship and the NUS Presidential Young Professorship Award.