

On the Rate-Distortion-Perception Function

Jun Chen, Lei Yu, Jia Wang, Wuxian Shi, Yiqun Ge, Wen Tong

Abstract—Rate-distortion-perception theory generalizes Shannon’s rate-distortion theory by introducing a constraint on the perceptual quality of the output. The perception constraint complements the conventional distortion constraint and aims to enforce distribution-level consistencies. In this new theory, the information-theoretic limit is characterized by the rate-distortion-perception function. Although a coding theorem for the rate-distortion-perception function has recently been established, the fundamental nature of the optimal coding schemes remains unclear, especially regarding the role of randomness in encoding and decoding. It is shown in the present work that except for certain extreme cases, the rate-distortion-perception function is achievable by deterministic codes. This paper also clarifies the subtle differences between two notions of perfect perceptual quality and explores some alternative formulations of the perception constraint.

Index Terms—Common randomness, divergence, maximal coupling, perceptual quality, rate-distortion, soft-covering lemma, squared error, total variation distance.

I. INTRODUCTION

For a Polish metric space \mathcal{X} , let $(\mathcal{X}, \mathbb{B}(\mathcal{X}))$ be the Borel measurable space induced by the metric. Let $\mathcal{P}(\mathcal{X})$ denote the set of distributions (i.e., probability measures) defined over $(\mathcal{X}, \mathbb{B}(\mathcal{X}))$, and let X be a random variable with distribution $p_X \in \mathcal{P}(\mathcal{X})$. Moreover, let $\Delta : \mathcal{X} \times \mathcal{X} \rightarrow [0, \infty)$ be a (measurable) distortion function with $\Delta(x, \hat{x}) = 0 \Leftrightarrow x = \hat{x}$, and let $d : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \rightarrow [0, \infty]$ be a divergence with $d(p_X, p_{\hat{X}}) = 0 \Leftrightarrow p_X = p_{\hat{X}}$. The rate-distortion-perception function for X is given by

$$R(D, P) := \inf_{p_{\hat{X}|X}} I(X; \hat{X})$$

$$\text{subject to } \mathbb{E}[\Delta(X, \hat{X})] \leq D, \quad (1)$$

$$d(p_X, p_{\hat{X}}) \leq P, \quad (2)$$

where the infimum is taken over all channels (i.e., Markov kernels) $p_{\hat{X}|X}$ from \mathcal{X} to itself satisfying (1) and (2). The rate-distortion-perception function generalizes the classical rate-distortion function by complementing the conventional distortion constraint (1) with a perception constraint (2). The rationale behind (2) is that probability distributions have important perceptual implications, thus constraining the divergence between

p_X and $p_{\hat{X}}$ helps enforce the perceptual consistencies between the input and the output [1].

Since its inception in the seminal paper by Blau and Michaeli [2], rate-distortion-perception theory has received considerable attention in the machine learning community. A coding theorem for the rate-distortion-perception function has also been established recently [3] (see [4], [5] for some variants of this coding theorem). It is worth mentioning that there exist closely related works in the signal processing and information theory literature on quantizer design with a prescribed output distribution motivated by perceptual considerations [6]–[9]. This line of research culminates in [10] with a rate-distortion theory for output constrained lossy source coding. These two theories are intimately connected in the sense that the perception constraint is exactly meant to constrain the output distribution. On the other hand, they also have noticeable differences. In particular, [10] focuses on the case where the output is an i.i.d. sequence whereas the formulation in [2], [3] does not directly impose this restriction. It will be seen that this difference has implications in terms of the rate-distortion tradeoff.

To understand the motivation behind the present work, it is instructive to consider the following example first introduced in [11]. Let S be uniformly distributed over the unit circle $\mathcal{S} := \{s \in \mathbb{R}^2 : \|s\|_2 = 1\}$, where $\|\cdot\|_p$ is the p -norm. The question is how to minimize the expected distortion $\mathbb{E}[\|S - \hat{S}\|_2^2]$ if S is encoded using 1 bit while the reconstruction \hat{S} is required to meet the perfect perceptual quality constraint (i.e., \hat{S} is also uniformly distributed over \mathcal{S}). Let $\theta(s)$ denotes the angle of s for any $s \in \mathcal{S}$. Two coding schemes are studied in [11]. For the first coding scheme, the encoding operation is given by

$$K := \begin{cases} 0, & \theta(S) \in [0, \pi), \\ 1, & \theta(S) \in [\pi, 2\pi), \end{cases}$$

and the decoding operation is given by

$$\hat{S} := (\cos((K + W)\pi), \sin((K + W)\pi)),$$

where W is uniformly distributed over $[0, 1)$ and is independent of S . One can readily verify that the resulting expected distortion

$$\mathbb{E}[\|S - \hat{S}\|_2^2] = 2 - \frac{8}{\pi^2}.$$

In fact, it is shown in [11] that this is the minimum achievable distortion with private randomness only (i.e., the random seed at the decoder is independent of that at the encoder). The second coding scheme makes use of W at both the encoder and the decoder. Specifically, the encoding operation is given by

$$K := \begin{cases} 0, & \frac{\theta(S)}{\pi} + W \in [0, 1) \cup [2, 3), \\ 1, & \frac{\theta(S)}{\pi} + W \in [1, 2), \end{cases}$$

while the decoding operation is given by

$$\hat{S} := (\cos((K - W)\pi), \sin((K - W)\pi)).$$

In this case, we have

$$\mathbb{E}[||S - \hat{S}||^2] = 2 - \frac{4}{\pi} < 2 - \frac{8}{\pi^2},$$

which clearly shows the advantage of common randomness over private randomness.

The above toy example in the one-shot setting naturally leads to the question whether the same phenomenon appears in the asymptotic setting where many data points are encoded at once. We shall show that the answer depends critically on the definition of perfect perceptual quality. Specifically, in the asymptotic setting, there are two notions of perfect perceptual quality: weak-sense and strong-sense; the advantage of common randomness over private randomness manifests under the strong-sense perfect perceptual quality constraint but not under the weak-sense version. Moreover, if the weak-sense perfect perceptual quality constraint is further relaxed by allowing slight imperfection, then no randomness is needed at all. We would like to point out that the difference between common randomness and private randomness has been investigated in the context of output constrained lossy source coding [10], which has important implications here, especially with respect to the case of strong-sense perfect perceptual quality.

The rest of this paper is organized as follows. Section II contains coding theorems for various types of coding systems; in particular, it is shown that except for certain extreme cases, the rate-distortion-perception function is achievable by deterministic codes. Section III is devoted to the clarification of the subtle differences between two different notions of perfect perceptual quality. Some alternative formulations of the perception constraint are explored in Section IV. Section V concludes the paper.

Notation: We use p_X^n to denote the product of n copies of p_X . We use $p_{Y|X}$ to denote a channel (a regular conditional distribution or a Markov kernel), which associates to each point $x \in \mathcal{X}$ a probability measure $p_{Y|X}(\cdot|x)$ such that, for every measurable set $B \subseteq \mathcal{Y}$, the map

$x \mapsto p_{Y|X}(B|x)$ is measurable with respect to the σ -algebra on \mathcal{X} . A distribution is discrete if it can be written as $\sum_i p_i \delta_{x_i}$ for some countable number of points x_i and positive values p_i such that $\sum_i p_i = 1$, where δ_x is the Dirac measure at x . For a discrete distribution, its support is defined as the set of points at which the probability masses are positive. For two distributions p_X and $p_{\hat{X}}$, we use $\Pi(p_X, p_{\hat{X}})$ to denote the set of couplings of p_X and $p_{\hat{X}}$ (i.e., the set of joint distributions $p_{X\hat{X}}$ with marginals p_X and $p_{\hat{X}}$). Let $\mathcal{B}(\rho)$ and $\mathcal{N}(\mu, \sigma^2)$ denote respectively the Bernoulli distribution with parameter ρ and the Gaussian distribution with mean μ and variance σ^2 . The cardinality of set \mathcal{S} is written as $|\mathcal{S}|$. The binary entropy function is represented by $H_b(\cdot)$, i.e., $H_b(a) := -a \log(a) - (1-a) \log(1-a)$ for $a \in [0, 1]$. Define $1_{\mathcal{E}}(\cdot, \cdot)$ to be an indicator function in the sense that $1_{\mathcal{E}}(x, \hat{x}) = 1$ if $(x, \hat{x}) \in \mathcal{E}$ and $1_{\mathcal{E}}(x, \hat{x}) = 0$ otherwise. Let $\text{Unif}[i : j]$ denote the uniform distribution over $[i : j]$, where $[i : j] := \{i, i+1, \dots, j\}$ for integers $i \leq j$. Throughout this paper, the base of the logarithm function is 2.

II. CODING THEOREMS

Let $\{X_t\}_{t=1}^{\infty}$ be an i.i.d. process with marginal distribution p_X .

Definition 1. *Given distortion constraint D and perception constraint P , rate R is said to be achievable with common randomness if for all sufficiently large n , there exist shared seed distribution p_Q on a Polish space \mathcal{Q} , encoding distribution $p_{Z|X^n Q}$ with \mathcal{Z} countable (equipped with the Hamming metric), and decoding distribution $p_{\hat{X}^n|Z Q}$ with $\hat{\mathcal{X}} = \mathcal{X}$ such that the induced joint distribution $p_{X^n Q Z \hat{X}^n} := p_X^n p_Q p_{Z|X^n Q} p_{\hat{X}^n|Z Q}$ satisfies*

$$\frac{1}{n} H(Z|Q) \leq R, \quad (3)$$

$$\frac{1}{n} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \hat{X}_t)] \leq D, \quad (4)$$

$$d(p_X, p_{\hat{X}_t}) \leq P, \quad t \in [1 : n]. \quad (5)$$

The infimum of such achievable R is denoted by $R_{\text{cr}}(D, P)$. The achievable rate with private randomness is defined in the same way except that the encoder and the decoder do not have access to a shared random seed (i.e., Q is set to be a constant); the corresponding fundamental limit is denoted $R_{\text{pr}}(D, P)$. If no randomness is allowed at all (i.e., the encoder output Z is required to be a deterministic function of X^n while the decoder output \hat{X}^n is required to be a deterministic function of Z), we denote the fundamental limit by $R_{\text{nr}}(D, P)$.

Remark 1. The rationale behind (3) is as follows. Given any realization $Q = q$, random variable Z can be represented by a variable-length code of average length no more than $H(Z|Q = q) + 1$. Normalizing $H(Z|Q = q) + 1$ by n and taking the expectation with respect to Q yields $\frac{1}{n}H(Z|Q) + \frac{1}{n}$. The extra factor $\frac{1}{n}$ is negligible as $n \rightarrow \infty$. Alternatively, one can replace (3) by $\frac{1}{n} \log |\mathcal{Z}| \leq R$, where \mathcal{Z} is the alphabet over which Z is defined. This variant is more suitable for fixed-length codes. As far as $R_{\text{cr}}(D, P)$, $R_{\text{pr}}(D, P)$, and $R_{\text{nr}}(D, P)$ are concerned, the difference between variable-length codes and fixed-length codes only manifests in certain extreme cases (say, $D = 0$).

Remark 2. It is easy to establish the following ordering by invoking the operational meanings of the relevant quantities:

$$R_{\text{cr}}(D, P) \leq R_{\text{pr}}(D, P) \leq R_{\text{nr}}(D, P). \quad (6)$$

In this paper, we will make some regularity assumptions along the way when they are needed for establishing certain technical results. This first one is as follows:

Assumption 1. $d(\cdot, \cdot)$ is convex in its second argument.

This assumption is quite mild as it is satisfied by f -divergence and Rényi divergence [12], [13] as well as those taking the form¹ of $(p_X, p_{\hat{X}}) \mapsto \inf_{p_{X, \hat{X}} \in \Pi(p_X, p_{\hat{X}})} \mathbb{E}[c(X, \hat{X})]$, where c is a (measurable) cost function and $\Pi(p_X, p_{\hat{X}})$ denotes the set of all couplings of p_X and $p_{\hat{X}}$.

The following result, due to Theis and Wagner [3, Theorem 3], provides a computable characterization of $R_{\text{cr}}(D, P)$ by linking it to $R(D, P)$.

Theorem 1. Under Assumption 1,

$$R_{\text{cr}}(D, P) = R(D, P)$$

for $D \geq 0$ and $P \geq 0$.

Remark 3. It can be shown using the standard converse argument that Theorem 1 continues to hold if (5) is weakened to

$$\frac{1}{n} \sum_{t=1}^n d(p_X, p_{\hat{X}_t}) \leq P$$

or even further weakened to

$$d\left(p_X, \frac{1}{n} \sum_{t=1}^n p_{\hat{X}_t}\right) \leq P.$$

¹Such divergences arise naturally in the theory of optimal transport. In particular, if c is the metric on \mathcal{X} , then $d(p_X, p_{\hat{X}}) := \inf_{p_{X, \hat{X}} \in \Pi(p_X, p_{\hat{X}})} \mathbb{E}[c(X, \hat{X})]$ is the 1-Wasserstein distance.

Remark 4. In fact, Theorem 1 is established in [3] under the more restrictive distortion constraint (as compared to (4))

$$\mathbb{E}[\Delta(X_t, \hat{X}_t)] \leq D, \quad t \in [1 : n],$$

without Assumption 1.

The proof of Theorem 1 in [3] relies on the strong functional representation lemma [14] and consequently makes use of common randomness in an essential way. Thus an open problem is posed in [3], which asks whether the same result can be established under weaker conditions. The next result provides an affirmative answer by showing that it suffices to use deterministic codes when $D > 0$ and $P > 0$. To state this result precisely, we need the following technical assumption:

Assumption 2. For any $D > 0$ and $P > 0$, we have

$$R(D, P) < \infty; \quad (7)$$

moreover, given any $\epsilon > 0$, there exists a discrete random variable \tilde{X} with its support $\tilde{\mathcal{X}}$ satisfying $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ and $|\tilde{\mathcal{X}}| < \infty$ such that

$$I(X; \tilde{X}) \leq R(D, P) + \epsilon, \quad (8)$$

$$\mathbb{E}[\Delta(X, \tilde{X})] \leq D + \epsilon, \quad (9)$$

$$\mathbb{E}\left[\max_{\tilde{x} \in \tilde{\mathcal{X}}} \Delta(X, \tilde{x})\right] < \infty, \quad (10)$$

$$d(p_X, p_{\tilde{X}}) \leq P + \epsilon, \quad (11)$$

$$d(p_X, \gamma) < \infty \text{ for all distributions } \gamma \text{ supported on } \tilde{\mathcal{X}}. \quad (12)$$

Note that (7) is a prerequisite for the existence of a coding theorem non-void for all $D > 0$ and $P > 0$. It clearly holds when $|\mathcal{X}| < \infty$ since $R(D, P) \leq R(0, 0) = H(X) < \infty$. When $|\mathcal{X}| < \infty$, (8)–(11) are also trivially true. In general, by the definition of $R(D, P)$, there exists a random variable \hat{X} such that

$$I(X; \hat{X}) \leq R(D, P) + \epsilon,$$

$$\mathbb{E}[\Delta(X, \hat{X})] \leq D,$$

$$d(p_X, p_{\hat{X}}) \leq P.$$

If we think of \tilde{X} as a quantized version of \hat{X} , then (8) is automatically satisfied due to the data processing inequality, (10) is to ensure that no quantization output might be catastrophically bad while (9) and (11) basically require that $\mathbb{E}[\Delta(X, \hat{X})]$ and $d(p_X, p_{\hat{X}})$ are not too sensitive to the discretization of $p_{\hat{X}}$ (which can be viewed as a form of weak convergence requirement). Finally, (12) is a natural consequence² of (11) for any

²Actually we only need $d(p_X, \gamma) < \infty$ for γ in a small neighborhood of $p_{\hat{X}}$ confined to $\mathcal{P}(\tilde{\mathcal{X}})$.

reasonably behaved divergence. So Assumption 2 is basically always true when $|\mathcal{X}| < \infty$. In addition, we verify in Appendix A that Assumption 2 holds for the case of square-integrable random variable, squared distortion measure, and squared quadratic Wasserstein distance (i.e., $\mathbb{E}[X^2] < \infty$, $\Delta(x, \hat{x}) := (x - \hat{x})^2$, and $d(p_X, p_{\hat{X}}) := \inf_{p_{X\hat{X}} \in \Pi(p_X, p_{\hat{X}})} \mathbb{E}[(X - \hat{X})^2]$). It is worth mentioning that with deterministic encoding and decoding performed at any finite rate, the reconstruction is inevitably discrete. Hence, there are reasons to believe that Assumption 2 cannot be substantially relaxed. Assumption 2 also has some nice implications. Specifically, together with Assumption 1, (7) implies that $R(D, P)$ is convex and consequently continuous in (D, P) for $D > 0$ and $P > 0$ while (12) implies that $d(p_X, \gamma)$ is continuous in γ over the interior of the probability simplex defined on $\tilde{\mathcal{X}}$.

Theorem 2. *Under Assumptions 1 and 2,*

$$R_{\text{nr}}(D, P) = R(D, P)$$

for $D > 0$ and $P > 0$.

Proof: The detailed proof can be found in Appendix B. The basic idea is that, in the asymptotic setting, it is possible to leverage the aggregated randomness to simultaneously shape the marginal distributions of all output symbols into the desired form via proper deterministic encoding and decoding even though the bit rate might be far below the corresponding entropy. Consider the toy example in Section I. In the one-shot setting, it is clearly impossible to simulate a uniform distribution over the unit circle using 1 bit if the decoder is required to be deterministic. However, in the asymptotic setting, even if the rate remains 1 bit per data point, we are still able to accumulate enough randomness, which can be shared strategically by the reconstructed points in such a way that they all acquire an approximate uniform distribution over the unit circle. ■

The case $D = 0$ corresponds to the conventional zero-error source coding problem [15], for which there is no loss of optimality in restricting the encoder and the decoder to be deterministic. Moreover, it is clear that the perception constraint becomes superfluous when $D = 0$.

Theorem 3. *For $P \geq 0$,*

$$R_{\text{cr}}(0, P) = R_{\text{pr}}(0, P) = R_{\text{nr}}(0, P) = R(0, P),$$

where

$$R(0, P) = \begin{cases} H(X), & p_X \text{ is a discrete distribution,} \\ \infty, & \text{otherwise.} \end{cases}$$

It remains to deal with the case $P = 0$. This is addressed by the next result, for which we need the following assumption:

Assumption 3. *For any $D > 0$ and $\epsilon > 0$, there exist a discrete random variable \tilde{X} and an arbitrary random variable \hat{X} on \mathcal{X} such that $X \leftrightarrow \tilde{X} \leftrightarrow \hat{X}$ form a Markov chain, the support of \tilde{X} , denoted $\tilde{\mathcal{X}}$, satisfies $|\tilde{\mathcal{X}}| < \infty$, and*

$$\begin{aligned} I(X; \tilde{X}) &\leq R(D, 0) + \epsilon, \\ \mathbb{E}[\Delta(X, \hat{X})] &\leq D, \\ p_{\hat{X}} &= p_X. \end{aligned}$$

Note that according to the definition of $R(D, 0)$, there exists a random variable \hat{X} such that

$$\begin{aligned} I(X; \hat{X}) &\leq R(D, 0) + \epsilon, \\ \mathbb{E}[\Delta(X, \hat{X})] &\leq D, \\ p_{\hat{X}} &= p_X. \end{aligned}$$

So Assumption 3 basically postulates the existence of a discrete random variable \tilde{X} sitting between X and \hat{X} with $I(X; \tilde{X}) \approx I(X; \hat{X})$. This is trivially true when $|\mathcal{X}| < \infty$. In general, this assumption is quite natural as even with the availability of private randomness, the interface between the encoder and the decoder remains discrete at any finite rate. We verify at the end of Appendix C that Assumption 3 holds for the case of square-integrable random variable and squared distortion measure.

Theorem 4. *Under Assumptions 1 and 3,*

$$R_{\text{pr}}(D, 0) = R(D, 0)$$

for $D > 0$.

Remark 5. *It is clear that deterministic encoder-decoder pairs are inadequate for achieving finite-valued $R(D, 0)$ when p_X is a continuous distribution or has an infinite entropy.*

Proof: See Appendix D. ■

III. ON DIFFERENT NOTIONS OF PERFECT PERCEPTUAL QUALITY

Note that setting $P = 0$ in (5) only ensures $p_{\hat{X}_t} = p_{X_t}$, $t \in [1 : n]$, which should be distinguished from the more restrictive constraint $p_{\hat{X}^n} = p_{X^n}$. We shall refer to the former as weak-sense perfect perceptual quality and the latter as strong-sense perfect perceptual quality. It is interesting to understand whether these two notions of perfect perceptual quality make any difference in terms of the rate-distortion tradeoff.

Let $R_{\text{cr}}(D) := R_{\text{cr}}(D, 0)$ and $R_{\text{pr}}(D) := R_{\text{pr}}(D, 0)$. In light of Theorems 1, 3, and 4, for $D \geq 0$,

$$R_{\text{cr}}(D) = R_{\text{pr}}(D) = \phi(D), \quad (13)$$

where

$$\begin{aligned} \phi(D) &:= R(D, 0) = \inf_{p_{\hat{X}|X}} I(X; \hat{X}) \\ &\text{subject to } \mathbb{E}[\Delta(X, \hat{X})] \leq D, \\ & p_{\hat{X}} = p_X. \end{aligned}$$

Now we proceed to define the counterparts of $R_{\text{cr}}(D)$ and $R_{\text{pr}}(D)$ under the strong-sense perfect perceptual quality constraint.

Definition 2. Given distortion constraint D , rate \tilde{R} is said to be achievable with common randomness under the strong-sense perfect perceptual quality constraint if for all sufficiently large n , there exist shared seed distribution p_Q (on a Polish space), encoding distribution $p_{Z|X^n Q}$ (with \mathcal{Z} countable), and decoding distribution $p_{\hat{X}^n|ZQ}$ (with $\hat{\mathcal{X}} = \mathcal{X}$) such that the induced joint distribution $p_{X^n Q Z \hat{X}^n} := p_X^n p_Q p_{\hat{X}^n|ZQ} p_{Z|X^n Q}$ satisfies

$$\begin{aligned} \frac{1}{n} H(Z|Q) &\leq \tilde{R}, \\ \frac{1}{n} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \hat{X}_t)] &\leq D, \\ p_{\hat{X}^n} &= p_X^n. \end{aligned}$$

The infimum of such achievable \tilde{R} is denoted $\tilde{R}_{\text{cr}}(D)$. In the case of private randomness only, the corresponding limit is denoted $\tilde{R}_{\text{pr}}(D)$.

Remark 6. Analogously to (6), we have

$$\tilde{R}_{\text{cr}}(D) \leq \tilde{R}_{\text{pr}}(D). \quad (14)$$

Moreover, since strong-sense perfect perceptual quality implies weak-sense perfect perceptual quality, it follows that

$$R_{\text{cr}}(D) \leq \tilde{R}_{\text{cr}}(D), \quad R_{\text{pr}}(D) \leq \tilde{R}_{\text{pr}}(D).$$

Remark 7. Under the strong-sense perfect perceptual quality constraint, requiring the encoder and the decoder to be deterministic trivializes the problem as the encoder-decoder pair is basically forced to establish a one-to-one mapping between the input and the output.

The following result, together with (13), shows that in the presence of common randomness, the difference between weak-sense perfect perceptual quality and strong-sense perfect perceptual quality has no impact on the fundamental rate-distortion tradeoff.

Theorem 5. For $D \geq 0$

$$R_{\text{cr}}(D) = \tilde{R}_{\text{cr}}(D) = \phi(D). \quad (15)$$

Proof: This result can be deduced from the proof of [9, Theorem 7] (see also [7]). ■

The following result, together with (13), shows that in the case of private randomness only, the two different notions of perfect randomness indeed lead to different rate-distortion tradeoffs. Along with Theorem 5, it also indicates that under the strong-sense perfect perceptual quality constraint, common randomness is generally more powerful than private randomness, which should be contrasted with the fact that under the weak-sense perfect perceptual quality constraint, the difference between common randomness and private randomness is immaterial (see Theorem 4).

We first introduce a definition, which is an extended version of [16, Definition 3].

Definition 3. A tuple $(p_X, p_{\hat{X}}, \Delta)$ of source distribution, reconstruction distribution, and distortion measure is said to be uniformly integrable if for every $\epsilon > 0$, there exists $\delta > 0$ such that $\sup_{p_{X\hat{X}}, \mathcal{E}} \mathbb{E}[\Delta(X, \hat{X}) 1_{\mathcal{E}}(X, \hat{X})] \leq \epsilon$, where the supremum is over all $p_{X\hat{X}} \in \Pi(p_X, p_{\hat{X}})$ and all measurable events \mathcal{E} with $\mathbb{P}((X, \hat{X}) \in \mathcal{E}) \leq \delta$.

Note that $(p_X, p_{\hat{X}}, \Delta)$ is uniformly integrable if Δ is bounded, i.e., $\sup_{x, \hat{x} \in \mathcal{X}} \Delta(x, \hat{x}) < \infty$, which is trivially true when $|\mathcal{X}| < \infty$. Moreover, in Appendix E, we verify uniform integrability for square-integrable X and \hat{X} paired with squared distortion measure.

Theorem 6. If (p_X, p_X, Δ) is uniformly integrable, then

$$\tilde{R}_{\text{pr}}(D) = \varphi(D) \quad (16)$$

for $D \geq 0$, where³

$$\varphi(D) := \inf_{p_{U\hat{X}|X}} \max\{I(X; U), I(\hat{X}; U)\}$$

$$\text{subject to } \mathbb{E}[\Delta(X, \hat{X})] \leq D, \quad (17)$$

$$p_{\hat{X}U|X} = p_{U|X} p_{\hat{X}|U}, \quad (18)$$

$$p_{\hat{X}} = p_X. \quad (19)$$

Moreover, under the squared distortion measure (assuming $\mathcal{X} \subseteq \mathbb{R}$),

$$\varphi(D) = R\left(\frac{D}{2}\right), \quad (20)$$

³Here $\hat{\mathcal{X}} = \mathcal{X}$, and the infimum above is taken over all $p_{U\hat{X}|X}$ with U being a Polish space such that (17)-(19) hold. A similar convention applies to the bound in Theorem 10.

where

$$R\left(\frac{D}{2}\right) := \inf_{p_{V|X}} I(X; V) \quad (21)$$

$$\text{subject to } \mathbb{E}[(X - V)^2] \leq \frac{D}{2}. \quad (22)$$

Remark 8. $R\left(\frac{D}{2}\right)$ is interpreted as $R\left(\frac{D}{2}, \infty\right)$ in [2], [17]. This interpretation is actually not completely accurate. Note that $R(D)$ is the rate-distortion function with the output alphabet being \mathbb{R} as V is allowed to take any real value. In contrast, $R(D, \infty)$ is the rate-distortion function with the output alphabet being \mathcal{X} . In general, under the squared distortion measure,

$$R(D, 0) \leq R\left(\frac{D}{2}\right) \leq R\left(\frac{D}{2}, \infty\right),$$

where the second inequality becomes an equality when $\mathcal{X} = \mathbb{R}$. Note that the first inequality follows by (14), (15), (16), and (20) (see also [2, Theorem 2]).

Proof: One can specialize (16) from [10, Theorem 1] and [16, Theorem 2]. Moreover, (20) is implied by [17, Theorem 2] (see also [11, Theorem]). We give a simple proof of this fact in Appendix F. ■

Theorem 7. For $X \sim \mathcal{B}(\rho)$ with $\rho \in (0, \frac{1}{2}]$,

$$\phi(D) = \begin{cases} 2H_b(\rho) + \frac{2-2\rho-D}{2} \log\left(\frac{2-2\rho-D}{2}\right) + D \log\left(\frac{D}{2}\right) \\ + \frac{2\rho-D}{2} \log\left(\frac{2\rho-D}{2}\right), & D \in [0, 2\rho(1-\rho)), \\ 0, & D \in [2\rho(1-\rho), \infty), \end{cases}$$

$$\varphi(D) = \begin{cases} H_b(\rho) - H_b\left(\frac{1-\sqrt{1-2D}}{2}\right), & D \in [0, 2\rho(1-\rho)), \\ 0, & D \in [2\rho(1-\rho), \infty), \end{cases}$$

under the Hamming distortion measure (which coincides with the squared distortion measure when $\mathcal{X} = \{0, 1\}$).

Remark 9. In this case, we can deduce from [2, Equation (6)] that

$$R\left(\frac{D}{2}, \infty\right) = \begin{cases} H_b(\rho) - H_b\left(\frac{D}{2}\right), & D \in [0, 2\rho), \\ 0, & D \in [2\rho, \infty), \end{cases}$$

which is in general different from $\varphi(D)$. So [17, Theorem 2] should be interpreted with great caution.

For illustrative purposes, we plot $\phi(D)$ and $\varphi(D)$ for $X \sim \mathcal{B}\left(\frac{1}{4}\right)$ in Fig. 1.

Proof: The expression of $\phi(D)$ can be specialized from [2, Equation (6)]. The derivation of $\varphi(D)$ is given in Appendix G. ■

Theorem 8. For $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$\phi(D) = \begin{cases} \frac{1}{2} \log\left(\frac{4\sigma^4}{4\sigma^2 D - D^2}\right), & D \in [0, 2\sigma^2), \\ 0, & D \in [2\sigma^2, \infty), \end{cases}$$

$$\varphi(D) = \begin{cases} \frac{1}{2} \log\left(\frac{2\sigma^2}{D}\right), & D \in [0, 2\sigma^2), \\ 0, & D \in [2\sigma^2, \infty), \end{cases}$$

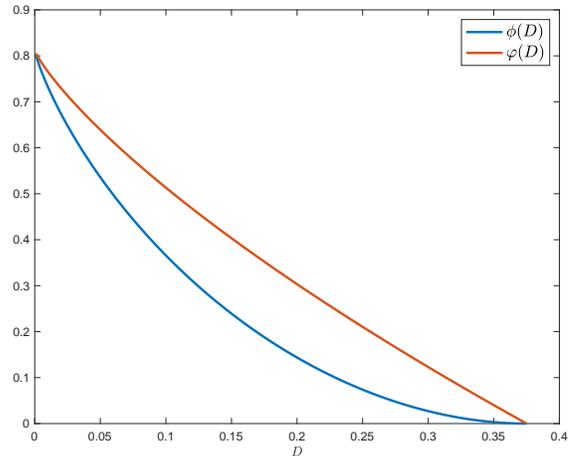


Fig. 1. Plots of $\phi(D)$ and $\varphi(D)$ for $X \sim \mathcal{B}\left(\frac{1}{4}\right)$.

under the squared distortion measure.

Remark 10. In this case, $\mathcal{X} = \mathbb{R}$ and consequently [17, Theorem 2] holds. Indeed, we can deduce from [18, Theorem 1] that

$$R\left(\frac{D}{2}, \infty\right) = \begin{cases} \frac{1}{2} \log\left(\frac{2\sigma^2}{D}\right), & D \in [0, 2\sigma^2), \\ 0, & D \in [2\sigma^2, \infty), \end{cases}$$

which coincides with $\varphi(D)$.

For illustrative purposes, we plot $\phi(D)$ and $\varphi(D)$ for $X \sim \mathcal{N}(0, 1)$ in Fig. 2.

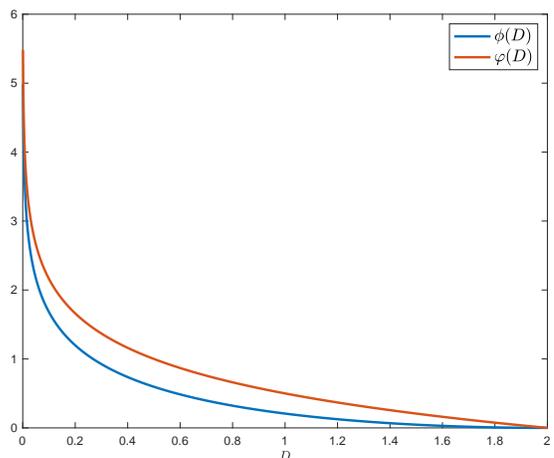


Fig. 2. Plots of $\phi(D)$ and $\varphi(D)$ for $X \sim \mathcal{N}(0, 1)$.

Proof: The expression of $\phi(D)$ can be specialized from [18, Theorem 1] (see also [7, Proposition 2]) while

the expression of $\varphi(D)$ can be obtained by invoking (20) and the fact [15] that for the quadratic Gaussian case,

$$R(D) = \begin{cases} \frac{1}{2} \log\left(\frac{\sigma^2}{D}\right), & D \in [0, \sigma^2), \\ 0, & D \in [2\sigma^2, \infty). \end{cases}$$

Various extensions can be found in [16], [19].

IV. ALTERNATIVE FORMULATIONS OF THE PERCEPTION CONSTRAINT

The fact that the perception constraint in (5) fails to capture the notion of strong-sense perfect perceptual quality motivates us to consider the following alternative formulation.

Definition 4. Given distortion constraint D and perception constraint P , rate \tilde{R} is said to be achievable with common randomness if for all sufficiently large n , there exist shared seed distribution p_Q (on a Polish space), encoding distribution $p_{Z|X^n Q}$ (with \mathcal{Z} countable), and decoding distribution $p_{\hat{X}^n|ZQ}$ (with $\hat{\mathcal{X}} = \mathcal{X}$) such that the induced joint distribution $p_{X^n Q Z \hat{X}^n} := p_{\hat{X}^n}^n p_Q p_{Z|X^n Q} p_{\hat{X}^n|ZQ}$ satisfies

$$\begin{aligned} \frac{1}{n} H(Z|Q) &\leq \tilde{R}, \\ \frac{1}{n} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \hat{X}_t)] &\leq D, \\ \frac{1}{n} d(p_{X^n}^n, p_{\hat{X}^n}) &\leq P. \end{aligned} \quad (23)$$

The infimum of such achievable \tilde{R} is denoted $\tilde{R}_{\text{cr}}(D, P)$. In the case of private randomness only, the corresponding limit is denoted $\tilde{R}_{\text{pr}}(D, P)$.

Note that the strong-sense perceptual quality constraint can be viewed as an extreme case of the new formulation since setting $P = 0$ in (23) forces $p_{\hat{X}^n} = p_{X^n}^n$. So we have $\tilde{R}_{\text{cr}}(D, 0) = \tilde{R}_{\text{cr}}(D)$ and $\tilde{R}_{\text{pr}}(D, 0) = \tilde{R}_{\text{pr}}(D)$. However, the multiletter nature of the perception constraint in (23) makes it challenging to obtain a computable characterization of $\tilde{R}_{\text{cr}}(D, P)$ and $\tilde{R}_{\text{pr}}(D, P)$ when $P > 0$. To ease the difficulty, we shall impose some further restrictions on $d(\cdot, \cdot)$:

Assumption 4. $d(\cdot, \cdot)$ is tensorizable in the sense that

$$d(\otimes_{t=1}^n p_{Y_t}, p_{\hat{Y}^n}) \geq \sum_{t=1}^n d(p_{Y_t}, p_{\hat{Y}_t}),$$

where $\otimes_{t=1}^n p_{Y_t}$ denotes the joint distribution formed by the product of marginals p_{Y_1}, \dots, p_{Y_n} .

Assumption 5. $d(\cdot, \cdot)$ is decomposable in the sense that

$$d(\otimes_{t=1}^n p_{Y_t}, \otimes_{t=1}^n p_{\hat{Y}_t}) = \sum_{t=1}^n d(p_{Y_t}, p_{\hat{Y}_t}).$$

Note that Assumptions 4 and 5 are satisfied by the Kullback-Leibler divergence and those taking the form of $\inf_{p_{Y^n \hat{Y}^n} \in \Pi(p_{Y^n}, p_{\hat{Y}^n})} \mathbb{E}[c(Y^n, \hat{Y}^n)]$ with an additive cost function c in the sense $c(y^n, \hat{y}^n) = \sum_{i=1}^n c'(y_i, \hat{y}_i)$ for some c' (e.g., $c(y^n, \hat{y}^n) := \|y^n - \hat{y}^n\|_p^p$).

Theorem 9. Under Assumptions 1, 4, and 5,

$$\tilde{R}_{\text{cr}}(D, P) = R(D, P)$$

for $D \geq 0$ and $P \geq 0$.

Proof: We first prove the converse part. Let \tilde{R} be an achievable rate with respect to distortion constraint D and perception constraint P . We have

$$\begin{aligned} \tilde{R} &\geq \frac{1}{n} H(Z|Q) \\ &\geq \frac{1}{n} I(X^n; Z|Q) \\ &\geq \frac{1}{n} I(X^n; \hat{X}^n|Q) \\ &= \frac{1}{n} \sum_{t=1}^n I(X_t; \hat{X}^n|Q, X^{t-1}) \\ &= \frac{1}{n} \sum_{t=1}^n I(X_t; \hat{X}^n, Q, X^{t-1}) \\ &\geq \frac{1}{n} \sum_{t=1}^n I(X_t; \hat{X}_t) \\ &= I(X_T; \hat{X}_T|T) \\ &= I(X_T; \hat{X}_T, T) \\ &\geq I(X_T; \hat{X}_T), \end{aligned} \quad (24)$$

where T is uniformly distributed over $[1 : n]$ and independent of (X^n, \hat{X}^n) . Moreover,

$$\begin{aligned} D &\geq \frac{1}{n} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \hat{X}_t)] \\ &= \mathbb{E}[\mathbb{E}[\Delta(X_T, \hat{X}_T)|T]] \\ &= \mathbb{E}[\Delta(X_T, \hat{X}_T)], \end{aligned} \quad (25)$$

and

$$\begin{aligned} P &\geq \frac{1}{n} d(p_{X^n}^n, p_{\hat{X}^n}) \\ &\geq \frac{1}{n} \sum_{t=1}^n d(p_{X_t}, p_{\hat{X}_t}) \end{aligned} \quad (26)$$

$$\begin{aligned} &= \frac{1}{n} \sum_{t=1}^n d(p_X, p_{\hat{X}_t}) \\ &\geq d\left(p_X, \frac{1}{n} \sum_{t=1}^n p_{\hat{X}_t}\right) \end{aligned} \quad (27)$$

$$= d(p_X, p_{\hat{X}_T}), \quad (28)$$

where (26) and (27) are due to respectively the tensorizability and convexity (in its second argument) of $d(\cdot, \cdot)$. Combining (24), (25), (28) and invoking the fact that $p_{X_T} = p_X$ proves $\tilde{R}_{\text{cr}}(D, P) \geq R(D, P)$.

Now it remains to prove the achievability part. We augment $\{X_t\}_{t=1}^{\infty}$ into a joint i.i.d. process $\{(X_t, \hat{X}_t)\}_{t=1}^{\infty}$ using a memoryless channel $p_{\hat{X}|X}$ satisfying $\mathbb{E}[\Delta(X, \hat{X})] \leq D$ and $d(p_X, p_{\hat{X}}) \leq P$. For any positive integer n ,

$$\frac{1}{n}d(p_{X^n}, p_{\hat{X}^n}) = d(p_X, p_{\hat{X}}) \leq P,$$

where the equality follows by the decomposability of $d(\cdot, \cdot)$. Moreover, by the strong functional representation lemma [14], there exists a random variable Q , independent of X^n , such that \hat{X}^n can be expressed as a deterministic function of (X^n, Q) and

$$\frac{1}{n}H(\hat{X}^n|Q) \leq I(X; \hat{X}) + \frac{1}{n}\log(nI(X; \hat{X}) + 1) + \frac{4}{n}.$$

Setting $Z = \hat{X}^n$ and sending $n \rightarrow \infty$ completes the proof. ■

Theorem 10. For $D \geq 0$ and $P \geq 0$,

$$\tilde{R}_{\text{pr}}(D, P) \leq R'(D, P),$$

where

$$R'(D, P) := \inf_{p_{U\hat{X}|X}} \max\{I(X; U), I(\hat{X}; U)\}$$

$$\text{subject to } \mathbb{E}[\Delta(X, \hat{X})] \leq D,$$

$$p_{\hat{X}U|X} = p_{U|X}p_{\hat{X}|U},$$

$$d(p_X, p_{\hat{X}}) \leq P,$$

$$(p_X, p_{\hat{X}}, \Delta) \text{ is uniformly integrable.}$$

Proof: Suppose that \hat{X}^n is constrained to be an i.i.d. sequence with marginal distribution $p_{\hat{X}}$ satisfying $d(p_X, p_{\hat{X}}) \leq P$. For such \hat{X}^n ,

$$\frac{1}{n}d(p_{X^n}, p_{\hat{X}^n}) = d(p_X, p_{\hat{X}}) \leq P,$$

where the equality follows by the decomposability of $d(\cdot, \cdot)$. Now the problem is converted to output constrained lossy source coding in the sense of [10], and consequently the desired upper bound can be deduced by following steps similar to those for the achievability part of [16, Theorem 2].

It can be shown that, with the output constrained to be i.i.d., this upper bound is in fact the best possible (assuming $d(\cdot, \cdot)$ is tensorizable, decomposable, and convex in its second argument). Moreover, when $P = 0$, the i.i.d. constraint is automatically satisfied and the upper bound is tight. Indeed, we have $R'(D, 0) = \varphi(D)$, which

is known to coincide with $\tilde{R}_{\text{pr}}(D, 0)$ (or equivalently, $\tilde{R}_{\text{pr}}(D)$) according to Theorem 6. On the other hand, it is unclear whether the i.i.d. constraint is redundant when $P > 0$. So the upper bound can potentially be loose in that regime. In contrast, as the proof of Theorem 9 indicates, with the presence of common randomness, the i.i.d. constraint incurs no penalty in terms of the rate-distortion-perception tradeoff. ■

Roughly speaking, distortion measures and perception measures in the rate-distortion-perception framework can be considered full-reference metrics and no-reference metrics, respectively. Take image compression as an example. To evaluate a compressed image, full-reference metrics compare it to the ground truth (i.e., the original version) while no-reference metrics quantify its quality using a prescribed criterion based on some statistical feature information. In the current setting, the realization of \hat{X}^n , the realization of X^n , and the distribution of X^n correspond to the object to be evaluated, the ground truth, and the statistical feature information, respectively. However, the perception constraint in the form of (5) or (23) is imposed on the distribution of \hat{X}^n , not on its realization. This is somewhat unsatisfactory since the perceptual quality should be defined for each individual object (say, an image) rather than for an ensemble only. Furthermore, it is often impossible to deduce a realization-based perceptual quality measure from a distribution-based measure because two different distributions may generate the same realization. So the operational meaning of (5) and (23) is not entirely clear.

To gain a better understanding, let us revisit the toy example in Section I. Suppose we want to evaluate the “perceptual quality” of a given realization of \hat{S} . Clearly, it is irrelevant here whether the distribution on the unit circle is uniform or not as we are dealing with a property of the realization itself. When a particular realization is concerned, the only role of the so-called “perfect perceptual quality constraint” is to force it to be on the unit circle. So the “perception constraint” should be better stated to specify a perceptually admissible set rather than the distribution(s) on it. Interestingly, if no restriction is imposed on how \hat{S} is distributed on the unit circle, then one can simply choose two antipodal reconstruction points (e.g., those highlighted with * in Fig. 3) and perform vector quantization without involving extra randomness. Indeed, deterministic encoding and

decoding

$$K := \begin{cases} 0, & \theta(S) \in [0, \pi), \\ 1, & \theta(S) \in [\pi, 2\pi), \end{cases}$$

$$\hat{S} := \begin{cases} (0, 1), & K = 0, \\ (0, -1), & K = 1, \end{cases}$$

yields

$$\mathbb{E}[\|S - \hat{S}\|^2] = 2 - \frac{4}{\pi},$$

previously only achievable with the aid of common randomness. Note that requiring \hat{S} to reside on the unit circle is not a superfluous ‘‘perception constraint’’ since otherwise it is possible [11] to achieve

$$\mathbb{E}[\|S - \hat{S}\|^2] = 1 - \frac{4}{\pi^2}$$

by using the following modified decoder

$$\hat{S} := \begin{cases} (0, \frac{2}{\pi}), & K = 0, \\ (0, -\frac{2}{\pi}), & K = 1, \end{cases}$$

with the two reconstruction points (highlighted with \circ in Fig. 3) sitting inside the unit circle. Therefore, this distribution-oblivious ‘‘perception constraint’’ also requires a compromise of distortion.

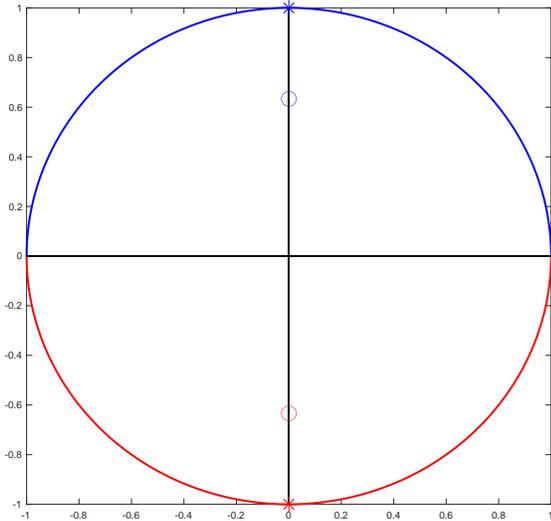


Fig. 3. A toy example. The reconstruction points under the perception constraint are highlighted with $*$ while the unconstrained counterparts are highlighted with \circ .

By now it should be clear that the role of realization-based perception constraints is to specify a collection of perceptually admissible sets parameterized by P (the smaller P is, the more restrictive the set becomes). One

can reconcile such constraints with their distribution-based counterparts using empirical distribution as a link. To this end, we quantify the perceptual quality of a given realization \hat{x}^n by the divergence $d(p_X, \gamma_{\hat{x}^n})$, where $\gamma_{\hat{x}^n}$ is the empirical distribution of \hat{x}^n . We say \hat{x}^n is P -typical with respect to p_X if $d(p_X, \gamma_{\hat{x}^n}) \leq P$, and let the perceptually admissible set be the set of P -typical sequences. It will be seen that the single-letter characterization of the rate-distortion-perception function with distribution-based perception constraints (especially in the form of (5)) can be largely recovered in the realization-based framework under the following definition.

Definition 5. Given distortion constraint D and perception constraint P , rate \bar{R} is said to be achievable with common randomness if for all sufficiently large n , there exist shared seed distribution p_Q (on a Polish space), encoding distribution $p_{Z|X^n Q}$ (with Z countable), and decoding distribution $p_{\hat{X}^n|Z Q}$ (with $\hat{X} = \mathcal{X}$) such that the induced joint distribution $p_{X^n Q Z \hat{X}^n} := p_X^n p_Q p_{Z|X^n Q} p_{\hat{X}^n|Z Q}$ satisfies

$$\frac{1}{n} H(Z|Q) \leq \tilde{R},$$

$$\frac{1}{n} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \hat{X}_t)] \leq D,$$

$$d(p_X, \gamma_{\hat{X}^n}) \leq P \text{ almost surely.} \quad (29)$$

The infimum of such achievable \bar{R} is denoted $\bar{R}_{\text{cr}}(D, P)$. The achievable rate with private randomness is defined in the same way except that the encoder and the decoder do not have access to a shared random seed (i.e., Q is set to be a constant); the corresponding fundamental limit is denoted $\bar{R}_{\text{pr}}(D, P)$. If the encoder and the decoder are further required to be deterministic (i.e., no randomness is allowed at all), we denote the fundamental limit by $\bar{R}_{\text{nr}}(D, P)$.

Theorem 11. Under Assumptions 1 and 2,

$$\bar{R}_{\text{cr}}(D, P) = \bar{R}_{\text{pr}}(D, P) = \bar{R}_{\text{nr}}(D, P) = R(D, P).$$

for $D > 0$ and $P > 0$

Remark 11. It is worth mentioning that $P = 0$ is in general not feasible under Definition 5 since p_X cannot be realized as an empirical distribution for all sufficiently large n . Moreover, $D = 0$ implies $\hat{X}^n = X^n$ almost surely. Note that in the finite alphabet case, $d(p_X, \gamma_{X^n}) \leq P$ almost surely for all sufficiently large n if and only if

$$P \geq \sup_{p_{X'}: p_{X'} \ll p_X} d(p_X, p_{X'}),$$

where $p_{X'} \ll p_X$ means $p_{\hat{X}'}$ is absolutely continuous with respect to p_X .

Proof: First consider the case of common randomness. Let \bar{R} be an achievable rate with respect to distortion constraint D and perception constraint P . Similarly to (24) and (25), we have

$$\bar{R} \geq I(X_T; \hat{X}_T), \quad (30)$$

$$D \geq \mathbb{E}[(\Delta(X_T, \hat{X}_T))], \quad (31)$$

where T is uniformly distributed over $[1 : n]$ and independent of (X^n, \hat{X}^n) . Note that given $\hat{X}^n = \hat{x}^n$, the distribution of \hat{X}_T is the same as the empirical distribution of \hat{x}^n , and consequently

$$d(p_X, \gamma_{\hat{x}^n}) = d(p_X, p_{\hat{X}_T | \hat{X}^n = \hat{x}^n}).$$

Since $d(p_X, \gamma_{\hat{x}^n}) \leq P$ almost surely and $d(\cdot, \cdot)$ is convex in its second argument, it follows that

$$d(p_X, p_{\hat{X}_T}) \leq P. \quad (32)$$

Combining (30), (31), (32) and invoking the fact that $p_{X_T} = p_X$ proves $\bar{R}_{\text{cr}}(D, P) \geq R(D, P)$.

Next consider the case of no randomness. According to Assumption 2, for any $\epsilon \in (0, \min\{\frac{D}{2}, \frac{P}{2}\})$, there exists a discrete random variable \tilde{X} with its support $\tilde{\mathcal{X}}$ satisfying $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ and $|\tilde{\mathcal{X}}| < \infty$ such that

$$I(X; \tilde{X}) \leq R(D - 2\epsilon, P - 2\epsilon) + \epsilon,$$

$$\mathbb{E}[\Delta(X, \tilde{X})] \leq D - \epsilon,$$

$$\mathbb{E} \left[\max_{\tilde{x} \in \tilde{\mathcal{X}}} \Delta(X, \tilde{x}) \right] < \infty,$$

$$d(p_X, p_{\tilde{X}}) \leq P - \epsilon.$$

By Lemma 1 in Appendix B, for any $\delta > 0$, when n is sufficiently large, we can find deterministic encoding function $f^{(n)} : \mathcal{X}^n \rightarrow [1 : M^{(n)}]$ and decoding function $g^{(n)} : [1 : M^{(n)}] \rightarrow \mathcal{C}^{(n)}$ with the properties

$$\frac{1}{n} \log M^{(n)} \leq I(X; \tilde{X}) + 2\delta(H(\tilde{X}) + 1),$$

$$\frac{1}{n} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \hat{X}_t)] \leq \mathbb{E}[\Delta(X, \tilde{X})] + \delta,$$

$$\mathcal{C}^{(n)} \subseteq \mathcal{T}_\delta^{(n)}(p_{\tilde{X}}),$$

where $\tilde{X}^n := g^{(n)}(f^{(n)}(X^n))$ and $\mathcal{T}_\delta^{(n)}(p_{\tilde{X}})$ is the set of δ -typical sequences with respect to $p_{\tilde{X}}$. Clearly, by choosing $\delta \leq \frac{\epsilon}{2(H(\tilde{X})+1)}$, we have

$$\frac{1}{n} \log M^{(n)} \leq R(D - 2\epsilon, P - 2\epsilon) + 2\epsilon,$$

$$\frac{1}{n} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \hat{X}_t)] \leq D.$$

Moreover, since $d(p_X, \gamma)$ is continuous in γ over the interior of the probability simplex defined on $\tilde{\mathcal{X}}$, it follows that

$$d(p_X, \gamma_{\tilde{x}^n}) \leq d(p_X, p_{\tilde{X}}) + \epsilon \leq P$$

for all $\tilde{x}^n \in \mathcal{T}_\delta^{(n)}(p_{\tilde{X}})$ when δ is small enough. Now sending $\epsilon \rightarrow 0$ and invoking the continuity of $R(D, P)$ for $D > 0$ and $P > 0$ shows $\bar{R}_{\text{nr}}(D, P) \leq R(D, P)$. In view of the fact that $\bar{R}_{\text{cr}}(D, P) \leq \bar{R}_{\text{pr}}(D, P) \leq \bar{R}_{\text{nr}}(D, P)$, the proof is complete. ■

Although defining the perceptual quality of a realization based on its empirical distribution is arguably justifiable for i.i.d. sources, this is at best a first-order approximation of what humans subconsciously adopt for image evaluation. How to take into account more sophisticated patterns while maintaining the theory at a manageable level is a challenge to be addressed in future works.

V. CONCLUSION

The roles of common randomness and private randomness in rate-distortion-perception theory have been investigated and shown to depend critically on how the perception constraint is formulated. The operational meanings of various perception constraints are also examined. It is our opinion that, as the distinguishing feature of the new theory and likely the decisive factor for its success, the notion of perceptual quality remains to be formalized more convincingly in an information-theoretic framework as a no-reference metric closely mimicking the relevant human intuition.

APPENDIX A

VERIFICATION OF ASSUMPTION 2

Here we verify Assumption 2 for the case $\mathbb{E}[X^2] < \infty$, $\Delta(x, \hat{x}) := (x - \hat{x})^2$, and $d(p_X, p_{\hat{X}}) := \inf_{p_{X\hat{X}} \in \Pi(p_X, p_{\hat{X}})} \mathbb{E}[(X - \hat{X})^2]$ (in this case, $d(p_X, p_{\hat{X}})$ is written as $W_2^2(p_X, p_{\hat{X}})$ according to the convention).

Let \tilde{X} be a quantized version of X , obtained by mapping X to its nearest point in $\frac{1}{\sqrt{N}}[-N : N]$, where N is a positive integer. Since $\mathbb{E}[X^2] < \infty$, for any $D > 0$, we can choose a sufficiently large N such that $\mathbb{E}[(X - \tilde{X})^2] \leq \frac{D}{2}$. Let \hat{X} be the mirror version of X with respect to $\tilde{X} := \mathbb{E}[X|\tilde{X}]$ in the sense that $p_{X\tilde{X}\hat{X}} = p_{X\tilde{X}}p_{\hat{X}|\tilde{X}}$ and $p_{X|\hat{X}} = p_{\hat{X}|\tilde{X}}$. Obviously, $X \leftrightarrow \tilde{X} \leftrightarrow \hat{X}$ form a Markov chain and $p_{\hat{X}} = p_X$. Note that

$$\begin{aligned} I(X; \hat{X}) &\leq I(X; \tilde{X}) \leq H(\tilde{X}) \leq H(\bar{X}) \leq \log(2N + 1), \\ \mathbb{E}[(X - \hat{X})^2] &= 2\mathbb{E}[(X - \tilde{X})^2] \leq 2\mathbb{E}[(X - \bar{X})^2] \leq D, \\ W_2^2(p_X, p_{\hat{X}}) &= 0. \end{aligned}$$

Therefore, we have⁴ $R(D, 0) < \infty$. Since $R(D, P) \leq R(D, 0)$ for $P > 0$, it follows that (7) holds.

By the definition of $R(D, P)$, for any $D > 0$ and $P > 0$, there exists a random variable \hat{X} such that

$$\begin{aligned} I(X; \hat{X}) &\leq R(D, P) + \epsilon, \\ \mathbb{E}[(X - \hat{X})^2] &\leq D, \\ W_2^2(p_X, p_{\hat{X}}) &\leq P. \end{aligned}$$

Let \tilde{X} be a quantized version of \hat{X} , obtained by mapping X to its nearest point in $\frac{1}{\sqrt{N}}[-N : N]$, where N is a positive integer. Clearly, we have $|\tilde{\mathcal{X}}| < \infty$ since $\tilde{\mathcal{X}} \subseteq \frac{1}{\sqrt{N}}[-N : N]$. With this construction, (8) is a simple consequence of the data processing inequality. It can be verified that

$$\begin{aligned} \mathbb{E}[(X - \tilde{X})^2] & \quad (33) \\ &= \mathbb{E}[(X - \hat{X} + \hat{X} - \tilde{X})^2] \\ &= \mathbb{E}[(X - \hat{X})^2] + \mathbb{E}[(\hat{X} - \tilde{X})^2] \\ &\quad + 2\mathbb{E}[(X - \hat{X})(\hat{X} - \tilde{X})] \\ &\leq \mathbb{E}[(X - \hat{X})^2] + \mathbb{E}[(\hat{X} - \tilde{X})^2] \\ &\quad + 2\sqrt{\mathbb{E}[(X - \hat{X})^2]\mathbb{E}[(\hat{X} - \tilde{X})^2]} \quad (34) \\ &\leq D + \mathbb{E}[(\hat{X} - \tilde{X})^2] + 2\sqrt{D\mathbb{E}[(\hat{X} - \tilde{X})^2]}, \end{aligned}$$

where (34) is due to the Cauchy-Schwarz inequality. Similarly,

$$W_2^2(p_X, p_{\tilde{X}}) \leq P + \mathbb{E}[(\hat{X} - \tilde{X})^2] + 2\sqrt{P\mathbb{E}[(\hat{X} - \tilde{X})^2]}.$$

Since $\mathbb{E}[\hat{X}^2] < \infty$ (implied by the fact $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[(X - \hat{X})^2] < \infty$), it follows that $\mathbb{E}[(\hat{X} - \tilde{X})^2] \rightarrow 0$ as $N \rightarrow \infty$. Therefore, we can ensure (9) and (11) by setting N large enough. Moreover, (10) is satisfied because

$$\begin{aligned} \mathbb{E} \left[\max_{\tilde{x} \in \tilde{\mathcal{X}}} (X - \tilde{x})^2 \right] &\leq \mathbb{E}[X^2] + 2\sqrt{N}\mathbb{E}[|X|] + N \\ &\leq \mathbb{E}[X^2] + 2\sqrt{N\mathbb{E}[X^2]} + N \quad (35) \\ &< \infty, \end{aligned}$$

where (35) is due to the Cauchy-Schwarz inequality. In view of the fact that $W_2^2(p_X, \gamma) \leq \mathbb{E}[\max_{\tilde{x} \in \tilde{\mathcal{X}}} (X - \tilde{x})^2]$ for all γ with support $\tilde{\mathcal{X}}$, (12) must hold as well. This completes the verification of Assumption 2.

⁴It is worth mentioning that the value of $R(D, 0)$ does not depend on the choice of divergence d .

APPENDIX B PROOF OF THEOREM 2

Lemma 1. *Let \tilde{X} be a discrete random variable defined on \mathcal{X} and jointly distributed with X and assume that its support $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ satisfies $|\tilde{\mathcal{X}}| < \infty$ and*

$$\mathbb{E} \left[\max_{\tilde{x} \in \tilde{\mathcal{X}}} \Delta(X, \tilde{x}) \right] < \infty.$$

Given any $D > 0$, $P > 0$, and $\delta > 0$, there exist deterministic encoding function $f^{(n)} : \mathcal{X}^n \rightarrow [1 : M^{(n)}]$ and decoding function $g^{(n)} : [1 : M^{(n)}] \rightarrow \mathcal{C}^{(n)}$ for all sufficiently large n such that

$$\begin{aligned} \frac{1}{n} \log M^{(n)} &\leq I(X; \tilde{X}) + 2\delta(H(\tilde{X}) + 1), \\ \frac{1}{n} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \tilde{X}_t)] &\leq \mathbb{E}[\Delta(X, \tilde{X})] + \delta, \\ \mathcal{C}^{(n)} &\subseteq \mathcal{T}_\delta^{(n)}(p_{\tilde{X}}), \end{aligned}$$

where $\tilde{X}^n := g^{(n)}(f^{(n)}(X^n))$ and $\mathcal{T}_\delta^{(n)}(p_{\tilde{X}})$ denotes the set of δ -typical sequences with respect to $p_{\tilde{X}}$, i.e., $\mathcal{T}_\delta^{(n)}(p_{\tilde{X}}) := \{\tilde{x}^n \in \tilde{\mathcal{X}}^n : |\gamma_{\tilde{x}^n}(\tilde{x}) - p_{\tilde{X}}(\tilde{x})| \leq \delta p_{\tilde{X}}(\tilde{x}) \text{ for all } \tilde{x} \in \tilde{\mathcal{X}}\}$.

Proof: This result has many known variants in the literature (see, e.g., [20, Theorem 9.6.2]). We include its proof for completeness.

We independently and uniformly choose $M^{(n)} := \lfloor 2^{n(I(X; \tilde{X}) + 2\delta')} \rfloor$ codewords $\mathcal{C}^{(n)} := \{\tilde{X}^n\}_{m=1}^{M^{(n)}}$ from $\mathcal{T}_\delta^{(n)}(p_{\tilde{X}})$. Given $X^n = x^n$, the encoder finds an \hat{m} such that

$$\frac{1}{n} \sum_{t=1}^n \Delta(x_t, \tilde{X}_t(\hat{m})) \leq \mathbb{E}[\Delta(X, \tilde{X})] + \frac{\delta}{2}. \quad (36)$$

If no such \hat{m} exists, the encoder simply sets $\hat{m} = 1$. Let

$$\begin{aligned} \mathcal{A}_{x^n} &:= \left\{ \tilde{x}^n \in \mathcal{T}_\delta^{(n)}(p_{\tilde{X}}) : \prod_{t=1}^n p_{\tilde{X}|X}(\tilde{x}_t|x_t) > \frac{2^{n\hat{R}}}{|\mathcal{T}_\delta^{(n)}(p_{\tilde{X}})|} \right. \\ &\quad \left. \text{or } \frac{1}{n} \sum_{t=1}^n \Delta(x_t, \tilde{x}_t) > \mathbb{E}[\Delta(X, \tilde{X})] + \frac{\delta}{2} \right\}, \end{aligned}$$

where $\hat{R} := I(X; \tilde{X}) + \delta'$. Given $X^n = x^n$, the encoder fails to find a codeword satisfying (36) only if $\tilde{X}^n(m) \in$

\mathcal{A}_{x^n} for all m . Therefore,

$$\begin{aligned} & \mathbb{P}\left(\frac{1}{n}\sum_{t=1}^n \Delta(x_t, \tilde{X}^n(\hat{m}))\right) \\ & > \mathbb{E}[\Delta(X, \tilde{X})] + \frac{\delta}{2} \text{ for all } m \mid X^n = x^n \\ & \leq \left(1 - \sum_{\tilde{x}^n \in \mathcal{T}_\delta^{(n)}(p_{\tilde{X}}) \setminus \mathcal{A}_{x^n}} \frac{1}{|\mathcal{T}_\delta^{(n)}(p_{\tilde{X}})|}\right)^{M^{(n)}} \\ & \leq \left(1 - 2^{-n\hat{R}} \sum_{\tilde{x}^n \in \mathcal{T}_\delta^{(n)}(p_{\tilde{X}}) \setminus \mathcal{A}_{x^n}} \prod_{t=1}^n p_{\tilde{X}|X}(\tilde{x}_t|x_t)\right)^{M^{(n)}} \\ & \leq 1 - \sum_{\tilde{x}^n \in \mathcal{T}_\delta^{(n)}(p_{\tilde{X}}) \setminus \mathcal{A}_{x^n}} \prod_{t=1}^n p_{\tilde{X}|X}(\tilde{x}_t|x_t) \\ & \quad + \exp(-M^{(n)}2^{-n\hat{R}}) \end{aligned} \quad (37)$$

$$\begin{aligned} & = \sum_{\tilde{x}^n \in (\mathcal{T}_\delta^{(n)}(p_{\tilde{X}})^c \cup \mathcal{A}_{x^n})} \prod_{t=1}^n p_{\tilde{X}|X}(\tilde{x}_t|x_t) \\ & \quad + \exp(-M^{(n)}2^{-n\hat{R}}), \end{aligned} \quad (38)$$

where (37) is due to [15, Lemma 13.5.3]. Let $\mathcal{A}^{(n)} := \{(x^n, \tilde{x}^n) : x^n \in \mathcal{X}^n, \tilde{x}^n \in (\mathcal{T}_\delta^{(n)}(p_{\tilde{X}})^c \cup \mathcal{A}_{x^n})\}$. It follows from (38) that

$$\begin{aligned} & \mathbb{P}\left(\frac{1}{n}\sum_{t=1}^n \Delta(X_t, \tilde{X}(\hat{m})) > \mathbb{E}[\Delta(X, \tilde{X})] + \frac{\delta}{2}\right) \\ & \leq \mathbb{P}((X^n, \tilde{X}^n) \in \mathcal{A}^{(n)}) + \exp(-M^{(n)}2^{-n\hat{R}}), \end{aligned}$$

where (X_t, \tilde{X}_t) , $t \in [1 : n]$, are independent and distributed according to $p_{X, \tilde{X}}$. Note that $\tilde{X}^n \in \mathcal{T}_\delta^{(n)}(p_{\tilde{X}})$ with high probability when n is large. Moreover, in light of the weak law of large numbers, $\frac{1}{n}\sum_{t=1}^n p_{\tilde{X}|X}(\tilde{X}_t|X_t)$ and $\frac{1}{n}\sum_{t=1}^n \Delta(X_t, \tilde{X}_t)$ converge, respectively, to $-H(\tilde{X}|X)$ and $\mathbb{E}[\Delta(X, \tilde{X})]$ in probability. It is easy to see that

$$\begin{aligned} \frac{1}{n} \log \frac{2^{n\hat{R}}}{|\mathcal{T}_\delta^{(n)}(p_{\tilde{X}})|} & \geq \frac{1}{n} \log \frac{2^{n\hat{R}}}{2^{n(1+\delta)H(\tilde{X})}} \\ & = I(X; \tilde{X}) + \delta' - (1+\delta)H(\tilde{X}) \\ & = -H(\tilde{X}|X) + \delta' - \delta H(\tilde{X}). \end{aligned}$$

So by choosing $\delta' = \delta(H(\tilde{X}) + 1)$, we have $\mathbb{P}((X^n, \tilde{X}^n) \in \mathcal{A}^{(n)}) \rightarrow 0$ as $n \rightarrow \infty$. It is also clear that $\exp(-M^{(n)}2^{-n\hat{R}}) \rightarrow 0$ as $n \rightarrow \infty$.

As shown by the above argument, for any $\delta > 0$ and $\epsilon > 0$, when n is sufficiently large, we can find

deterministic encoding function $f^{(n)} : \mathcal{X}^n \rightarrow [1 : M^{(n)})$ and decoding function $g^{(n)} : [1 : M^{(n)}) \rightarrow \mathcal{C}^{(n)}$ such that

$$\begin{aligned} & \frac{1}{n} \log M^{(n)} \leq I(X; \tilde{X}) + 2\delta(H(\tilde{X}) + 1), \\ & \mathbb{P}\left(\frac{1}{n}\sum_{t=1}^n \Delta(X_t, \hat{X}_t) > \mathbb{E}[\Delta(X, \tilde{X})] + \frac{\delta}{2}\right) \leq \epsilon, \\ & \mathcal{C}^{(n)} \subseteq \mathcal{T}_\delta^{(n)}(p_{\tilde{X}}). \end{aligned}$$

Let

$$W := \begin{cases} 1, & \frac{1}{n}\sum_{t=1}^n \Delta(X_t, \hat{X}_t) > \mathbb{E}[\Delta(X, \tilde{X})] + \frac{\delta}{2}, \\ 0, & \text{otherwise,} \end{cases}$$

and $V_t := \max_{\tilde{x} \in \tilde{\mathcal{X}}} \Delta(X_t, \tilde{x})$, $t \in [1 : n]$. We have

$$\begin{aligned} & \frac{1}{n} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \hat{X}_t)] \\ & = \frac{1}{n} \sum_{t=1}^n \mathbb{P}(W = 0) \mathbb{E}[\Delta(X_t, \hat{X}_t) | W = 0] \\ & \quad + \frac{1}{n} \sum_{t=1}^n \mathbb{P}(W = 1) \mathbb{E}[\Delta(X_t, \hat{X}_t) | W = 1] \\ & \leq \frac{1}{n} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \hat{X}_t) | W = 0] \\ & \quad + \frac{1}{n} \sum_{t=1}^n \mathbb{E}[W \Delta(X_t, \hat{X}_t)] \\ & \leq \mathbb{E}[\Delta(X, \tilde{X})] + \frac{\delta}{2} + \frac{1}{n} \sum_{t=1}^n \mathbb{E}[W V_t]. \end{aligned}$$

Note that

$$\begin{aligned} \mathbb{E}[W V_t] & = \mathbb{P}\left(V_t \leq \frac{1}{\sqrt{\epsilon}}\right) \mathbb{E}\left[W V_t \mid V_t \leq \frac{1}{\sqrt{\epsilon}}\right] \\ & \quad + \mathbb{P}\left(V_t > \frac{1}{\sqrt{\epsilon}}\right) \mathbb{E}\left[W V_t \mid V_t > \frac{1}{\sqrt{\epsilon}}\right] \\ & \leq \mathbb{P}\left(V_t \leq \frac{1}{\sqrt{\epsilon}}\right) \mathbb{E}\left[W \frac{1}{\sqrt{\epsilon}} \mid V_t \leq \frac{1}{\sqrt{\epsilon}}\right] \\ & \quad + \mathbb{P}\left(V_t > \frac{1}{\sqrt{\epsilon}}\right) \mathbb{E}\left[V_t \mid V_t > \frac{1}{\sqrt{\epsilon}}\right] \\ & \leq \frac{1}{\sqrt{\epsilon}} \mathbb{E}[W] + \mathbb{P}\left(V_t > \frac{1}{\sqrt{\epsilon}}\right) \mathbb{E}\left[V_t \mid V_t > \frac{1}{\sqrt{\epsilon}}\right] \\ & \leq \sqrt{\epsilon} + \mathbb{P}\left(V_t > \frac{1}{\sqrt{\epsilon}}\right) \mathbb{E}\left[V_t \mid V_t > \frac{1}{\sqrt{\epsilon}}\right]. \end{aligned}$$

Since

$$\mathbb{E}[V_t] = \mathbb{E}\left[\max_{\tilde{x} \in \tilde{\mathcal{X}}} \Delta(X, \tilde{x})\right] < \infty,$$

it follows by the dominated convergence theorem that $\mathbb{P}(V_t > v) \mathbb{E}[V_t | V_t > v] \rightarrow 0$ as $v \rightarrow \infty$. Therefore,

by choosing a sufficiently small ϵ , we can ensure that $\mathbb{E}[WV_t] \leq \frac{\delta}{2}$, $t \in [1 : n]$, and consequently

$$\frac{1}{n} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \hat{X}_t)] \leq \mathbb{E}[\Delta(X, \tilde{X})] + \delta.$$

This completes the proof of Lemma 1. \blacksquare

Now we proceed to prove Theorem 2. It suffices to consider the case where p_X does not assign all probability mass to a single atom since otherwise the problem is trivial.

According to Assumption 2, for any $\epsilon \in (0, \min\{\frac{D}{2}, \frac{P}{2}\})$, there exists a random variable \tilde{X} with its support $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ and $|\tilde{\mathcal{X}}| < \infty$ such that

$$\begin{aligned} I(X; \tilde{X}) &\leq R(D - 2\epsilon, P - 2\epsilon) + \epsilon, \\ \mathbb{E}[\Delta(X, \tilde{X})] &\leq D - \epsilon, \\ \mathbb{E} \left[\max_{\tilde{x} \in \tilde{\mathcal{X}}} \Delta(X, \tilde{x}) \right] &< \infty, \\ d(p_X, p_{\tilde{X}}) &\leq P - \epsilon. \end{aligned} \quad (39)$$

By Lemma 1, for any $\delta > 0$, when n is sufficiently large, we can find deterministic encoding function $f^{(n)} : \mathcal{X}^n \rightarrow [1 : M^{(n)}]$ and decoding function $g^{(n)} : [1 : M^{(n)}] \rightarrow \mathcal{C}^{(n)}$ with the properties

$$\begin{aligned} \frac{1}{n} \log M^{(n)} &\leq I(X; \tilde{X}) + 2\delta(H(\tilde{X}) + 1), \\ \frac{1}{n} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \tilde{X}_t)] &\leq \mathbb{E}[\Delta(X, \tilde{X})] + \delta, \\ \mathcal{C}^{(n)} &\subseteq \mathcal{T}_\delta^{(n)}(p_{\tilde{X}}). \end{aligned}$$

Clearly, by choosing $\delta \leq \frac{\epsilon}{2(H(\tilde{X})+1)}$, we have

$$\begin{aligned} \frac{1}{n} \log M^{(n)} &\leq R(D - 2\epsilon, P - 2\epsilon) + 2\epsilon, \\ \frac{1}{n} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \tilde{X}_t)] &\leq D - \frac{\epsilon}{2}. \end{aligned}$$

Moreover, $\frac{1}{n} \sum_{t=1}^n p_{\tilde{X}_t}$ must converge to $p_{\tilde{X}}$ under the total variation distance (i.e., $d_{\text{TV}}(\frac{1}{n} \sum_{t=1}^n p_{\tilde{X}_t}, p_{\tilde{X}}) \rightarrow 0$) as $\delta \rightarrow 0$ given the fact that

$$\begin{aligned} \frac{1}{n} \sum_{t=1}^n p_{\tilde{X}_t | \tilde{X}^n = \tilde{x}^n} &= \gamma_{\tilde{x}^n}, \\ \mathbb{P}(\tilde{X}^n \in \mathcal{T}_\delta^{(n)}(p_{\tilde{X}})) &= 1. \end{aligned}$$

Since $d(p_X, \gamma)$ is continuous in γ over the interior of the probability simplex defined on $\tilde{\mathcal{X}}$, it follows that

$$d\left(p_X, \frac{1}{n} \sum_{t=1}^n p_{\tilde{X}_t}\right) \leq d(p_X, p_{\tilde{X}}) + \frac{\epsilon}{2} \leq P - \frac{\epsilon}{2} \quad (40)$$

when δ is sufficiently close to zero.

However, in general (40) does not imply the stronger requirement

$$d(p_X, p_{\tilde{X}_t}) \leq P - \frac{\epsilon}{2}, \quad t \in [1 : n].$$

Nevertheless, there is a simple remedy with the availability of common randomness. For any integer q , let $s_q^{(n)}$ be a circular shift operator in the sense that $s_q^{(n)}(x^n) = (x_{1 \oplus_n q}, x_{2 \oplus_n q}, \dots, x_{n \oplus_n q})$ for all x^n , where \oplus_n is modulo- n addition⁵. Let $\mathcal{C}_q^{(n)}$ denote the codebook obtained by applying $s_q^{(n)}$ to every codeword⁶ of $\mathcal{C}^{(n)}$, $q \in [0 : n - 1]$. Moreover, we equip each codebook $\mathcal{C}_q^{(n)}$ with encoding function $f_q^{(n)} : \mathcal{X}^n \rightarrow [1 : M^{(n)}]$ and decoding function $g_q^{(n)} : [1 : M^{(n)}] \rightarrow \mathcal{C}_q^{(n)}$ defined as follows:

$$\begin{aligned} f_q^{(n)}(x^n) &:= f^{(n)}(s_{-q}^{(n)}(x^n)), \quad x^n \in \mathcal{X}^n, \\ g_q^{(n)}(m) &= s_q^{(n)}(g^{(n)}(m)), \quad m \in [1 : M^{(n)}]. \end{aligned}$$

Note that we have $\mathcal{C}_0^{(n)} = \mathcal{C}^{(n)}$, $f_0^{(n)} = f^{(n)}$, and $g_0^{(n)} = g^{(n)}$. Let Q be uniformly distributed over $[0 : n - 1]$, independent of X^n , and available at both the encoder and decoder. The role of Q is to specify which codebook (as well as the associated encoding and decoding functions) to use. Let \tilde{X}^n denote the reconstruction. Based on our construction, it is clear that

$$\begin{aligned} p_{\tilde{X}_t} &= \frac{1}{n} \sum_{t'=1}^n p_{\tilde{X}_{t'}}, \quad t \in [1 : n], \\ \frac{1}{n} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \tilde{X}_t) | Q = q] &= \frac{1}{n} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \tilde{X}_t)], \\ & \quad q \in [0 : n - 1], \end{aligned}$$

and consequently

$$\begin{aligned} d(p_X, p_{\tilde{X}_t}) &\leq P - \frac{\epsilon}{2}, \quad t \in [1 : n], \\ \frac{1}{n} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \tilde{X}_t)] &\leq D - \frac{\epsilon}{2}. \end{aligned}$$

It remains to eliminate common randomness. The key observation here is that Q can be simulated using a negligible fraction of source symbols as $H(Q)$ is sublinear in n . Specifically, consider the case where $n_0 := \lfloor n\alpha \rfloor$ source symbols are leveraged to simulate Q , where $\alpha > 0$ is small enough. Let p_{\max} denote the maximum probability value of p_X (i.e., the maximum of the probabilities of atoms if atoms exist; otherwise, $p_{\max} = 0$). We assume $p_{\max} > 0$ since otherwise Q

⁵We assume the output of the modulo- n operation is in $[1 : n]$.

⁶The resulting codeword still retains the index of the original codeword.

can be exactly generated from $X \sim p_X$. Obviously, the maximum probability value of $p_X^{n_0}$ is $p_{\max}^{n_0}$, which vanishes exponentially as $n \rightarrow \infty$ (for any given $\alpha > 0$). This ensures that there is a map $\omega : \mathcal{X}^{n_0} \rightarrow [0 : n - 1]$ such that the distribution $p_{\omega(X^{n_0})}$ of $\omega(X^{n_0})$ satisfies $|p_{\omega(X^{n_0})}(i) - \frac{1}{n}| \leq p_{\max}^{n_0}$ for all $i \in [0 : n - 1]$. For example, first map as many atoms as possible to the elements of $[0 : n - 1]$ while ensuring that the total mass of atoms mapped to each element is no greater than $\frac{1}{n}$; then use the remaining atoms and atomless measurable sets to fill the gap between the total mass for each element and the target $\frac{1}{n}$ so that the total mass for each element is within the range $\frac{1}{n} \pm p_{\max}^{n_0}$. Consequently, we have $d_{\text{TV}}(p_{\omega(X^{n_0})}, \text{Unif}[0 : n - 1]) \leq n p_{\max}^{n_0} \rightarrow 0$ as $n \rightarrow \infty$.

We assume the encoder uses this simulation code on the $(n + 1)$ -th to $(n + n_0)$ -th source symbols $X_{n+1}^{n+n_0}$ to generate $\hat{Q} := \omega(X_{n+1}^{n+n_0})$. It then transmits \hat{Q} to the decoder losslessly with $\lceil \log n \rceil$ bits and applies the aforementioned randomly shifted code to the first n source symbols X^n but with Q replaced by its approximate version \hat{Q} . Denote the reconstructions of the first n source symbols by \hat{X}^n . For the $(n + 1)$ -th to $(n + n_0)$ -th source symbols, the decoder generates reconstructions $\hat{X}_{n+j} := \hat{X}_j$, $j \in [1 : n_0]$. Under this construction, both the encoder and decoder are deterministic.

We shall show that the deterministic code has the desired properties. Since

$$\begin{aligned} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \hat{X}_t) | Q = q] &= \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \tilde{X}_t)], \\ & \quad q \in [0 : n - 1], \\ \mathbb{E}[\Delta(X_{n+j}, \hat{X}_{n+j})] &\leq \mathbb{E} \left[\max_{\tilde{x} \in \tilde{\mathcal{X}}} \Delta(X, \tilde{x}) \right] < \infty, \\ & \quad j \in [1 : n_0], \end{aligned}$$

it follows that

$$\frac{1}{n + n_0} \sum_{t=1}^{n+n_0} \mathbb{E}[\Delta(X_t, \hat{X}_t)] \leq D$$

for all sufficiently large n as long as α is chosen small enough. Moreover, in view of the fact that $p_{\hat{X}_t | \hat{Q}} = p_{\tilde{X}_t | Q}$, $t \in [1 : n]$, and $d_{\text{TV}}(p_{\hat{Q}}, p_Q) \rightarrow 0$ as $n \rightarrow \infty$, we must have $d_{\text{TV}}(p_{\hat{X}_t}, p_{\tilde{X}_t}) \rightarrow 0$ uniformly for all $t \in [1 : n]$ as $n \rightarrow \infty$, which, together with the uniform convergence of $p_{\hat{X}_t}$, $t \in [1 : n]$, to $p_{\tilde{X}}$ under the total variation distance as $\delta \rightarrow 0$ and the continuity of $d(p_X, q)$ in this second argument at $q = p_{\tilde{X}}$, implies

$$d(p_X, p_{\tilde{X}_t}) \leq P, \quad t \in [1 : n + n_0],$$

for all sufficiently large n as long as δ is chosen small enough. Finally, invoking the fact that $\frac{\lceil \log n \rceil}{n+n_0} \rightarrow 0$ as

$n \rightarrow \infty$ and the continuity of $R(D, P)$ for $D > 0$ and $P > 0$ completes the proof of Theorem 2.

Remark 12. *It is possible to avoid using the simulation code via a more careful construction.*

Let $\mathcal{A} := \{x^n \in \mathcal{X}^n : x^n = s_i^{(n)}(x^n) \text{ for some } i \in [1 : n - 1]\}$ and $\mathcal{B} := \mathcal{X}^n \setminus \mathcal{A}$. Moreover, we partition \mathcal{B} into $\mathcal{B}_0, \mathcal{B}_1, \dots, \mathcal{B}_{n-1}$ such that $\mathcal{B}_q = \{s_i^{(n)}(x^n) : x^n \in \mathcal{B}_0\}$, $q \in [1 : n - 1]$. (The existence of such a partition will be discussed later.)

Take codebook $\mathcal{C}^n \subseteq \mathcal{T}_\delta^{(n)}(p_{\tilde{X}})$ and its associated deterministic encoding function $f^{(n)} : \mathcal{X}^n \rightarrow [1 : M^{(n)}]$ and decoding function $g^{(n)} : [1 : M^{(n)}] \rightarrow \mathcal{C}^{(n)}$ as specified in the proof of Theorem 2. Given $x^n \in \mathcal{A} \cup \mathcal{B}_0$, the encoder finds⁷ (m^*, q^*) such that

$$(m^*, q^*) = \arg \min_{m \in [1 : M^{(n)}], i \in [0 : n - 1]} \sum_{t=1}^n \Delta(x_t, \tilde{x}_t),$$

where $\tilde{x}^n := s_q^{(n)}(g^{(n)}(m))$. It then sends (m^*, q^*) to the decoder, which produces $s_{i^*}^{(n)}(g^{(n)}(m^*))$ as the reconstruction. For $x^n \in \mathcal{B}_q$, $q \in [1 : n - 1]$, the encoder finds the index pair for the corresponding $s_{-q}^{(n)}(x^n)$ in \mathcal{B}_0 according to the aforesaid encoding rule and sends it to the decoder. The decoder first produces the reconstruction for $s_{-q}^{(n)}(x^n)$ according to the aforesaid decoding rule, then outputs the final result by shifting this reconstruction using $s_q^{(n)}$. Let \tilde{X}^n denote the output induced by this new coding scheme. Since the distortion does not increase for any realization x^n as compared to the original scheme⁸, it follows that

$$\sum_{t=1}^n \mathbb{E}[\Delta(X_t, \tilde{X}_t)] \leq \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \tilde{X}_t)], \quad (41)$$

where $\tilde{X}^n := g^{(n)}(f^{(n)}(X^n))$. Moreover, as shown at the end of Appendix B,

$$\lim_{n \rightarrow \infty} \mathbb{P}(X^n \in \mathcal{A}) = 0. \quad (42)$$

So with δ chosen small enough, the circular symmetry of the coding scheme (conditioned on $X^n \in \mathcal{B}$) ensures that $p_{\tilde{X}_t}$, $t \in [1 : n]$, are uniformly close to $p_{\tilde{X}}$ under the total variation distance and consequently

$$d(p_X, p_{\tilde{X}_t}) \leq P, \quad t \in [1 : n],$$

for all sufficiently large n . As the rate overhead for transmitting q^* is negligible, the new coding scheme indeed has the desired properties.

⁷Use a prescribed deterministic tie-break rule if the minimizer is not unique.

⁸If we modify the coding scheme by setting $(m^*, q^*) = (f^{(n)}(x^n), 0)$ for $x^n \in \mathcal{B}_0$, then the distortion might increase for $x^n \in \mathcal{B} \setminus \mathcal{B}_0$.

However, there is a subtle issue regarding the partition of \mathcal{B} into $\mathcal{B}_0, \mathcal{B}_1, \dots, \mathcal{B}_{n-1}$. The existence of such a partition is in a certain sense guaranteed. Note that \mathcal{B} is the union of a collection of disjoint equivalent classes, where each equivalent class consists of n different sequences that can be converted from one to another via circular shifting. We can form \mathcal{B}_0 by choosing one sequence from each equivalent class in \mathcal{B} ; then \mathcal{B}_i is uniquely specified due to the requirement $\mathcal{B}_q = \{s_q^{(n)}(x^n) : x^n \in \mathcal{B}_0\}$, $q \in [1 : n-1]$. It can be seen that there is considerable freedom in creating this kind of partitions. When \mathcal{X} is finite or countably infinite, the resulting sets are always measurable. But in a more general setting, certain partitions might yield non-measurable sets.

Here we show that this issue can be resolved by performing the partition judiciously when \mathcal{X} is a Polish space (as assumed throughout this paper). First consider the case $\mathcal{X} = \mathbb{R}$. We shall start with a simple example where $n = 3$. There are totally 6 permutations on $[1 : 3]$, namely, $(1, 2, 3)$, $(1, 3, 2)$, $(2, 1, 3)$, $(2, 3, 1)$, $(3, 1, 2)$, and $(3, 2, 1)$. Pick two permutations that are not related via circular shifting⁹. Each permutation (a, b, c) can be used to specify a region in \mathbb{R}^3 according to the following rule: $(a, b, c) \mapsto \{x^3 \in \mathbb{R}^3 : x_a \geq x_b \geq x_c\}$. Now we can let \mathcal{B}_0 be the union of the regions specified by the two picked permutations with the elements in \mathcal{A} excluded. Then \mathcal{B}_1 and \mathcal{B}_2 are also uniquely determined. In general, there are $n!$ permutations on $[1 : n]$. We can pick $(n-1)!$ permutations that are not related via circular shifting, and use these permutations to specify $(n-1)!$ regions in \mathbb{R}^n (according to the obvious extension of the aforementioned rule). Then define \mathcal{B}_0 by taking the union of these $(n-1)!$ regions and excluding the elements in \mathcal{A} . It can be verified that the induced partition $\mathcal{B}_0, \mathcal{B}_1, \dots, \mathcal{B}_{n-1}$ is free of the measurability issue.

This method can be generalized to any Polish space. Note that any Borel space on a Polish space is Borel isomorphic to a measurable subspace of \mathbb{R} . We can simply map the Polish space \mathcal{X} to a measurable subspace \mathcal{Y} of \mathbb{R} via an isomorphism f , which naturally induces a map from \mathcal{X}^n to \mathcal{Y}^n via the isomorphism $f^{(n)} := (f, f, \dots, f)$. We then perform the partition $\{\mathcal{B}_i\}$ for the high probability subset \mathcal{A}^c of \mathbb{R}^n as above, and take intersections $\{\mathcal{B}_i \cap \mathcal{Y}^n\}$ to produce a partition for the high probability subset of \mathcal{Y}^n . Mapping this back to the Polish space \mathcal{X}^n , we obtain a desired partition.

Proof of (42): Note that given $q \in [1 : n-1]$, we

⁹For example, $(1, 2, 3)$ and $(1, 3, 2)$ are not related via circular shifting while $(1, 2, 3)$ and $(3, 1, 2)$ are since applying $s_2^{(3)}$ to $(1, 2, 3)$ gives $(3, 1, 2)$.

can divide $[1 : n]$ into $k := \gcd(n, q)$ subsets of size $\frac{n}{k}$, and the elements in each subset differ by a multiple of q (modulo- n). For example, when $n = 6$ and $q = 4$, we have two subsets $\{1, 3, 5\}$ and $\{2, 4, 6\}$. In this case, $X^n = s_q^{(n)}(X^n)$ means $X_1 = X_3 = X_5$ and $X_2 = X_4 = X_6$; as a consequence

$$\begin{aligned} \mathbb{P}(X^n = s_q^{(n)}(X^n)) &\leq \mathbb{P}(X_1 = X_3 \text{ and } X_2 = X_4) \\ &= \tau^2, \end{aligned}$$

where $\tau := \mathbb{P}(X_i = X_j)$ for $i \neq j$. In general, it can be verified that $X^n = s_q^{(n)}(X^n)$ implies at least $\lfloor \frac{n}{2k} \rfloor k$ independent events of the kind $X_i \neq X_j$, which, together with the fact $\lfloor \frac{n}{2k} \rfloor k \geq \frac{n}{4}$, further implies

$$\mathbb{P}(X^n = s_q^{(n)}(X^n)) \leq \tau^{\frac{n}{4}}.$$

Now one can readily prove (42) since

$$\begin{aligned} \mathbb{P}(X^n \in \mathcal{A}) &\leq \sum_{q=1}^{n-1} \mathbb{P}(X^n = s_q^{(n)}(X^n)) \\ &\leq (n-1)\tau^{\frac{n}{4}} \\ &\rightarrow 0 \text{ as } n \rightarrow \infty, \end{aligned}$$

where the last step is due to $\tau \in [0, 1)$.

APPENDIX C VERIFICATION OF ASSUMPTION 3

Here we verify Assumption 3 for the case $\mathbb{E}[X^2] < \infty$ and $\Delta(x, \hat{x}) := (x - \hat{x})^2$.

According to the definition of $R(D, 0)$, for any $D > 0$ and $\delta \in (0, D)$, there exists a random variable \check{X} such that

$$\begin{aligned} I(X; \check{X}) &\leq R(D - \delta, 0) + \delta, \\ \mathbb{E}[(X - \check{X})^2] &\leq D - \delta, \\ p_{\check{X}} &= p_X. \end{aligned}$$

Let \tilde{X} be a quantized version of \check{X} , obtained by mapping \check{X} to its nearest point in $\frac{1}{\sqrt{N}}[-N : N]$, where N is a positive integer. By the data processing inequality,

$$I(X; \tilde{X}) \leq I(X; \check{X}) \leq R(D - \delta, 0) + \delta.$$

As $R(D, 0)$ is convex in D and consequently¹⁰ continuous for $D > 0$, we have $R(D - \delta, 0) + \delta \leq R(D, 0) + \epsilon$ by choosing a sufficiently small δ . Let \hat{X} be the mirror version of \tilde{X} with respect to \tilde{X} in the sense that $p_{X\tilde{X}\hat{X}} = p_{X\tilde{X}}p_{\hat{X}|\tilde{X}}$ and $p_{\hat{X}|\tilde{X}} = p_{\tilde{X}|\hat{X}}$.

¹⁰It is shown in Appendix A that $R(D, 0) < \infty$ for $D > 0$.

Obviously, $X \leftrightarrow \check{X} \leftrightarrow \tilde{X} \leftrightarrow \hat{X}$ form a Markov chain and $p_{\check{X}} = p_{\tilde{X}} = p_X$. Moreover,

$$\begin{aligned}
& \mathbb{E}[(X - \hat{X})^2] \\
&= \mathbb{E}[(X - \check{X}) + (\check{X} - \tilde{X}) + (\tilde{X} - \hat{X})^2] \\
&= \mathbb{E}[(X - \check{X})^2] + \mathbb{E}[(\check{X} - \tilde{X})^2] + \mathbb{E}[(\tilde{X} - \hat{X})^2] \\
&\quad + 2\mathbb{E}[(X - \check{X})(\check{X} - \tilde{X})] + 2\mathbb{E}[(X - \check{X})(\tilde{X} - \hat{X})] \\
&\quad + 2\mathbb{E}[(\check{X} - \tilde{X})(\tilde{X} - \hat{X})] \\
&\leq \mathbb{E}[(X - \check{X})^2] + \mathbb{E}[(\check{X} - \tilde{X})^2] + \mathbb{E}[(\tilde{X} - \hat{X})^2] \\
&\quad + 2\sqrt{\mathbb{E}[(X - \check{X})^2]\mathbb{E}[(\check{X} - \tilde{X})^2]} \\
&\quad + 2\sqrt{\mathbb{E}[(X - \check{X})^2]\mathbb{E}[(\tilde{X} - \hat{X})^2]} \\
&\quad + 2\sqrt{\mathbb{E}[(\check{X} - \tilde{X})^2]\mathbb{E}[(\tilde{X} - \hat{X})^2]} \tag{43} \\
&= \mathbb{E}[(X - \check{X})^2] + 4\mathbb{E}[(\check{X} - \tilde{X})^2] \\
&\quad + 4\sqrt{\mathbb{E}[(X - \check{X})^2]\mathbb{E}[(\check{X} - \tilde{X})^2]} \tag{44} \\
&\leq D - \delta + 4\mathbb{E}[(\check{X} - \tilde{X})^2] + 4\sqrt{(D - \delta)\mathbb{E}[(\check{X} - \tilde{X})^2]},
\end{aligned}$$

where (43) is due to the Cauchy-Schwarz inequality, and (44) is due to $\mathbb{E}[(\check{X} - \tilde{X})^2] = \mathbb{E}[(\tilde{X} - \hat{X})^2]$ (implied by the fact that $p_{\check{X}\tilde{X}} = p_{\tilde{X}\hat{X}}$). Since $p_{\check{X}} = p_X$, it follows that $\mathbb{E}[\check{X}^2] < \infty$, which further implies $\mathbb{E}[(\check{X} - \tilde{X})^2] \rightarrow 0$ as $N \rightarrow \infty$. Therefore, we can ensure $\mathbb{E}[(X - \hat{X})^2] \leq D$ by setting N large enough. This completes the verification.

APPENDIX D PROOF OF THEOREM 4

According to Assumption 3, for any $\epsilon > 0$, there exist \tilde{X} and \hat{X} such that $X \leftrightarrow \tilde{X} \leftrightarrow \hat{X}$ form a Markov chain, the support of \tilde{X} , denoted $\tilde{\mathcal{X}}$, satisfies $|\tilde{\mathcal{X}}| < \infty$, and

$$\begin{aligned}
I(X; \tilde{X}) &\leq R(D, 0) + \epsilon, \\
\mathbb{E}[\Delta(X, \tilde{X})] &\leq D, \\
p_{\tilde{X}} &= p_X.
\end{aligned}$$

We shall treat the conditional distribution $p_{X|\tilde{X}}$ induced by $p_{X\tilde{X}\hat{X}}$ specified above as a memoryless channel and establish a soft-covering lemma that is needed for the proof of Theorem 4.

Definition 6. Given a codebook $\mathcal{C}^{(n)} \subseteq \tilde{\mathcal{X}}^n$, let $p_{\mathcal{C}^{(n)}}^{\text{out}}$ denote its induced distribution of the output sequence generated through memoryless channel $p_{X|\tilde{X}}$ by a code-word randomly picked from $\mathcal{C}^{(n)}$ according to the uniform distribution.

Lemma 2. For any $R > I(\tilde{X}; X)$ and $\delta > 0$, there exists a sequence of codebooks $\{\mathcal{C}^{(n)}\}_{n=1}^{\infty}$ with $\mathcal{C}^{(n)} \subseteq \mathcal{T}_{\delta}^{(n)}(p_{\tilde{X}})$ and $|\mathcal{C}^{(n)}| \leq 2^{nR}$ such that

$$\lim_{n \rightarrow \infty} d_{\text{TV}}(p_{\mathcal{C}^{(n)}}^{\text{out}}, p_{\tilde{X}}^n) = 0,$$

where $\mathcal{T}_{\delta}^{(n)}(p_{\tilde{X}})$ denotes the set of δ -typical sequences with respect to $p_{\tilde{X}}$.

Proof: We first briefly review some basic definitions in information-spectrum methods. The *limsup in probability* of a sequence of random variables $\{W_n\}_{n=1}^{\infty}$ is defined as

$$\text{p-lim sup } W_n := \inf \left\{ \tau : \lim_{n \rightarrow \infty} \mathbb{P}\{W_n > \tau\} = 0 \right\}.$$

Correspondingly, the *liminf in probability* is defined as

$$\text{p-lim inf } W_n := -\text{p-lim sup } -W_n.$$

For a sequence of pairs of random variables $(\mathbf{W}, \mathbf{V}) := \{(W^n, V^n)\}_{n=1}^{\infty}$, the *sup-information rate* of (\mathbf{W}, \mathbf{V}) is defined as

$$\bar{I}(\mathbf{W}, \mathbf{V}) := \text{p-lim sup}_{n \rightarrow \infty} \frac{1}{n} \iota_{W^n; V^n}(W^n; V^n)$$

where $\iota_{W^n; V^n} := \log \frac{dp_{W^n V^n}}{d(p_{W^n} p_{V^n})}$ denotes the information density of (W^n, V^n) . For a sequence of distributions $\{(p_{W^n}, p_{V^n})\}_{n=1}^{\infty}$, the *inf-relative-entropy rate* of $\{(p_{W^n}, p_{V^n})\}_{n=1}^{\infty}$ is defined as

$$\underline{D}(\{p_{W^n}\}_{n=1}^{\infty} \| \{p_{V^n}\}_{n=1}^{\infty}) := \text{p-lim inf}_{n \rightarrow \infty} \frac{1}{n} \log \frac{dp_{W^n}}{dp_{V^n}}(W^n).$$

Now we are in a position to prove Lemma 2. Construct a sequence of pairs of random variables $(\tilde{\mathbf{X}}, \mathbf{X}) := \{(\tilde{X}^n, X^n)\}_{n=1}^{\infty}$ with $(\tilde{X}^n, X^n) \sim p_{\tilde{X}X}^n$ and another sequence of pairs of random variables $(\tilde{\mathbf{Y}}, \mathbf{Y}) := \{(\tilde{Y}^n, Y^n)\}_{n=1}^{\infty}$ with $(\tilde{Y}^n, Y^n) \sim p_{\tilde{Y}Y} p_{Y^n|\tilde{Y}^n}$, where $p_{Y^n|\tilde{Y}^n} := p_{X^n|\tilde{X}^n}$ and $p_{\tilde{Y}^n}$ is a truncated version $p_{\tilde{X}^n}^n$ in the sense that

$$p_{\tilde{Y}^n}(y^n) := \begin{cases} \frac{p_{\tilde{X}^n}^n(y^n)}{\mathbb{P}(\tilde{X}^n \in \mathcal{T}_{\delta}^{(n)}(p_{\tilde{X}}))}, & y^n \in \mathcal{T}_{\delta}^{(n)}(p_{\tilde{X}}), \\ 0 & y^n \in \tilde{\mathcal{X}}^n \setminus \mathcal{T}_{\delta}^{(n)}(p_{\tilde{X}}). \end{cases}$$

By the general soft-covering lemma [21, Theorem 4] [22, Corollary VII.4], for any $R > \bar{I}(\tilde{\mathbf{Y}}; \mathbf{Y})$, there exists a sequence of codebooks with $\mathcal{C}^{(n)} \subseteq \mathcal{T}_{\delta}^{(n)}(p_{\tilde{X}})$ and $|\mathcal{C}^{(n)}| \leq 2^{nR}$ such that

$$\lim_{n \rightarrow \infty} d_{\text{TV}}(p_{\mathcal{C}^{(n)}}^{\text{out}}, p_{Y^n}) = 0.$$

Note that

$$\begin{aligned}
d_{\text{TV}}(p_{Y^n}, p_{\tilde{X}^n}^n) &\leq d_{\text{TV}}(p_{\tilde{Y}^n}, p_{\tilde{X}^n}^n) \tag{45} \\
&= \mathbb{P}(\tilde{X} \notin \mathcal{T}_{\delta}^{(n)}) \\
&\rightarrow 0 \text{ as } n \rightarrow \infty,
\end{aligned}$$

where (45) is due to the data processing inequality for the total variation distance. Moreover, in light of the triangle inequality for the total variation distance,

$$d_{\text{TV}}(p_{\mathcal{C}^{(n)}}^{\text{out}}, p_X^n) \leq d_{\text{TV}}(p_{\mathcal{C}^{(n)}}^{\text{out}}, p_{Y^n}) + d_{\text{TV}}(p_{Y^n}, p_X^n).$$

So it suffices to prove $I(\tilde{X}; X) \geq \bar{I}(\tilde{Y}; Y)$.

It can be verified that

$$\begin{aligned} \bar{I}(\tilde{Y}; Y) &= \mathbf{p}\text{-}\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{dp_{\tilde{Y}^n Y^n}}{d(p_{\tilde{Y}^n} p_{Y^n})}(\tilde{Y}^n, Y^n) \\ &= \mathbf{p}\text{-}\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{dp_{X|\tilde{X}}^n}{dp_X^n}(\tilde{Y}^n, Y^n) \\ &\quad + \frac{1}{n} \log \frac{dp_X^n}{dp_{Y^n}}(Y^n) \\ &\leq \mathbf{p}\text{-}\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{dp_{X|\tilde{X}}^n}{dp_X^n}(\tilde{Y}^n, Y^n) \\ &\quad + \mathbf{p}\text{-}\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{dp_X^n}{dp_{Y^n}}(Y^n) \quad (46) \\ &= \mathbf{p}\text{-}\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{dp_{X|\tilde{X}}^n}{dp_X^n}(\tilde{Y}^n, Y^n) \\ &\quad - \underline{D}(\{p_{Y^n}\}_{n=1}^\infty \| \{p_X^n\}_{n=1}^\infty), \end{aligned}$$

where (46) follows by [23, p. 14, the third inequality from the bottom]. Since $\lim_{n \rightarrow \infty} \mathbb{P}(\tilde{X}^n \in \mathcal{T}_\delta^{(n)}(p_{\tilde{X}})) = 1$ and the conditional distribution of (\tilde{X}^n, X^n) given $\tilde{X}^n \in \mathcal{T}_\delta^{(n)}(p_{\tilde{X}})$ is the same as $p_{\tilde{Y}^n, Y^n}$, we must have

$$\begin{aligned} &\mathbf{p}\text{-}\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{dp_{X|\tilde{X}}^n}{dp_X^n}(\tilde{Y}^n, Y^n) \\ &= \mathbf{p}\text{-}\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{dp_{X|\tilde{X}}^n}{dp_X^n}(\tilde{X}^n, X^n). \end{aligned}$$

Moreover, by the weak law of large numbers,

$$\mathbf{p}\text{-}\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{dp_{X|\tilde{X}}^n}{dp_X^n}(\tilde{X}^n, X^n) = I(\tilde{X}; X).$$

Finally, invoking the fact that $\underline{D}(\{p_{Y^n}\}_{n=1}^\infty \| \{p_X^n\}_{n=1}^\infty) \geq 0$ [23, Lemma 3.2.1] completes the proof of Lemma 2. \blacksquare

Now we proceed to prove Theorem 4. Given $R > I(\tilde{X}; X)$ and $\delta > 0$, let $\{\mathcal{C}^{(n)}\}_{n=1}^\infty$ be a sequence of codebooks with the properties specified in Lemma 2. Construct $\mathcal{C}_q^{(n)}$ by applying the shift operator $s_q^{(n)}$ to every codeword of $\mathcal{C}^{(n)}$, $q \in [0 : n-1]$. Let Q be uniformly distributed over $[0 : n-1]$, and \tilde{Y}^n be uniformly distributed over $\mathcal{C}_i^{(n)}$ given $Q = q$, $q \in [0 : n-1]$. Denote the output sequence generated by \tilde{Y}^n through memoryless channel $p_{X|\tilde{X}}$ as \tilde{X}^n . For $q \in [0 : n-1]$,

$$\begin{aligned} d_{\text{TV}}(p_{\tilde{Y}^n|Q=q}, p_X^n) &= d_{\text{TV}}(p_{\tilde{Y}^n|Q=0}, p_X^n) \quad (47) \\ &= d_{\text{TV}}(p_{\mathcal{C}^{(n)}}^{\text{out}}, p_X^n), \end{aligned}$$

where (47) holds because $p_{\tilde{Y}^n|Q=q}$ is simply a shifted version of $p_{\tilde{Y}^n|Q=0}$ while p_X^n is shift-invariant. Since $d_{\text{TV}}(\cdot, \cdot)$ is convex in its first argument, it follows that

$$d_{\text{TV}}(p_{\tilde{Y}^n}, p_X^n) \leq d_{\text{TV}}(p_{\mathcal{C}^{(n)}}^{\text{out}}, p_X^n)$$

and consequently must converge to 0 as $n \rightarrow \infty$.

Our coding scheme can be illustrated using the following probabilistic graphical model:

$$\begin{array}{ccc} X^n & \searrow & \\ \downarrow & & \tilde{X}^n \longrightarrow \hat{X}^n \\ \tilde{X}^n & \nearrow & \end{array}$$

The encoder first leverages the conditional distribution $p_{\tilde{Y}^n|\tilde{Y}^n}$ induced by $p_{\tilde{Y}^n}$ to generate \tilde{X}^n based on X^n . Note that we have $X^n \in \mathcal{X}^n$, $\tilde{X}^n \in \cup_{q=0}^{n-1} \mathcal{C}_q^{(n)}$, and $p_{X^n \tilde{X}^n} = p_X^n p_{\tilde{X}^n|X^n} = p_X^n p_{\tilde{Y}^n|\tilde{Y}^n}$. For $t \in [1 : n]$,

$$\begin{aligned} d_{\text{TV}}(p_{X_t \tilde{X}_t}, p_{X \tilde{X}}) & \\ &\leq d_{\text{TV}}(p_{X_t \tilde{X}_t}, p_{\tilde{Y}_t \tilde{Y}_t}) + d_{\text{TV}}(p_{\tilde{Y}_t \tilde{Y}_t}, p_{X \tilde{X}}) \quad (48) \end{aligned}$$

$$\begin{aligned} &\leq d_{\text{TV}}(p_{X^n \tilde{X}^n}, p_{\tilde{Y}^n \tilde{Y}^n}) + d_{\text{TV}}(p_{\tilde{Y}_t \tilde{Y}_t}, p_{X \tilde{X}}) \quad (49) \\ &= d_{\text{TV}}(p_X^n, p_{\tilde{Y}^n}) + d_{\text{TV}}(p_{\tilde{Y}_t}, p_{\tilde{X}}), \end{aligned}$$

where (48) and (49) are due to the triangle inequality and the data processing inequality for the total variation distance, respectively. Furthermore, we have $d_{\text{TV}}(p_X^n, p_{\tilde{Y}^n}) \rightarrow 0$ as $n \rightarrow \infty$, and $d_{\text{TV}}(p_{\tilde{Y}_t}, p_{\tilde{X}}) \rightarrow 0$ uniformly for all t as $\delta \rightarrow 0$. Therefore,

$$d_{\text{TV}}(p_{X_t \tilde{X}_t}, p_{X \tilde{X}}) \leq \epsilon_\delta$$

for $t \in [1 : n]$ and all sufficiently large n , where $\epsilon_\delta \rightarrow 0$ as $\delta \rightarrow 0$.

For each $t \in [1 : n]$, let¹¹

$$\begin{aligned} &p_{\tilde{X}_t \tilde{X}_t | X_t}(\tilde{x}, \tilde{x} | x) \\ &:= b_{X_t | X_t}(\tilde{x} | x) 1_{\tilde{x}=\tilde{x}}(\tilde{x}, \tilde{x}) \\ &\quad + \begin{cases} \frac{1}{\kappa_t(x)} r_{\tilde{X}_t | X_t}(\tilde{x} | x) r_{\tilde{X}_t | X_t}(\tilde{x} | x), & \kappa_t(x) > 0, \\ 0, & \kappa_t(x) = 0, \end{cases} \\ &\quad (x, \tilde{x}, \tilde{x}) \in \mathcal{X} \times \tilde{\mathcal{X}} \times \tilde{\mathcal{X}}, \end{aligned}$$

¹¹Since $p_{\tilde{Y}^n}$ is shift-invariant and $p_{X|\tilde{X}}$ is memoryless, it follows that $p_{\tilde{Y}^n \tilde{Y}^n}$ and consequently $p_{\tilde{Y}^n|\tilde{Y}^n}$ are shift-invariant as well, which further implies the shift-invariance of $p_{X^n \tilde{X}^n}$ in view of the fact $p_{X^n \tilde{X}^n} = p_X^n p_{\tilde{Y}^n|\tilde{Y}^n}$. Therefore, $p_{\tilde{X}_t | X_t}$, $b_{X_t | X_t}$, $r_{\tilde{X}_t | X_t}$, $r_{\tilde{X}_t | X_t}$, κ_t , and $p_{\tilde{X}_t \tilde{X}_t | X_t}$ are all time-invariant.

where

$$\begin{aligned}
b_{X'_t|X_t}(x'|x) &:= \min\{p_{\tilde{X}_t|X_t}(x'|x), p_{\tilde{X}|X}(x'|x)\}, \\
r_{\tilde{X}_t|X_t}(\tilde{x}|x) &:= p_{\tilde{X}_t|X_t}(\tilde{x}|x) - b_{X'_t|X_t}(\tilde{x}|x), \\
r_{\tilde{X}|X}(\tilde{x}|x) &:= p_{\tilde{X}|X}(\tilde{x}|x) - b_{X'_t|X_t}(\tilde{x}|x), \\
\kappa_t(x) &:= d_{\text{TV}}(p_{\tilde{X}_t|X_t=x}, p_{\tilde{X}|X=x}) \\
&= \sum_{\tilde{x} \in \tilde{\mathcal{X}}} r_{\tilde{X}_t|X_t}(\tilde{x}|x) \\
&= \sum_{\tilde{x} \in \tilde{\mathcal{X}}} r_{\tilde{X}|X}(\tilde{x}|x) \\
&= 1 - \sum_{x' \in \tilde{\mathcal{X}}} b_{X'_t|X_t}(x'|x).
\end{aligned}$$

We claim that $p_{\tilde{X}_t \tilde{X}_t|X_t}$ is a regular conditional distribution. Indeed, on one hand, given each x , $p_{\tilde{X}_t \tilde{X}_t|X_t=x}$ is a distribution (more precisely, a probability mass function) since $p_{\tilde{X}_t \tilde{X}_t|X_t}(\tilde{x}, \tilde{x}|x) \geq 0$, $(\tilde{x}, \tilde{x}) \in \tilde{\mathcal{X}}^2$, and

$$\sum_{\tilde{x}, \tilde{x} \in \tilde{\mathcal{X}}} p_{\tilde{X}_t \tilde{X}_t|X_t}(\tilde{x}, \tilde{x}|x) = \sum_{\tilde{x}} b_{X'_t|X_t}(\tilde{x}|x) + \kappa_t(x) = 1;$$

on the other hand, given each (\tilde{x}, \tilde{x}) , $x \mapsto p_{\tilde{X}_t \tilde{X}_t|X_t}(\tilde{x}, \tilde{x}|x)$ is measurable (which is due to the fact that both $x \mapsto p_{\tilde{X}_t|X_t}(\tilde{x}|x)$ and $x \mapsto p_{\tilde{X}|X}(\tilde{x}|x)$ are measurable), and so is $x \mapsto \check{p}_{\tilde{X}_t \tilde{X}_t|X_t}(\mathcal{B}|x)$ for each $\mathcal{B} \subseteq \tilde{\mathcal{X}}^2$. Moreover, given each x , $\check{p}_{\tilde{X}_t \tilde{X}_t|X_t=x}$ is in fact a maximal coupling of $p_{\tilde{X}_t|X_t=x}$ and $p_{\tilde{X}|X=x}$ since

$$\begin{aligned}
&\sum_{\tilde{x}, \tilde{x} \in \tilde{\mathcal{X}}: \tilde{x} \neq \tilde{x}} p_{\tilde{X}_t \tilde{X}_t|X_t}(\tilde{x}, \tilde{x}|x) \\
&= \frac{1}{\kappa_t(x)} \sum_{\tilde{x} \in \tilde{\mathcal{X}}} r_{\tilde{X}_t|X_t}(\tilde{x}|x) \sum_{\tilde{x} \in \tilde{\mathcal{X}}: \tilde{x} \neq \tilde{x}} r_{\tilde{X}|X}(\tilde{x}|x) \\
&= \frac{1}{\kappa_t(x)} \sum_{\tilde{x} \in \tilde{\mathcal{X}}} r_{\tilde{X}_t|X_t}(\tilde{x}|x) (\kappa_t(x) - r_{\tilde{X}_t|X_t}(\tilde{x}|x)) \\
&= \kappa_t(x) - \frac{1}{\kappa_t(x)} \sum_{\tilde{x} \in \tilde{\mathcal{X}}} r_{\tilde{X}_t|X_t}(\tilde{x}|x) r_{\tilde{X}|X}(\tilde{x}|x) \\
&= \kappa_t(x) \\
&= d_{\text{TV}}(\check{p}_{\tilde{X}_t \tilde{X}_t|X_t=x}, p_{\tilde{X}|X=x}), \quad \kappa_t(x) > 0,
\end{aligned}$$

which clearly also holds when $\kappa_t(x) = 0$.

The encoder leverages $p_{\tilde{X}_t|X_t \tilde{X}_t}$ induced by $p_{X_t \tilde{X}_t \tilde{X}_t} := p_X p_{\tilde{X}_t \tilde{X}_t|X_t}$ to generate \tilde{X}_t from (X_t, \tilde{X}_t) , $t \in [1 : n]$. Note that $p_{X_t \tilde{X}_t} = p_{X \tilde{X}}$ and

$$\begin{aligned}
\mathbb{P}(\tilde{X}_t \neq \tilde{X}_t) &= \int \mathbb{P}(\tilde{X}_t \neq \tilde{X}_t | X_t = t) dp_X(x) \\
&= \int d_{\text{TV}}(p_{\tilde{X}_t|X_t=x}, p_{\tilde{X}|X=x}) dp_X(x) \\
&= d_{\text{TV}}(p_{X_t \tilde{X}_t}, p_{X \tilde{X}}) \\
&\leq \epsilon_\delta, \quad t \in [1 : n],
\end{aligned}$$

when n is sufficiently large.

The encoder then sends \tilde{X}^n to the decoder. We have

$$\begin{aligned}
\frac{1}{n} H(\tilde{X}^n) &\leq \frac{1}{n} H(\tilde{X}^n) + \frac{1}{n} H(\tilde{X}^n | \tilde{X}^n) \\
&\leq R + \frac{\log n}{n} + \frac{1}{n} H(\tilde{X}^n | \tilde{X}^n) \\
&\leq R + \frac{\log n}{n} + \frac{1}{n} \sum_{t=1}^n H(\tilde{X}_t | \tilde{X}_t) \\
&\leq R + \frac{\log n}{n} + \frac{1}{n} \sum_{t=1}^n (H_b(\mathbb{P}(\tilde{X}_t \neq \tilde{X}_t)) \\
&\quad + \mathbb{P}(\tilde{X}_t \neq \tilde{X}_t) \log |\tilde{\mathcal{X}}|) \\
&\leq R + \frac{\log n}{n} + H_b(\epsilon_\delta) + \epsilon_\delta \log |\tilde{\mathcal{X}}|, \quad (51)
\end{aligned}$$

where (50) is due to Fano's inequality, and (51) holds when $\epsilon_\delta \leq \frac{1}{2}$ and n is sufficiently large.

Given \tilde{X}^n , the decoder simply generates \hat{X}^n using the conditional distribution $p_{\hat{X}^n | \tilde{X}^n} := p_{\tilde{X}^n}^n$. It is clear that $p_{X_t \hat{X}_t} = p_{X \hat{X}}$, $t \in [1 : n]$, and consequently

$$\begin{aligned}
\frac{1}{n} \sum_{t=1}^n \mathbb{E}[\Delta(X_t, \hat{X}_t)] &= \mathbb{E}[\Delta(X, \hat{X})] \leq D, \\
\check{p}_{\hat{X}_t} &= p_X, \quad t \in [1 : n].
\end{aligned}$$

Finally, by choosing ϵ, δ sufficiently small, R sufficiently close to $I(X; \tilde{X})$, and n sufficiently large, we can make

$$R + \frac{\log n}{n} + H_b(\epsilon_\delta) + \epsilon_\delta \log |\tilde{\mathcal{X}}|$$

as close to $R(D, 0)$ as we want. This completes the proof of Theorem 4.

APPENDIX E

VERIFICATION OF UNIFORM INTEGRABILITY

Here we verify uniform integrability for the case $\mathbb{E}[X^2] < \infty$, $\mathbb{E}[\hat{X}^2] < \infty$, and $\Delta(x, \hat{x}) := (x - \hat{x})^2$.

By the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\mathbb{E}[(X - \hat{X})^2 1_{\mathcal{E}}(X, \hat{X})] \\
&\leq (\sqrt{\mathbb{E}[X^2 1_{\mathcal{E}}(X, \hat{X})]} + \sqrt{\mathbb{E}[\hat{X}^2 1_{\mathcal{E}}(X, \hat{X})]})^2.
\end{aligned}$$

Note that for any $\chi > 0$,

$$\begin{aligned}
&\mathbb{E}[X^2 1_{\mathcal{E}}(X, \hat{X})] \\
&= \mathbb{P}(X^2 \leq \chi) \mathbb{E}[X^2 1_{\mathcal{E}}(X, \hat{X}) | X^2 \leq \chi] \\
&\quad + \mathbb{P}(X^2 > \chi) \mathbb{E}[X^2 1_{\mathcal{E}}(X, \hat{X}) | X^2 > \chi] \\
&\leq \chi \mathbb{P}((X, \hat{X}) \in \mathcal{E}) + \mathbb{P}(X^2 > \chi) \mathbb{E}[X^2 | X^2 > \chi] \\
&\leq \chi \delta + \mathbb{P}(X^2 > \chi) \mathbb{E}[X^2 | X^2 > \chi].
\end{aligned}$$

Similarly, we have

$$\mathbb{E}[\hat{X}^2 1_{\mathcal{E}}(X, \hat{X})] \leq \chi \delta + \mathbb{P}(\hat{X}^2 > \chi) \mathbb{E}[\hat{X}^2 | \hat{X}^2 > \chi].$$

Since $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[\hat{X}^2] < \infty$, it follows by the dominated convergence theorem that both $\mathbb{P}(X^2 > \chi) \mathbb{E}[X^2 | X^2 > \chi]$ and $\mathbb{P}(\hat{X}^2 > \chi) \mathbb{E}[\hat{X}^2 | \hat{X}^2 > \chi]$ converge to 0 as $\chi \rightarrow \infty$. Therefore, there exists $\chi^* > 0$ such that $\mathbb{P}(X^2 > \chi^*) \mathbb{E}[X^2 | X^2 > \chi^*] \leq \frac{\epsilon}{8}$ and $\mathbb{P}(\hat{X}^2 > \chi^*) \mathbb{E}[\hat{X}^2 | \hat{X}^2 > \chi^*] \leq \frac{\epsilon}{8}$. Setting $\delta = \frac{\epsilon}{8\chi^*}$ ensures

$$\mathbb{E}[(X - \hat{X})^2 1_{\mathcal{E}}(X, \hat{X})] \leq \left(2\sqrt{\chi^* \frac{\epsilon}{8\chi^*}} + \frac{\epsilon}{8}\right)^2 = \epsilon.$$

APPENDIX F PROOF OF THEOREM 6

We shall first prove $\varphi(D) \geq R(\frac{D}{2})$. For any $p_{U\hat{X}|X}$ satisfying (17), (18), and (19), let $V := \mathbb{E}[X|U]$ and $\hat{V} := \mathbb{E}[\hat{X}|U]$. Since $X \leftrightarrow U \leftrightarrow \hat{X}$ form a Markov chain (see (18)), it follows that

$$\begin{aligned} & \mathbb{E}[(X - \hat{X})^2] \\ &= \mathbb{E}[(X - V)^2] + \mathbb{E}[(V - \hat{V})^2] + \mathbb{E}[(\hat{X} - \hat{V})^2], \end{aligned}$$

which, together with (17), implies

$$\min\{\mathbb{E}[(X - V)^2], \mathbb{E}[(\hat{X} - \hat{V})^2]\} \leq \frac{D}{2}.$$

Now consider the case $\mathbb{E}[(X - V)^2] \leq \frac{D}{2}$. Note that

$$\begin{aligned} \max\{I(X; U); I(\hat{X}; U)\} &\geq I(X; U) \\ &\geq I(X; V) \\ &\geq R\left(\frac{D}{2}\right), \end{aligned} \quad (52)$$

where (52) is due to the data processing inequality. By symmetry (see (19)), $\max\{I(X; U); I(\hat{X}; U)\} \geq R(\frac{D}{2})$ continues to hold if $\mathbb{E}[(\hat{X} - \hat{V})^2] \leq \frac{D}{2}$. This proves $\varphi(D) \geq R(\frac{D}{2})$.

Next we proceed to prove $\varphi(D) \leq R(\frac{D}{2})$. For any $p_{V|X}$ satisfying (22), let $U := \mathbb{E}[X|V]$. We have

$$\mathbb{E}[(X - U)^2] \leq \mathbb{E}[(X - V)^2],$$

which, together with (22), implies $\mathbb{E}[(X - U)^2] \leq \frac{D}{2}$. Now construct $p_{\hat{X}U|X}$ such that $p_{\hat{X}U|X} = p_{U|X} p_{\hat{X}|U}$ and $p_{\hat{X}|U} = p_{X|U}$. Note that (18) and (19) are satisfied. Moreover,

$$\begin{aligned} \mathbb{E}[(X - \hat{X})^2] &= \mathbb{E}[(X - U)^2] + \mathbb{E}[(\hat{X} - U)^2] \\ &= 2\mathbb{E}[(X - U)^2] \\ &\leq D. \end{aligned}$$

So (17) is also satisfied. As a consequence,

$$\varphi(D) \leq \max\{I(X; U), I(\hat{X}; U)\}.$$

The proof is complete in view of the fact that

$$\max\{I(X; U), I(\hat{X}; U)\} = I(X; U) \leq I(X; V).$$

APPENDIX G

PROOF OF THEOREM 7

In view of (20), the problem boils down to determining $R(\frac{D}{2})$ by solving the optimization problem in (21). To this end, we need the following lemma.

Lemma 3. *If there exist p_V^* over a finite set $\mathcal{V} \subseteq [0, 1]$ with $p_V^*(v) > 0$, $v \in \mathcal{V}$, and $\lambda \geq 0$ such that*

$$\frac{(1 - \rho)2^{-\lambda v^2}}{\sum_{\tilde{v} \in \mathcal{V}} p_V^*(\tilde{v})2^{-\lambda \tilde{v}^2}} + \frac{\rho 2^{-\lambda(1-v)^2}}{\sum_{\tilde{v} \in \mathcal{V}} p_V^*(\tilde{v})2^{-\lambda(1-\tilde{v})^2}} = 1, \quad v \in \mathcal{V}, \quad (53)$$

$$\frac{(1 - \rho)2^{-\lambda v^2}}{\sum_{\tilde{v} \in \mathcal{V}} p_V^*(\tilde{v})2^{-\lambda \tilde{v}^2}} + \frac{\rho 2^{-\lambda(1-v)^2}}{\sum_{\tilde{v} \in \mathcal{V}} p_V^*(\tilde{v})2^{-\lambda(1-\tilde{v})^2}} \leq 1, \quad v \in [0, 1] \setminus \mathcal{V}, \quad (54)$$

$$\begin{aligned} & (1 - \rho) \frac{\sum_{v \in \mathcal{V}} p_V^*(v)2^{-\lambda v^2}}{\sum_{\tilde{v} \in \mathcal{V}} p_V^*(\tilde{v})2^{-\lambda \tilde{v}^2}} \\ & + \rho \frac{\sum_{v \in \mathcal{V}} p_V^*(v)2^{-\lambda(1-v)^2}(1-v)^2}{\sum_{\tilde{v} \in \mathcal{V}} p_V^*(\tilde{v})2^{-\lambda(1-\tilde{v})^2}} = \frac{D}{2}, \end{aligned} \quad (55)$$

then $p_{V|X}^*$ given by

$$p_{V|X}^*(v|x) := \frac{p_V^*(v)2^{-\lambda(x-v)^2}}{\sum_{\tilde{v} \in \mathcal{V}} p_V^*(\tilde{v})2^{-\lambda(x-\tilde{v})^2}}, \quad x \in \{0, 1\}, \quad v \in \mathcal{V},$$

is an optimal solution to (21).

Proof: It is clear that there is no loss of generality in assuming that V only takes value from $[0, 1]$. Let \mathcal{V}' be an arbitrary finite subset of $[0, 1]$. In view of the standard Karush-Kuhn-Tucker conditions [15, pp. 362–364], (53)–(55) ensures that $p_{V|X}^*$ attains the infimum in (21) when the alphabet of V is restricted¹² to be $\mathcal{V} \cup \mathcal{V}'$. Moreover, according to the support lemma [24, p. 631], it suffices to consider the finite alphabet case; in fact, the alphabet size of V does not need to exceed 3 for the purpose of preserving p_X , $H(X|V)$, and $\mathbb{E}[(X - V)^2]$. So $p_{V|X}^*$ must be an optimal solution to (21). ■

Now we are in a position to solve (21). It suffices to consider the case $D \in (0, 2\rho(1 - \rho))$ since obviously $R(\frac{D}{2})$ equals $H_b(\rho)$ when $D = 0$ and equals 0 when $D \geq 2\rho(1 - \rho)$.

Let $\mathcal{V} := \{a, 1 - a\}$ with

$$a := \frac{1 - \sqrt{1 - 2D}}{2}.$$

Note that $a \in (0, \rho)$. Define p_V^* over \mathcal{V} such that

$$p_V^*(a) = \frac{1 - a - \rho}{1 - 2a}, \quad p_V^*(1 - a) = \frac{\rho - a}{1 - 2a}.$$

¹²We set $p_{V|X}^*(v|x) = 0$ for $v \in \mathcal{V}' \setminus \mathcal{V}$.

Moreover, let

$$\lambda := \frac{1}{1-2a} \log\left(\frac{1-a}{a}\right),$$

which is clearly positive. We shall proceed to verify that the constructed p_V^* and λ satisfying (53)–(55).

Note that

$$\begin{aligned} & \frac{(1-\rho)2^{-\lambda a^2}}{p_V^*(a)2^{-\lambda a^2} + p_V^*(1-a)2^{-\lambda(1-a)^2}} \\ & + \frac{\rho 2^{-\lambda(1-a)^2}}{p_V^*(a)2^{-\lambda(1-a)^2} + p_V^*(1-a)2^{-\lambda a^2}} \\ & = \frac{(1-\rho)}{p_V^*(a) + p_V^*(1-a)\frac{a}{1-a}} + \frac{\rho \frac{a}{1-a}}{p_V^*(a)\frac{a}{1-a} + p_V^*(1-a)} \\ & = 1, \end{aligned} \tag{56}$$

where (56) is due to

$$2^{-\lambda(1-a)^2} = \frac{a}{1-a} 2^{-\lambda a^2}.$$

Similarly,

$$\begin{aligned} & \frac{(1-\rho)2^{-\lambda(1-a)^2}}{p_V^*(a)2^{-\lambda a^2} + p_V^*(1-a)2^{-\lambda(1-a)^2}} \\ & + \frac{\rho 2^{-\lambda a^2}}{p_V^*(a)2^{-\lambda(1-a)^2} + p_V^*(1-a)2^{-\lambda a^2}} \\ & = \frac{(1-\rho)\frac{a}{1-a}}{p_V^*(a) + p_V^*(1-a)\frac{a}{1-a}} + \frac{\rho}{p_V^*(a)\frac{a}{1-a} + p_V^*(1-a)} \\ & = 1. \end{aligned}$$

So (53) indeed holds.

Next let

$$\begin{aligned} \eta(v) & := \frac{(1-\rho)2^{-\lambda v^2}}{p_V^*(a)2^{-\lambda a^2} + p_V^*(1-a)2^{-\lambda(1-a)^2}} \\ & + \frac{\rho 2^{-\lambda(1-v)^2}}{p_V^*(a)2^{-\lambda(1-a)^2} + p_V^*(1-a)2^{-\lambda a^2}}. \end{aligned}$$

We have

$$\begin{aligned} \frac{d}{dv}\eta(v) & = -\frac{\frac{2}{\log e}(1-\rho)\lambda v 2^{-\lambda v^2}}{p_V^*(a)2^{-\lambda a^2} + p_V^*(1-a)2^{-\lambda(1-a)^2}} \\ & + \frac{\frac{2}{\log 2}\rho\lambda(1-v)2^{-\lambda(1-v)^2}}{p_V^*(a)2^{-\lambda(1-a)^2} + p_V^*(1-a)2^{-\lambda a^2}}. \end{aligned}$$

Clearly,

$$\frac{d}{dv}\eta(v) \stackrel{\geq}{\leq} 0$$

if and only if

$$\xi(v) \stackrel{\geq}{<} \log\left(\frac{p_V^*(a)2^{-\lambda(1-a)^2} + p_V^*(1-a)2^{-\lambda a^2}}{p_V^*(a)2^{-\lambda a^2} + p_V^*(1-a)2^{-\lambda(1-a)^2}}\right),$$

where

$$\xi(v) := \log\left(\frac{\rho(1-v)2^{-\lambda(1-v)^2}}{(1-\rho)v2^{-\lambda v^2}}\right).$$

It can be verified that

$$\begin{aligned} & \log\left(\frac{p_V^*(a)2^{-\lambda(1-a)^2} + p_V^*(1-a)2^{-\lambda a^2}}{p_V^*(a)2^{-\lambda a^2} + p_V^*(1-a)2^{-\lambda(1-a)^2}}\right) \\ & = \log\left(\frac{p_V^*(a)\frac{a}{1-a} + p_V^*(1-a)}{p_V^*(a) + p_V^*(1-a)\frac{a}{1-a}}\right) \\ & = \log\left(\frac{\rho}{1-\rho}\right). \end{aligned}$$

On the other hand,

$$\xi(v)|_{v=a, \frac{1}{2}, 1-a} = \log\left(\frac{\rho}{1-\rho}\right).$$

Moreover,

$$\frac{d^2}{dv^2}\xi(v) = \frac{(1-2v)}{v^2(1-v)^2} \log e,$$

which shows that $\xi(v)$ is a strictly convex function for $v \in (0, \frac{1}{2})$ and a strictly concave function for $v \in (\frac{1}{2}, 1)$.

So we must have

$$\xi(v) \begin{cases} > \log\left(\frac{\rho}{1-\rho}\right), & v \in [0, a) \cup (\frac{1}{2}, 1-a), \\ = \log\left(\frac{\rho}{1-\rho}\right), & v = a, \frac{1}{2}, 1-a, \\ < \log\left(\frac{\rho}{1-\rho}\right), & v \in (a, \frac{1}{2}) \cup (1-a, 1], \end{cases}$$

and consequently

$$\frac{d}{dv}\eta(v) \begin{cases} > 0, & v \in [0, a) \cup (\frac{1}{2}, 1-a), \\ = 0, & v = a, \frac{1}{2}, 1-a, \\ < 0, & v \in (a, \frac{1}{2}) \cup (1-a, 1]. \end{cases}$$

This together with (53) implies (54).

Finally, we have

$$\begin{aligned} & (1-\rho)\frac{p_V^*(a)2^{-\lambda a^2} a^2 + p_V^*(1-a)2^{-\lambda(1-a)^2} (1-a)^2}{p_V^*(a)2^{-\lambda a^2} + p_V^*(1-a)2^{-\lambda(1-a)^2}} \\ & + \rho\frac{p_V^*(a)2^{-\lambda(1-a)^2} (1-a)^2 + p_V^*(1-a)2^{-\lambda a^2} a^2}{p_V^*(a)2^{-\lambda(1-a)^2} + p_V^*(1-a)2^{-\lambda a^2}} \\ & = (1-\rho)\frac{p_V^*(a)a^2 + p_V^*(1-a)a(1-a)}{p_V^*(a) + p_V^*(1-a)\frac{a}{1-a}} \\ & + \rho\frac{p_V^*(a)a(1-a) + p_V^*(1-a)a^2}{p_V^*(a)\frac{a}{1-a} + p_V^*(1-a)} \\ & = \frac{D}{2}, \end{aligned}$$

which verifies (55).

In light of Lemma 3, $p_{V|X}^*$ is an optimal solution to (21). Note that

$$\begin{aligned} p_{V|X}^*(a|0) &= \frac{p_V^*(a)2^{-\lambda a^2}}{p_V^*(a)2^{-\lambda a^2} + p_V^*(1-a)2^{-\lambda(1-a)^2}} \\ &= \frac{p_V^*(a)}{p_V^*(a) + p_V^*(1-a)\frac{a}{1-a}} \\ &= \frac{(1-a)(1-a-\rho)}{(1-\rho)(1-2a)}, \\ p_{V|X}^*(a|1) &= \frac{p_V^*(a)2^{-\lambda(1-a)^2}}{p_V^*(a)2^{-\lambda(1-a)^2} + p_V^*(1-a)2^{-\lambda a^2}} \\ &= \frac{p_V^*(a)\frac{a}{1-a}}{p_V^*(a)\frac{a}{1-a} + p_V^*(1-a)} \\ &= \frac{a(1-a-\rho)}{\rho(1-2a)}, \end{aligned}$$

and

$$\begin{aligned} p_{V|X}^*(1-a|0) &= \frac{a(\rho-a)}{(1-\rho)(1-2a)}, \\ p_{V|X}^*(1-a|1) &= \frac{(1-a)(\rho-a)}{\rho(1-2a)}. \end{aligned}$$

The induced $p_{X|V}^*$ is given by

$$\begin{aligned} p_{X|V}^*(0|a) &= 1-a, & p_{X|V}^*(1|a) &= a, \\ p_{X|V}^*(0|1-a) &= a, & p_{X|V}^*(1|1-a) &= 1-a. \end{aligned}$$

As a consequence,

$$\begin{aligned} R\left(\frac{D}{2}\right) &= H_b(\rho) - H_b(a) \\ &= H_b(\rho) - H_b\left(\frac{1 - \sqrt{1-2D}}{2}\right). \end{aligned}$$

REFERENCES

- [1] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE Conf. Comp. Vision and Pattern Recog. (CVPR)*, 2018, pp. 6288–6237.
- [2] Y. Blau and T. Michaeli, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *International Conference on Machine Learning*, pp. 675–685, 2019.
- [3] L. Theis and A. B. Wagner, "A coding theorem for the rate-distortion-perception function," *ICLR 2021 neural compression workshop*.
- [4] R. Matsumoto, "Introducing the perception-distortion tradeoff into the rate-distortion theory of general information sources," *IEICE Comm. Express*, vol. 7, no. 11, pp. 427–431, 2018.
- [5] R. Matsumoto, "Rate-distortion-perception tradeoff of variable-length source coding for general information sources," *IEICE Comm. Express*, vol. 8, no. 2, pp. 38–42, 2019.
- [6] M. Li, J. Klejsa, and W. B. Kleijn, "Distribution preserving quantization with dithering and transformation," *IEEE Signal Process. Lett.*, vol. 17, no. 12, pp. 1014–1017, Dec. 2010.
- [7] M. Li, J. Klejsa, and W. B. Kleijn. (2011). "On distribution preserving quantization. [Online]. Available: <https://arxiv.org/abs/1108.3728>
- [8] J. Klejsa, G. Zhang, M. Li, and W. B. Kleijn, "Multiple description distribution preserving quantization," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6410–6422, Dec. 2013.
- [9] N. Saldi, T. Linder, and S. Yüksel, "Randomized quantization and source coding with constrained output distribution," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 91–106, Jan. 2015.
- [10] N. Saldi, T. Linder, and S. Yüksel, "Output constrained lossy source coding with limited common randomness," *IEEE Trans. Inf. Theory*, vol. 61, no. 9, pp. 4984–4998, Sep. 2015.
- [11] L. Theis and E. Agustsson, "On the advantages of stochastic encoders," *ICLR 2021 neural compression workshop*.
- [12] I. Csiszár and P. C. Shields, "Information theory and statistics: A Tutorial", *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, pp. 417–528, 2004.
- [13] V. van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, Jul. 2014.
- [14] C. T. Li and A. El Gamal, "Strong functional representation lemma and applications to coding theorems," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 6967–6978, Nov. 2018.
- [15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.
- [16] A. B. Wagner, "The rate-distortion-perception tradeoff: The role of common randomness," 2022, arXiv:2202.04147. [Online] Available: <https://arxiv.org/abs/2202.04147>
- [17] Z. Yan, F. Wen, R. Ying, C. Ma, and P. Liu, "On perceptual lossy compression: The cost of perceptual reconstruction and an optimal training framework," in *International Conference on Machine Learning*, 2021.
- [18] G. Zhang, J. Qian, J. Chen, and A. Khisti, "Universal Rate-Distortion-Perception Representations for Lossy Compression," in *Conference on Neural Information Processing Systems*, 2021.
- [19] H. Liu, G. Zhang, J. Chen, A. Khisti, "Lossy compression with distribution shift as entropy constrained optimal transport," in *International Conference on Learning Representations*, 2022.
- [20] R. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.
- [21] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 752–772, Mar. 1993.
- [22] P. Cuff, "Distributed channel synthesis," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7071–7096, Nov. 2013.
- [23] T. S. Han, *Information-Spectrum Methods in Information Theory*. Springer, 2003.
- [24] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.