CROSS-DOMAIN LOSSY COMPRESSION AS OPTIMAL TRANSPORT WITH AN ENTROPY BOTTLENECK

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the problem of cross-domain lossy compression where the reconstruction distribution is different from the source distribution in order to account for distributional shift due to processing. We formulate this as a generalization of optimal transport with an entropy bottleneck to account for the rate constraint due to compression. We provide expressions for the tradeoff between compression rate and the achievable distortion with and without shared common randomness between the encoder and decoder and demonstrate using the example of a binary source that shared randomness can strictly improve the tradeoff. For the case without common randomness and squared-Euclidean distortion, we show that the optimal solution partially decouples into the problem of optimal compression and transport and also characterize the penalty associated with fully decoupling them. We provide experimental results by training deep learning end-to-end compression systems for performing denoising on SVHN and super-resolution on MNIST, and demonstrate consistency with our theoretical results.

1 INTRODUCTION

Using deep neural networks for lossy image compression has proven to be effective, with rate-distortion performance capable of dominating general-purpose image codecs like JPEG, WebP or BPG (Rippel & Bourdev, 2017; Agustsson et al., 2017; Mentzer et al., 2018). More recently, many of these works have included generative aspects within the compression to synthesize realistic elements when the rate is otherwise too low to represent fine-grained details (Tschannen et al., 2018; Agustsson et al., 2019; Mentzer et al., 2020). Though this has been found to deteriorate rate-distortion performance, it has generally resulted in more perceptually-pleasing image reconstruction by reducing artifacts such as pixelation and blur. Using a distributional constraint as a proxy for perceptual measure, several works have subsequently formalized this in a mathematical framework known as the rate-distortion-perception tradeoff (Blau & Michaeli, 2018; 2019; Matsumoto, 2018; 2019; Theis & Wagner, 2021; Yan et al., 2021; Zhang et al., 2021). As is conventional in lossy compression, these works address the scenario in which both low distortion, whereby each individual image reconstruction resembles the ground truth image, and closeness in distribution in which it is not easy to discriminate between image samples from the data-generating distribution and reconstruction distribution, are desirable.

The underlying ideal in conventional compression systems is to have perfect reconstruction with respect to some ground truth input. However this is not the case in applications such as denoising, deblurring, or super-resolution (SR), which require restoration from a degraded input image. In fact, in these cases a ground truth may not even be available. In such applications naturally the reconstruction distribution must match the original source rather than the degraded input distribution. A large body of literature has been devoted to various image restoration tasks, including several methods based on deep learning including both supervised (e.g., (Blau & Michaeli, 2018)) and unsupervised learning methods (e.g., (Wang et al., 2021b)). Although most of the literature exclusively treats compression and restoration separately, in many application they can co-occur. For example, the encoder which records a degraded image may not be co-located with the decoder, but must transmit a compressed version of the image over a digital network. In turn, the decoder must perform both decompression and restoration.

To that end, we study the problem of cross-domain lossy compression in which the reconstruction distribution is different than the source distribution to account for distributional shift due to processing. The problem can be described as a transformation from some source domain to a new target domain

under a rate constraint, which generalizes optimal transport. This readily extends other works which view image restoration under the perception-distortion tradeoff (Blau & Michaeli, 2018) or under optimal transport (Wang et al., 2021b). It also provides a generalization of the rate-distortion-perception problem (Blau & Michaeli, 2019) where the reconstruction distribution must be close to the input distribution. Following (Theis & Agustsson, 2021; Theis & Wagner, 2021), we also utilize *common randomness* as a tool for compression in our setting. Concretely, our results are summarized as follows.

- We provide a formulation of cross-domain lossy compression as a generalization of optimal transport with an entropy constraint and identify the tradeoff between the compression rate and minimum achievable distortion both with and without common randomness at the encoder and decoder. We identify conditions under which the structure of the optimal solution partially decouples the problems of compression and transport, and discuss their architectural implications. In particular for squared error distortion when only individual randomness is present, it suffices to consider transport over optimal rate-distortion encodings between the source and target distributions
- We train deep learning end-to-end compression systems for performing super-resolution on MNIST and denoising on SVHN. Our setup is *unsupervised* and to the best of our knowledge the first to integrate both compression and restoration at once using deep learning. We first demonstrate that by having common randomness at the encoder and decoder the achievable distortion-rate tradeoffs are lower than the case when such randomness is not present. Furthermore, we provide experimental validation of the architectural principle suggested by our theoretical analysis.

2 Theoretical Formulation

We consider a setting where an input $X \sim p_X$ is observed at the encoder, which is a degraded (e.g., noisy, lower resolution, etc) version of the original source. It must be restored to an output $Y \sim p_Y$ at the decoder, where p_Y denotes the target distribution of interest. For example, if X denotes a noise-corrupted image and Y denotes the associated clear reconstruction, then p_Y can be selected to match the distribution of the original source. We will assume p_X and p_Y are probability distributions over $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^n$ and require X and Y to be close with respect to some fidelity metric, which will be measured using a non-negative cost function $d(x, y)$ over $\mathcal{X} \times \mathcal{Y}$. We will refer to $d(\cdot, \cdot)$ as the distortion measure and assume that it satisfies $d(x, y) = 0$ if and only if $x = y$. We further assume that X cannot be directly revealed to the decoder, but instead must be transmitted over a bit interface with an average rate constraint. Such a scenario occurs naturally in many practical systems when the encoder and decoder are not co-located such as communication systems or storage systems. We remark that our approach of measuring the distortion between the degraded input and reconstruction output is applicable in unsupervised image restoration settings where paired training data are not available (e.g. Zhang et al. (2017); Pan et al. (2021); Wang et al. (2021b); Menon et al. (2020)). These works also fix the reconstruction distribution $Y \sim p_Y$ and propose to minimize a distortion metric between the output and the degraded input, but do not consider compression. We next discuss how optimal transport and its extensions are related to our setting of interest.

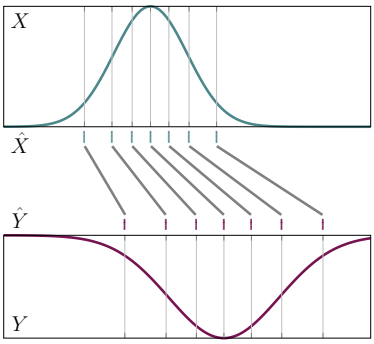


Figure 1: Illustration of Theorem 1 (no common randomness). Given source distribution p_X, target reconstruction distribution p_Y and rate R, we can find one-shot rate-distortion quantizers $\hat{X}$ of X and $\hat{Y}$ of Y and consider transport between them.

2.1 Optimal Transport and Extensions

Definition 1 (Optimal Transport)**.** Let $\Gamma(p_X, p_Y)$ be the set of all joint distributions $p_{X,Y}$ with marginals p_X and p_Y. The classical optimal transport problem is defined as

$$D(p_X, p_Y) = \inf_{p_{X,Y} \in \Gamma(p_X, p_Y)} \mathbb{E}[d(X, Y)], \tag{1}$$

where we refer to each $p_{X,Y} \in \Gamma(p_X, p_Y)$ as a transport plan.

Operationally the optimal transport plan in (1) minimizes the average distortion between the input and output while keeping the output distribution fixed to p_Y. This may generate a transport plan with potentially unbounded entropy, which may not be amenable in a rate-constrained setting. We therefore suggest a generalization to Def. 1 which constrains the entropy of the transport plan. It turns out that having common randomness at the encoder and decoder can help in this setting, so we will distinguish between when it is available and unavailable.

Definition 2 (Optimal Transport with Entropy Bottleneck — no common randomness)**.** Let $M_{\text{ncr}}(p_X, p_Y)$ denote the set of joint distributions $p_{X,Z,Y}$ compatible with the given marginal distributions p_X, p_Y satisfying $p_{X,Z,Y} = p_X p_{Z|X} p_{Y|Z}$. The optimal transport from p_X to p_Y with an entropy bottleneck of R and without common randomness is defined as:

$$\begin{aligned} D_{\text{ncr}}(p_X, p_Y, R) \triangleq & \inf_{p_{X,Z,Y} \in M_{\text{ncr}}(p_X, p_Y)} \mathbb{E}[d(X,Y)] \\ & \text{s.t.} \quad H(Z) \leq R, \end{aligned} \tag{2}$$

where $H(\cdot)$ denotes the Shannon entropy of a random variable.

We note that when the rate constraint R is sufficiently large such that one can select $Z = X$ in (2), then $D_{\text{ncr}}(p_X, p_Y, R) = D(p_X, p_Y)$ in (1). More generally, $D(p_X, p_Y)$ serves as a lower bound for $D_{\text{ncr}}(p_X, p_Y, R)$ for any $R > 0$. Def. 2 also has a natural operational interpretation in our setting. We can view the encoder as implementing the conditional distribution $p_{Z|X}$ to output a representation Z given the input X, and the decoder as implementing the conditional distribution $p_{Y|Z}$ to output the reconstruction Y given the representation Z. The entropy constraint in (2) essentially guarantees that the representation Z can be losslessly transmitted at a rate close to R[1].

It turns out that when we specialize the distortion measure to be the squared Euclidean distance, we can without loss of optimality impose a more structured architecture for implementing the encoder and the decoder. Let $W_2^2(\cdot, \cdot)$ be the squared quadratic Wasserstein distance obtained by setting $d(X, Y) = ||X - Y||^2$ in Def. 1.

Theorem 1. *Let*

$$\begin{aligned} D_{\text{mse}}(p_X, p_Y, R) \triangleq & \inf_{p_{\hat{X}|X}, p_{\hat{Y}|Y}} \mathbb{E}[\|X - \hat{X}\|^2] + \mathbb{E}[\|Y - \hat{Y}\|^2] + W_2^2(p_{\hat{X}}, p_{\hat{Y}}) \\ & \textit{s.t.} \quad \mathbb{E}[X|\hat{X}] = \hat{X}, \quad \mathbb{E}[Y|\hat{Y}] = \hat{Y}, \quad H(\hat{X}) \leq R, \quad H(\hat{Y}) \leq R, \end{aligned} \tag{3}$$

and

$$\begin{aligned} D_{\text{mse}}(p_X, R) \triangleq & \inf_{p_{\hat{X}|X}} \mathbb{E}[\|X - \hat{X}\|^2] \\ & \textit{s.t.} \quad H(\hat{X}) \leq R. \end{aligned} \tag{4}$$

Moreover, let

$$\overline{D}_{\text{ncr}}(p_X, p_Y, R) \triangleq D_{\text{mse}}(p_X, R) + D_{\text{mse}}(p_Y, R) + W_2^2(p_{\hat{X}^*}, p_{\hat{Y}^*}), \tag{5}$$

$$\underline{D}_{\text{ncr}}(p_X, p_Y, R) \triangleq D_{\text{mse}}(p_X, R) + D_{\text{mse}}(p_Y, R), \tag{6}$$

where $p_{\hat{X}^*}$ and $p_{\hat{Y}^*}$ are the marginal distributions induced by the minimizers $p_{\hat{X}^*|X}$ and $p_{\hat{Y}^*|Y}$ that attain $D_{\text{mse}}(p_X, R)$ and $D_{\text{mse}}(p_Y, R)$, respectively (assuming the existence and uniqueness of such minimizers). Then under the squared Eucledian distortion measure,

$$D_{\text{ncr}}(p_X, p_Y, R) = D_{\text{mse}}(p_X, p_Y, R). \tag{7}$$

In addition, we have

$$\overline{D}_{\text{ncr}}(p_X, p_Y, R) \geq D_{\text{ncr}}(p_X, p_Y, R) \geq \underline{D}_{\text{ncr}}(p_X, p_Y, R), \tag{8}$$

and both inequalities are tight when $p_X = p_Y$.

[1]The source coding theorem guarantees that any discrete random variable Z can be losslessly compressed using a variable length code with average length of no more than $H(Z) + 1$ bits.

Theorem 1 deconstructs Z into the auxiliary variables $\hat{X}$ and $\hat{Y}$ which are geometrically representative of X and Y, and decomposes the overall distortion in (2) in terms of the losses due to quantization, transport, and dequantization in (3). It also suggests a natural architecture that partially decouples compression and transport without loss of optimality. First, the sender uses the distribution $p_{\hat{X}|X}$ to produce the compressed representation $\hat{X}$ from X. This is then passed through a "converter" $p_{\hat{Y}|\hat{X}}$ to transform $\hat{X}$ to an optimal representation $\hat{Y}$ of Y. Finally, the receiver maps $\hat{Y}$ back to Y using the conditional distribution $p_{Y|\hat{Y}}$. This is illustrated in Figure 1. The entropy constraint $H(\hat{X}) \leq R$ in (2) essentially guarantees that $\hat{X}$ can be losslessly transmitted to the decoder where the converter can be applied to map $\hat{X}$ to $\hat{Y}$ before outputting Y. Alternately the constraint $H(\hat{Y}) \leq R$ guarantees that the converter could also be implemented at the encoder and then $\hat{Y}$ can be compressed and transmitted to the decoder. Finally note that our proposed architecture is symmetric[2] with respect to the encoder and the decoder and in particular the procedure to transport Y to X would simply be the inverse of transporting X to Y, and indeed the distortion incurred by dequantizing $p_{Y|\hat{Y}}$ is the same as the distortion incurred by quantizing $p_{\hat{Y}|Y}$.

For the special case of same source and target distribution, we have $D_{\text{mse}}(p_X, p_X, R) = 2D_{\text{mse}}(p_X, R)$, implying that the rate required to achieve distortion D under no output distribution constraint (and with the output alphabet relaxed to $\mathbb{R}^n$) achieves distortion $2D$ under the constraint that Y equals X in distribution. This recovers the result of Theorem 2 in Yan et al. (2021) for the one-shot setting. More generally, (8) shows that we may lower bound $D_{\text{mse}}(p_X, p_Y, R)$ by the distortion incurred when compressing X and Y individually, each at rate R, through ignoring the cost of transport. On the other hand, the upper bound corresponds to choosing the optimal rate-distortion representations $\hat{X}^*, \hat{Y}^*$ for X, Y, then considering transport between them. The advantage of this approach is that knowledge of the other respective distribution is not necessary for design. Although not optimal in general, we will, in fact, provide an example where this is optimal in Theorem 4.

Finally, it is worth noting that under mild regularity conditions, the optimal converter $p_{\hat{Y}|\hat{X}}$ can be realized as a (deterministic) bijection, as is in Figure 1.

Theorem 2. *Assume that $D_{\text{ncr}}(p_X, p_Y, R^*)$ is a strictly decreasing function in a neighborhood of R^* and $D_{\text{ncr}}(p_X, p_Y, R^*)$ is attained by $p_{X,\hat{X},\hat{Y},Y}$ as in Theorem 1. Then*

$$H(\hat{X}) = H(\hat{Y}) = R^*, \tag{9}$$

$$\mathbb{E}[\|\hat{X} - \hat{Y}\|^2] = W_2^2(p_{\hat{X}}, p_{\hat{Y}}), \tag{10}$$

and there is a bijection between $\hat{X}$ and $\hat{Y}$.

So far we have focused on the setting when there is no shared common randomness between the encoder and the decoder. We will now consider the setting when a shared random variable denoted by U is present at the encoder and decoder. We assume that the variable U is independent of the input X so that the decoder has no apriori information of the input. We begin with the following definition involving optimal transport.

Definition 3 (Optimal Transport with Entropy Bottleneck — with Common Randomness)**.** Let $M_{\text{cr}}(p_X, p_Y)$ denote the set of joint distributions $p_{U,X,Z,Y}$ compatible with the given marginal distributions p_X, p_Y and satisfying $p_{U,X,Z,Y} = p_U p_X p_{Z|X,U} p_{Y|Z,U}$, where p_U represents the distribution of shared randomness. The optimal transport from p_X to p_Y with entropy bottleneck R and common randomness is defined as

$$\begin{aligned} D_{\text{cr}}(p_X, p_Y, R) \triangleq \inf_{p_{U,X,Z,Y} \in M_{\text{cr}}(p_X, p_Y)} & \mathbb{E}[d(X, Y)] \\ \text{s.t.} \quad & H(Z|U) \leq R. \end{aligned} \tag{11}$$

[2]We say that the problem is symmetric if it is invariant under reversing p_X, p_Y with a new distortion measure defined by reversing the arguments of $d(\cdot, \cdot)$.

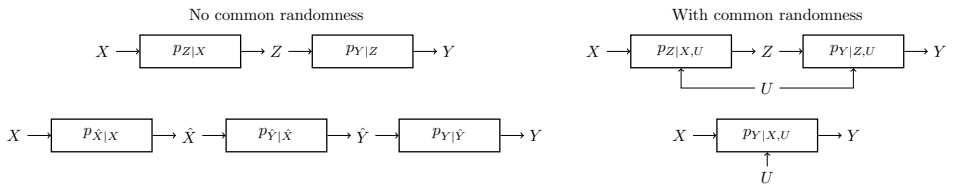


Figure 2: Architectures. Top left: Definition 2. Bottom left: Theorem 1. Top right: Definition 3. Bottom right: Theorem 3. Entropy coding of intermediate representations $Z, \hat{X}, \hat{Y}$ is not shown. For Theorem 3, the division between sender and receiver is across an encoder $C = f(X, U)$ and decoder $Y = g(C, U)$ performing entropy coding along $p_{Y|X,U}$.

Note that we optimize over $p_U, p_{Z|X,U}$, and $p_{Y|Z,U}$ in (11). Note that by setting U to be a constant Def. 3 reduces to the case without shared randomness in Def. 2. Thus it follows that $D_{\text{cr}}(p_X, p_Y, R) \leq D_{\text{ncr}}(p_X, p_Y, R)$ in general. Also from the same argument that was made following Def. 2, we have that $D_{\text{cr}}(p_X, p_Y, R) \geq D(p_X, p_Y)$ in Def. 1. Furthermore as with Def. 2, we can also provide a natural operational interpretation. In particular, given the input X and common randomness U the encoder can output a compressed representation Z using the conditional distribution $p_{Z|X,U}$. The representation Z can be losslessly compressed approximately to an average rate of R again by exploiting the shared randomness U. Finally the decoder, given Z and U can output the reconstruction Y using the conditional distribution $p_{Y|Z,U}$. An interesting difference with Def. 2 is that the setup is no longer symmetric between encoder and decoder, as X is independent of U but Y is not. The following result provides a simplification to the architecture in Def. 3.

Theorem 3. *Let $Q_{cr}(p_X, p_Y)$ denote the set of joint distributions $p_{U,X,Y}$ compatible with the given marginals p_X, p_Y satisfying $p_{U,X,Y} = p_U p_X p_{Y|U,X}$ as well as $H(Y|U, X) = 0$. Then*

$$D_{\text{cr}}(p_X, p_Y, R) = \inf_{p_{U,X,Y} \in Q_{cr}(p_X, p_Y)} \mathbb{E}[d(X, Y)] \quad \textit{s.t.} \quad H(Y|U) \leq R. \tag{12}$$

Before discussing the implications of Theorem 3 we remark on a technical point. Because the Shannon entropy is defined only for discrete random variables, U must be chosen in a way such that $Y|U = u$ is discrete for each u, even for continuous (X, Y). This is known to be possible, e.g. Li & El Gamal (2018) have provided a general construction for a U with this property, with additional quantitative guarantees to ensure that U is informative of Y. In the finite alphabet case we show in Appendix A.3 that optimization of U can be formulated as a linear program.

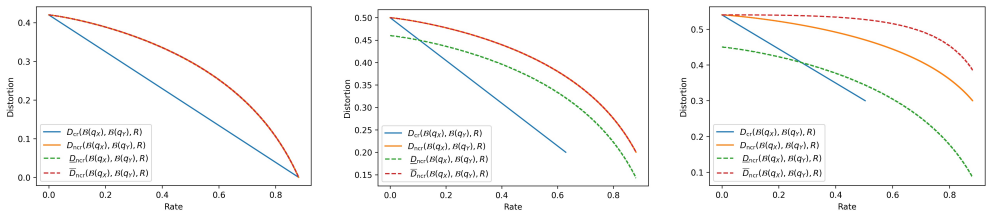
We next discuss the implication of Theorem 3. First note that the problem can be modelled with only $p_{Y|U,X}$ producing a reconstruction Y without the need for the intermediate representation Z, much like the conventional optimal transport in Def. 1. The condition $H(Y|U, X) = 0$ also implies that the transport plan is deterministic when conditioned on the shared randomness, which plays the role of stochasticity. Furthermore in this architecture the encoder should compute the representation Y given the source X and the shared random-variable U (which corresponds to the transport problem) and then compress it lossless at a rate close to $H(Y|U)$ (which corresponds to the compression problem). The receiver only needs to decompress and reconstruct Y. This is in contrast to the case without common randomness in Theorem 1 where the reconstruction Y must be generated at the decoder.

2.2 Example: Binary Case

Let $X \sim \mathcal{B}(q_X)$ and $Y \sim \mathcal{B}(q_Y)$ be two Bernoulli random variables with $q_X, q_Y \in (0, 1)$, and let $d(\cdot, \cdot)$ be the Hamming distortion measure $d_H(\cdot, \cdot)$ (i.e., $d_H(x, y) = 0$ if $x = y$ and $d_H(x, y) = 1$ otherwise). Let $D^{(B)}_{\min} \triangleq |q_X - q_Y|$ and $D^{(B)}_{\max} \triangleq q_X + q_Y - 2q_X q_Y$. Note that $D^{(B)}_{\min}$ is the total variation distance between $\mathcal{B}(q_X)$ and $\mathcal{B}(q_Y)$, which is the minimum $\mathbb{E}[d_H(X, Y)]$ achievable by coupling X and Y. On the other hand, we have $\mathbb{E}[d_H(X, Y)] = D^{(B)}_{\max}$ for X, Y independent.

Theorem 4. *Assume Hamming distortion measure. Under no common randomness, we have*

$$D_{\text{ncr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), R) = \begin{cases} -\frac{2(1-q_X)(1-q_Y)}{1-H_b^{-1}(R)} + 2 - q_X - q_Y, & q_X + q_Y \leq 1, \\ -\frac{2q_X q_Y}{1-H_b^{-1}(R)} + q_X + q_Y, & q_X + q_Y > 1, \end{cases} \tag{13}$$



(a) (b) (c)

Figure 3: Binary case distortion-rate tradeoffs. (a) $q_X = q_Y = 0.3$, where $\overline{D}_{\text{ncr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), R)$ and $\underline{D}_{\text{ncr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), R)$ coincide with $D_{\text{ncr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), R)$; (b) $q_X = 0.3, q_Y = 0.5$, where $\overline{D}_{\text{ncr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), R)$ is tight but $\underline{D}_{\text{ncr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), R)$ is loose; (b) $q_X = 0.3, q_Y = 0.6$, where both bounds are loose. In general, as long as $\mathcal{B}(q_X)$ and $\mathcal{B}(q_Y)$ are biased toward the same symbol (i.e., $q_X, q_Y \leq 1/2$ or $q_X, q_Y \geq 1/2$), the upper bound $\overline{D}_{\text{ncr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), R)$ is tight and blindly using optimal quantizer and dequantizer in the conventional rate-distortion sense incurs no penalty. Moreover, it can be seen from all these examples that common randomness strictly improves the distortion-rate tradeoff (except at some extreme point(s)).

for $R \in [0, \min\{H_b(q_X), H_b(q_Y)\}]$, *and* $= D_{\min}^{(B)}$ *for* $R > \min\{H_b(q_X), H_b(q_Y)\}$. *Moreover,*

$$D_{\text{ncr}}(\mathcal{B}(q_X) = \overline{D}_{\text{ncr}}(\mathcal{B}(q_X) \tag{14}$$

for $R \in [0, \min\{H_b(q_X), H_b(q_Y)\}]$ *if* $q_X, q_Y \leq 1/2$ *or* $q_X, q_Y \geq 1/2$. *Here,* $H_b^{-1}(R)$ *denotes the inverse of the binary entropy function on* $[0, 1/2]$. *With common randomness,*

$$D_{\text{cr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), R) = -\frac{2(1-q_X)q_X R}{H_b(q_X)} + D_{\max}^{(B)} \tag{15}$$

for $R \in [0, \rho H_b(q_X)]$, *and* $= D_{\min}^{(B)}$ *for* $R > \rho H_b(q_X)$. *Here,* $\rho \triangleq \min\{q_Y/q_X, (1-q_Y)/(1-q_X)\}$.

A comparison of these functions can be viewed in Figure 3.

3 Experimental Results

We use two sets of results to illustrate that the principles derived from our theoretical results are applicable to practical compression with deep learning. Importantly, we assume an *unsupervised* setting in which we have only unpaired noisy and clean images available to us, as in Wang et al. (2021b). Our first experiment is, to the best of our knowledge, the first in which restoration and compression are performed jointly using deep learning. We will furthermore demonstrate the utility of common randomness in this setting. Stochasticity was necessary in our theoretical results both with and without common randomness. In practice, we will use generative models to induce domain shift, and stochasticity is necessary to train an effective (rate-constrained) generative model at low rates to produce variety in the reconstructions. The first set of experiments compare the rate-distortion tradeoffs achieved by quantization schemes with and without common randomness using a rate constrained architecture designed for image denoising.

The second set of experiments are designed on the principle of Theorem 1. In addition to the generator trained from our first experiment, we will construct a helper network to allow us to estimate the decomposition (3). This is then compared with the direct loss between the noisy image and rate-constrained denoising reconstruction. If the losses are close, this would suggest that the decomposition is not only without loss of optimality but also effective in practice.

3.1 Rate-Distortion Comparison with Common Randomness

Let p_X be a degraded source distribution and p_Y be the target distribution. Our goal is to compress and reconstruct some noisy X so that the reconstruction semantically resembles X within the target distribution p_Y. For our applications, we will take the MSE loss as a fidelity criterion. Let f be an encoder, Q, and g be a decoder. For a given rate R with common randomness U available, we have a problem of the form

$$\min_{f,g,Q} \quad \|X - g(Q(f(X,U)))\|_2^2$$

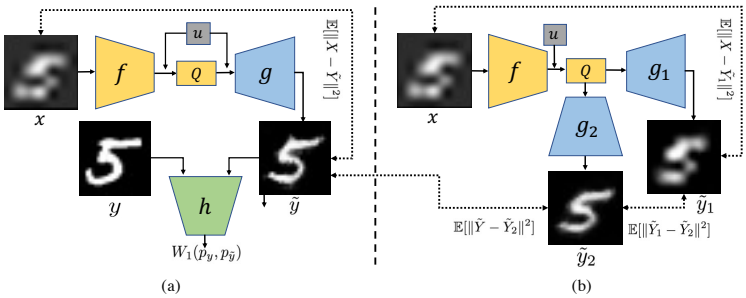


Figure 4: Illustration of our experimental setup. (a) shows the end-to-end learning system with common randomness, where the encoder and decoder have access to the same randomness u. (b) presents the network setup for verifying the architecture principle given in Theorem 1.

$$\text{s.t.} \quad p_{g(Q(f(X,U)))} = p_Y, \quad H(Q(f(X,U))|U) \leq R,$$

which uses parameterized neural networks to implement (11). In practice, we do not impose a hard equality constraint and allow some relaxation. We also fix Q and assume f and g are sufficiently expressive to map to these fixed quantization points. Let $\tilde{Y} \triangleq g(Q(f(X,U)))$. We will use a penalty on the Wasserstein-1 distance between $p_{\tilde{Y}}$ and p_Y in accordance with the Wasserstein GAN (Arjovsky et al., 2017) framework, so that our system is a stochastic rate-constrained autoencoder with GAN regularization. Specifically, we follow the network shown in Fig. 4(a) which in addition to f, Q, and g contains critic h.

For the realization of common randomness in Definition 3, we adopt the universal quantization scheme of Ziv (1985); Theis & Agustsson (2021). Given trained f and g and degraded image X, we generate restored image $\tilde{Y}$ through

$$\tilde{Y} = g(Q(f(X) + U) - U), \tag{16}$$

where U is the stochastic noise shared by the sender and receiver. Details about the quantization are provided in Appendix B.2. To find an appropriate f and g, we use the relaxed objective

$$L_1 = \mathbb{E}[\|X - \tilde{Y}\|^2] + \lambda W_1(p_Y, p_{\tilde{Y}}), \tag{17}$$

which is the sum of the MSE and Wasserstein-1 losses weighted by λ. By optimizing our network using this objective, we see two favorable properties. First, the Wasserstein-1 loss ensures the distribution of output is close to that of target images, i.e. $p_{\tilde{Y}} \approx p_Y$ for sufficiently large λ. Moreover, the MSE loss that pushes the output $\tilde{Y}$ to input X ensures that the output structurally resembles X. Consequently, the training objective allows the output $\tilde{Y}$ to be clear and preserves content from input.

To generate a rate-distortion trade-off curve, we modify the encoder to produce a different number of symbols ranging from low bit rate to high bit rate and record the MSE distortion loss between noisy inputs and denoised outputs. Fig. 5(a) and Fig. 5(c) show the curves for image super-resolution and image denoising. We also show some qualitative results in Fig. 5(b) and 5(d). As the rate increases, the generated high-quality images are clearer.

As exemplified by the Bernoulli distribution in Theorem 4, common randomness can help reduce the rate that is needed for reconstruction given a specific distortion. Equivalently, given a fixed rate, a system with common randomness can perform better than one without common randomness. To demonstrate this in practice, we conduct the following experiment. We remove the common randomness setup from the framework in Section 3.1 and alternatively add two independent noises U_1 and U_2 to the encoder and decoder sides. Concretely, under the new setting, (16) becomes

$$\tilde{Y} = g(Q(f(X) + U_1) - U_2) \tag{18}$$

Then we conduct training using the objective (17) as in the common randomness. The tradeoff curve without common randomness for both tasks are shown in Fig. 5(a) and 5(c) with orange dots. Performance of the framework is better when there is common randomness.

3.2 Architectural Principle

In the case without common randomness, Theorem 1 implies that (under the rate constraint) the overall distortion $\mathbb{E}[\|X - Y\|^2]$ can be decomposed to the summation of the three distortion terms

$$\mathbb{E}[\|X - Y\|^2] = \mathbb{E}[\|X - \hat{X}\|^2] + \mathbb{E}[\|Y - \hat{Y}\|^2] + W_2^2(p_{\hat{X}}, p_{\hat{Y}}), \tag{19}$$

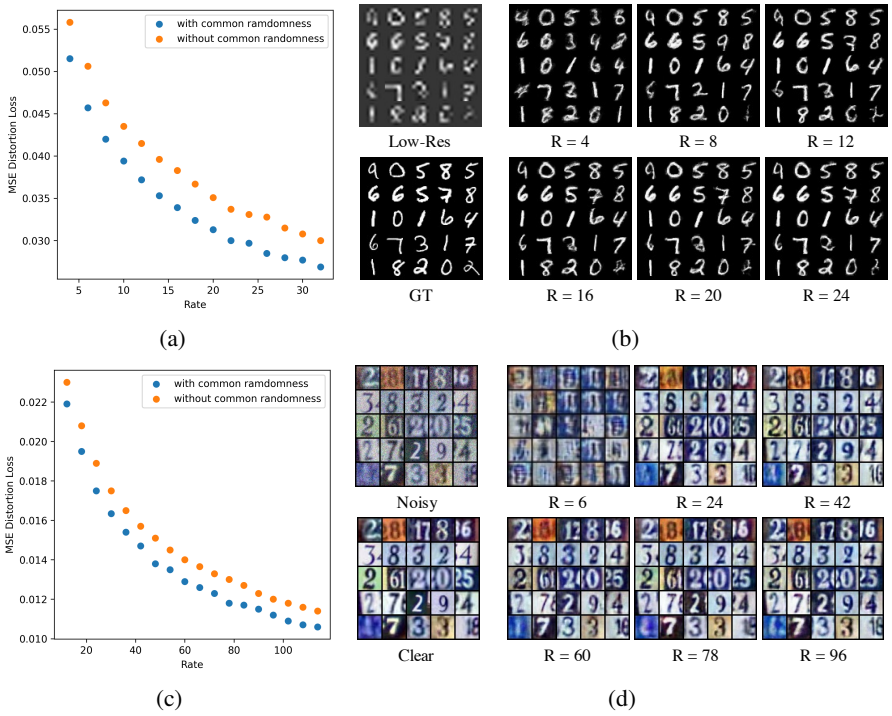


Figure 5: (a)(b) The experimental results of 4 times image super-resolution. (c)(d) The experimental results of image denoising. The noise pattern is synthesized by additive Gaussian noise with standard deviation set to 20. (a)(c) Rate-distortion trade-offs. Blue points are the MSE distortion loss for a particular rate under the setting of using common randomness, while orange points illustrate the same trade-off without using common randomness. For both tasks, at any rate, the performance of using common randomness is better than the case without common randomness. (b)(d) Examples for outputs from several models with different rates. As the rate increases, the outputs become clearer.

where $\hat{X}$ and $\hat{Y}$ are some representations of X and Y under MSE distortion. The chosen rate-distortion representations $\hat{X}$ for X and $\hat{Y}$ for Y must not only be representative of X and Y, but also enjoy low cost of transport between one another. We now seek to estimate the overhead of this decomposition in practice.

However, due to the nature of the deep learning framework, the distortion measure between images and compressed representations cannot be explicitly measured. Thus, we alternatively develop a two-branch network to compare the summation of the three distortion components (19) to the overall distortion. First, we take trained f and g from the previous experiment and freeze their weights. Given noisy input X, we encode it through f, then decoder g_1 is trained to minimize the distortion with X, and decoder g_2 is trained to minimize the distortion with $\tilde{Y} = f(g(X))$ which distributionally approximates a clean restoration (we use $\tilde{Y}$ instead of ground truth because we assume an unsupervised setting). Let

$$\tilde{Y}_1 = g_1(f(X)), \quad \tilde{Y}_2 = g_2(f(X)).$$

The idea here is that $\tilde{Y}_1$ is a rate-constrained reconstruction of X and $\tilde{Y}_2$ is a rate-constrained reconstruction of $\tilde{Y}$, both of which are produced from compressing X using f. We assume this is reasonable as in light of Theorem 2, there is no loss of optimality in doing so given sufficiently expressive neural networks. The overall decomposed loss is then given by

$$L_2 = \underbrace{\mathbb{E}[\|X - \tilde{Y}_1\|^2]}_{\text{(a)}} + \underbrace{\mathbb{E}[\|\tilde{Y} - \tilde{Y}_2\|^2]}_{\text{(b)}} + \underbrace{\mathbb{E}[\|\tilde{Y}_1 - \tilde{Y}_2\|^2]}_{\text{(c)}}, \tag{20}$$

in which $\tilde{Y}_1$ approximates $\hat{X}$ and $\tilde{Y}_2$ approximates $\hat{Y}$. Training is performed jointly over g_1 and g_2.

One additional point is that g_1 and g_2 can also be trained separately, although in this case we can no longer assume that f can be reused without loss of optimality, as this objective would not be equivalent to minimizing (20) (there is no control over (c)). We develop an additional experiment in which we optimize encoder-decoder pairs f_1, g_1 to minimize (a) and f_2, g_2 to minimize (b),

Table 1: Results of architecture principle in Theorem 1. At any rate, it can be observed that the total losses of our approximation system (L_2) are very close to that of end-to-end learning system under the setting without common randomness ($\mathbb{E}[\|X - \tilde{Y}\|^2]$).

	Super-resolution				Denoising			
Rate	4	10	20	30	12	30	60	90
End-to-end	0.0558	0.0435	0.0351	0.0308	0.0230	0.0175	0.0140	0.0123
Joint	0.0586	0.0453	0.0349	0.0309	0.0243	0.0192	0.0158	0.0139

where now $\tilde{Y}_1$ is produced using f_1, g_1 and $\tilde{Y}_2$ using f_2, g_2. In this setting, we aim to approximate the rate-distortion optimal $\tilde{X}^*$ and $\tilde{Y}^*$ corresponding to $\overline{D}_{\mathrm{ncr}}(p_X, p_Y, \cdot)$ from Theorem 1 using $\tilde{Y}_1$ and $\tilde{Y}_2$, and in doing this separate optimization it is clear that we will drive down (a) and (b) but increase (c). As it turns out, the resultant values (a) and (b) obtained during joint optimization are not much worse than the values from separate optimization. This suggests that in practice, the optimal rate-distortion representations (i.e. under objective (4)) can be leveraged for the general objective (20) without much loss of optimality, which further suggests that the encoder f_1 can be potentially trained without knowledge of p_Y without much performance loss. These can be viewed in Table 2.

Table 2: Comparison of separate vs. joint training for (20). See the above paragraph for explanation.

Super-resolution

Rate	4		10		20		30	
Method	Joint	Separate	Joint	Separate	Joint	Separate	Joint	Separate
$\mathbb{E}[\|X - \tilde{Y}_1\|^2]$	0.0355	0.0356	0.0223	0.0214	0.0136	0.0113	0.0092	0.0083
$\mathbb{E}[\|\tilde{Y} - \tilde{Y}_2\|^2]$	0.0227	0.0216	0.0222	0.0206	0.0191	0.0191	0.0172	0.0155

Denoising

Rate	12		30		60		90	
Method	Joint	Separate	Joint	Separate	Joint	Separate	Joint	Separate
$\mathbb{E}[\|X - \tilde{Y}_1\|^2]$	0.0191	0.0190	0.0146	0.0145	0.0117	0.0123	0.0104	0.0107
$\mathbb{E}[\|\tilde{Y} - \tilde{Y}_2\|^2]$	0.0050	0.0046	0.0044	0.0040	0.0038	0.0035	0.0032	0.0030

4 Related Works

Sinkhorn distances (Cuturi, 2013) are a formulation of optimal transport with a penalty term corresponding to the mutual information between the source and target distributions. This has been studied in information theory literature (e.g. (Bai et al., 2020; Wang et al., 2021a)). For source coding in particular, Saldi et al. (2014; 2015) consider common randomness with constrained output distribution. Blau & Michaeli (2018) evaluated a number of deep image restoration techniques and somewhat counter-intuitively demonstrated a tradeoff between optimizing for distortion and "perceptual quality", i.e. realism. This is explained by the fact that the output of the conventional rate-distortion objective can differ significantly from the source distribution. Wang et al. (2021b) model the shift in distribution due to degradation as an optimal transport problem. Meanwhile, output-constrained lossy compression has also been shown to improve perceptual quality (Tschannen et al., 2018), leading to the rate-distortion-perception framework (Blau & Michaeli, 2019).

5 Conclusion

We consider the setting of cross-domain lossy compression, in which we compress across different source and target distributions. We formulate this as an entropy-constrained optimal transport problem and provide expressions for characterizing the tradeoff between compression rate and the minimum achievable distortion. Using the example of binary sources, we demonstrate that shared common randomness between the encoder and decoder can strictly improve this tradeoff. We also develop a number of architectural principles through our theoretical results and provide experimental validations by training deep learning models for super-resolution and denoising tasks over compressed representations.

REFERENCES

Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc Van Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. pp. 1142–1152, 2017.

Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 221–231, 2019.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Yikun Bai, Xiugang Wu, and Ayfer Özgür. Information constrained optimal transport: From talagrand, to marton, to cover. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2210–2215. IEEE, 2020.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237, 2018.

Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pp. 675–685, 2019.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

Abbas El Gamal and Young-Han Kim. *Network information theory*. Cambridge university press, 2011.

Robert M Gray and Thomas G Stockham. Dithered quantizers. *IEEE Transactions on Information Theory*, 39(3):805–812, 1993.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pp. 5767–5777, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Cheuk Ting Li and Abbas El Gamal. Strong functional representation lemma and applications to coding theorems. *IEEE Transactions on Information Theory*, 64(11):6967–6978, 2018.

Ryutaroh Matsumoto. Introducing the perception-distortion tradeoff into the rate-distortion theory of general information sources. *IEICE Communications Express*, 7(11):427–431, 2018.

Ryutaroh Matsumoto. Rate-distortion-perception tradeoff of variable-length source coding for general information sources. *IEICE Communications Express*, 8(2):38–42, 2019.

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 2437–2445, 2020.

Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Conditional probability models for deep image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4394–4402, 2018.

Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Oren Rippel and Lubomir Bourdev. Real-time adaptive image compression. In *International Conference on Machine Learning*, pp. 2922–2930, 2017.

Naci Saldi, Tamás Linder, and Serdar Yüksel. Randomized quantization and source coding with constrained output distribution. *IEEE Transactions on Information Theory*, 61(1):91–106, 2014.

Naci Saldi, Tamás Linder, and Serdar Yüksel. Output constrained lossy source coding with limited common randomness. *IEEE Transactions on Information Theory*, 61(9):4984–4998, 2015.

Shibani Santurkar, David Budden, and Nir Shavit. Generative compression. In *2018 Picture Coding Symposium (PCS)*, pp. 258–262. IEEE, 2018.

Lucas Theis and Eirikur Agustsson. On the advantages of stochastic encoders. *arXiv preprint arXiv:2102.09270*, 2021.

Lucas Theis and Aaron B Wagner. A coding theorem for the rate-distortion-perception function. *arXiv preprint arXiv:2104.13662*, 2021.

Michael Tschannen, Eirikur Agustsson, and Mario Lucic. Deep generative models for distribution-preserving lossy compression. In *Advances in Neural Information Processing Systems*, pp. 5929–5940, 2018.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Shuchan Wang, Photios A Stavrou, and Mikael Skoglund. Generalized talagrand inequality for sinkhorn distance using entropy power inequality. *arXiv preprint arXiv:2109.08430*, 2021a.

Wei Wang, Fei Wen, Zeyu Yan, Rendong Ying, and Peilin Liu. Optimal transport for unsupervised restoration learning. *arXiv preprint arXiv:2108.02574*, 2021b.

Zeyu Yan, Fei Wen, Rendong Ying, Chao Ma, and Peilin Liu. On perceptual lossy compression: The cost of perceptual reconstruction and an optimal training framework. In *International Conference on Machine Learning*, 2021.

George Zhang, Jingjing Qian, Jun Chen, and Ashish Khisti. Universal rate-distortion-perception representations for lossy compression. *arXiv preprint arXiv:2106.10311*, 2021.

Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3929–3938, 2017.

Jacob Ziv. On universal quantization. *IEEE Transactions on Information Theory*, 31(3):344–347, 1985.

A THEORETICAL RESULTS

A.1 DISTORTION-RATE VS RATE-DISTORTION FORMULATION

In addition to Definitions 2 and 3, we can equivalently define

$$R_{\mathrm{ncr}}(p_X, p_Y, D) \triangleq \inf_{p_{X,Z,Y} \in M_{\mathrm{ncr}}(p_X, p_Y)} H(Z) \quad \text{s.t.} \quad \mathbb{E}[d(X,Y)] \le D, \tag{21}$$

$$R_{\mathrm{cr}}(p_X, p_Y, D) \triangleq \inf_{p_{U,X,Z,Y} \in M_{\mathrm{cr}}(p_X, p_Y)} H(Z|U) \quad \text{s.t.} \quad \mathbb{E}[d(X,Y)] \le D. \tag{22}$$

$D_{\mathrm{ncr/cr}}(p_X, p_Y, R)$ and $R_{\mathrm{ncr/cr}}(p_X, p_Y, D)$ are monotonically decreasing in R and D, respectively, so they are the inverse of each other. Sometimes it is more convenient to work with this rate-distortion formulation.

A.2 PROOFS OF THEORETICAL RESULTS

Proof of Theorem 1. For any $p_{XZY} \in M_{\mathrm{ncr}}(p_X, p_Y)$ with $H(Z) \le R$,

$$\begin{aligned}\mathbb{E}[\|X-Y\|^2] &= \mathbb{E}[\|X - \mathbb{E}[X|Z]\|^2] + \mathbb{E}[\|Y - \mathbb{E}[Y|Z]\|^2] + \mathbb{E}[\|\mathbb{E}[X|Z] - \mathbb{E}[Y|Z]\|^2] \\ &\ge D_{\mathrm{mse}}(p_X, p_Y, R),\end{aligned}$$

where the last equality follows from the definition of $D_{\mathrm{mse}}(p_X, p_Y, R)$ and the fact that

$$\max\{H(\mathbb{E}[X|Z]), H(\mathbb{E}[Y|Z]\} \le H(Z) \le R.$$

As a consequence, we must have $D_{\mathrm{ncr}}(p_X, p_Y, R) \ge D_{\mathrm{mse}}(p_X, p_Y, R)$. On the other hand, for any $p_{\hat{X}|X}$, $p_{\hat{Y}|Y}$ with $\mathbb{E}[X|\hat{X}] = \hat{X}$, $\mathbb{E}[Y|\hat{Y}] = \hat{Y}$, $H(\hat{X}) \le R$, and $H(\hat{Y}) \le R$, we can construct a joint distribution $p_{X\hat{X}\hat{Y}Y}$ such that $X \leftrightarrow \hat{X} \leftrightarrow \hat{Y} \leftrightarrow Y$ form a Markov chain, $p_{X\hat{X}} = p_X p_{\hat{X}|X}$, $p_{Y\hat{Y}} = p_Y p_{\hat{Y}|Y}$, and $p_{\hat{X}\hat{Y}}$ satisfying $\mathbb{E}[\|\hat{X} - \hat{Y}\|^2] = W_2^2(p_{\hat{X}}, p_{\hat{Y}})$. Note that

$$\begin{aligned}\mathbb{E}[\|X-Y\|^2] &= \mathbb{E}[\|X-\hat{X}\|^2] + \mathbb{E}[\|Y-\hat{Y}\|^2] + \mathbb{E}[\|\hat{X}-\hat{Y}\|^2] \\ &= \mathbb{E}[\|X-\hat{X}\|^2] + \mathbb{E}[\|Y-\hat{Y}\|^2] + W_2^2(p_{\hat{X}}, p_{\hat{Y}}).\end{aligned} \tag{23}$$

Let $Z \triangleq \hat{X}$. It can be verified that $p_{XZY} \in M_{\mathrm{ncr}}(p_X, p_Y)$ and $H(Z) = H(\hat{X}) \le R$, which, together with (23), implies $D_{\mathrm{ncr}}(p_X, p_Y, R) \le D_{\mathrm{mse}}(p_X, p_Y, R)$. This completes the proof of (7).

Dropping the term $W_2^2(p_{\hat{X}}, p_{\hat{Y}})$ in (3) yields

$$D_{\mathrm{ncr}}(p_X, p_Y, R) \ge \tilde{D}_{\mathrm{mse}}(p_X, R) + \tilde{D}_{\mathrm{mse}}(p_Y, R),$$

where

$$\tilde{D}_{\mathrm{mse}}(p_X, R) \triangleq \inf_{p_{\hat{X}|X}} \mathbb{E}[\|X-\hat{X}\|^2] \quad \text{s.t.} \quad \mathbb{E}[X|\hat{X}] = \hat{X}, \quad H(\hat{X}) \le R.$$

and $\tilde{D}_{\mathrm{mse}}(p_Y, R)$ is definely analogously. On the other hand, choosing $p_{\hat{X}|X} = p_{\hat{X}'|X}$ and $p_{\hat{Y}|Y} = p_{\hat{Y}'|Y}$ in (3) gives

$$D_{\mathrm{ncr}}(p_X, p_Y, R) \le \tilde{D}_{\mathrm{mse}}(p_X, R) + \tilde{D}_{\mathrm{mse}}(p_Y, R) + W_2^2(p_{\hat{X}'}, p_{\hat{Y}'}),$$

where $p_{\hat{X}'|X}$ and $p_{\hat{Y}'|Y}$ are the minimizers that attain $\tilde{D}_{\mathrm{mse}}(p_X, R)$ and $\tilde{D}_{\mathrm{mse}}(p_Y, R)$ respectively while $p_{\hat{X}'}$ and $p_{\hat{Y}'}$ are their induced marginal distributions. It is clear that $p_{\hat{X}'|X}$ and $p_{\hat{Y}'|Y}$ coincide

with $p_{\hat{X}^*|X}$ and $p_{\hat{Y}^*|Y}$ respectively as the constraints $\mathbb{E}[X|\hat{X}] = \hat{X}$ and $\mathbb{E}[Y|\hat{Y}] = \hat{Y}$ are automatically satisfied by $p_{\hat{X}^*|X}$ and $p_{\hat{Y}^*|Y}$. This proves (8). For the special case $p_X = p_Y$, we have $p_{\hat{X}^*|X} = p_{\hat{Y}^*|Y}$ and consequently the upper bound and the lower bound in (8) coincide.

Note that due to the involvement of conditional expectation, $\hat{X}$ is not necessarily defined over $\mathcal{X}$ if $\mathcal{X}$ is a strict subset of $\mathbb{R}^n$ (for the same reason, $\hat{Y}$ is not necessarily defined over $\mathcal{Y}$). In other words, the output of the quantizer is not consrained to the input alphabet and needs to be relaxed to $\mathbb{R}^n$. As such, $D_{\text{mse}}(p_X, R)$ should be interpreted as the one-shot distortion-rate function with the reconstruction alphabet being $\mathbb{R}^n$, which is in general strictly below its counterpart with the reconstruction alphabet being $\mathcal{X}$ (also known as the distortion-rate-perception function with an inactive perception constraint) if $\mathcal{X}$ is a strictly subset of $\mathbb{R}^n$. This subtle issue, which is often overlooked in the literature, arises when one deals with discrete X and Y (see the binary example in Section 2.2 and A.4). □

Proof of Theorem 2. We have that $\max\{H(\hat{X}), H(\hat{Y})\} \leq R^*$. If one of them, say $H(Y)$, is less than R^*, this will lead to a contradiction by the following argument. Note that

$$D^* = \mathbb{E}[\|X - Y\|^2] = \mathbb{E}[\|X - \hat{X}\|^2] + \mathbb{E}[\|Y - \hat{Y}\|^2] + \mathbb{E}[\|\hat{X} - \hat{Y}\|^2],$$

which depends on $p_{X\hat{X}\hat{Y}Y}$ only through $p_{X\hat{X}}$, $p_{\hat{X}\hat{Y}}$, and $p_{Y\hat{Y}}$. We can construct a new joint distribution $p_{X\hat{X}'\hat{Y}'Y}$ such that $p_{X\hat{X}'} = p_{X\hat{X}}$, $p_{\hat{X}'\hat{Y}'} = p_{\hat{X}\hat{Y}}$, $p_{Y\hat{Y}'} = p_{Y\hat{Y}}$, and $X \leftrightarrow \hat{X}' \leftrightarrow \hat{Y}' \leftrightarrow Y$ form a Markov chain. Denote $\hat{X}'$ by Z'. It is clear that the induced $p_{XZ'Y}$ belongs to $M_{\text{ncr}}(p_X, p_Y)$, preserves $\mathbb{E}[\|X - Y\|^2]$, and $H(Z') < R^*$, which is contradictory with the fact that $D_{\text{ncr}}(p_X, p_{\hat{X}}, R)$ is a strictly decreasing function in a neighborhood of $R = R^*$ since R can be set slightly below R^* without violating the constraint $H(Z') \leq R$. This proves (9), which futher implies the existence of a bijection between $\hat{X}$ and $\hat{Y}$.

It remains to prove (10). If (10) does not hold, then we can find some $p_{\hat{X}''\hat{Y}''}$ with $p_{\hat{X}''} = p_{\hat{X}}$ and $p_{\hat{Y}''} = p_{\hat{Y}}$ such that $\mathbb{E}[\|\hat{X}'' - \hat{Y}''\|^2] < \mathbb{E}[\|\hat{X} - \hat{Y}\|^2]$. Leverage this $p_{\hat{X}''\hat{Y}''}$ to construct a new joint distribution $p_{X\hat{X}''\hat{Y}''Y}$ such that $p_{X\hat{X}''} = p_{X\hat{X}}$, $p_{Y\hat{Y}''} = p_{Y\hat{Y}}$, and $X \leftrightarrow \hat{X}'' \leftrightarrow \hat{Y}'' \leftrightarrow Y$ form a Markov chain. Denote $\hat{X}''$ by Z''. It is clear that the induced $p_{XZ''Y}$ belongs to $M_{\text{ncr}}(p_X, p_Y)$, $H(Z'') = R^*$, and

$$\begin{aligned}
\mathbb{E}[\|X - Y\|^2] &= \mathbb{E}[\|X - \hat{X}''\|^2] + \mathbb{E}[\|Y - \hat{Y}''\|^2] + \mathbb{E}[\|\hat{X}'' - \hat{Y}''\|^2] \\
&< \mathbb{E}[\|X - \hat{X}\|^2] + \mathbb{E}[\|Y - \hat{Y}\|^2] + \mathbb{E}[\|\hat{X} - \hat{Y}\|^2] \\
&= D^*,
\end{aligned}$$

which is contradictory with the fact that $R_{\text{ncr}}(p_X, p_Y, D)$ is a strictly decreasing function in a neighborhood of $D = D^*$ since D can be set slightly below D^* without violating the constraint $\mathbb{E}[\|X - Y\|^2] \leq D$. So in conclusion, the converter is a one-to-one mapping, which induces an optimal coupling that attains $W_2^2(p_{\hat{X}}, p_{\hat{Y}})$. □

Proof of Theorem 3. Choosing $p_{U,X,Y}$ from $W_2^2(p_X, p_Y)$ and setting $Z = Y$ shows that

$$\begin{aligned}
D_{\text{cr}}(p_X, p_Y, R) \leq \inf_{p_{UXY} \in W(p_X, p_Y)} & \mathbb{E}[d(X, Y)] \\
\text{s.t.} \quad & H(Y|U) \leq R.
\end{aligned}$$

So it remains to prove that this upper bound is tight. In the light of the functional representation lemma, for any (U, X, Z, Y) with $p_{UXZY} \in M_{\text{cr}}(p_X, p_Y)$, there exist V_1, independent of (U, X), and V_2, independent of (U, X, V_1), as well as determintic mappings ϕ_1 and ϕ_2 such that $Z = \phi_2(U, X, V_1)$ and $Y = \phi_2(Z, V_2)$. Let $U' \triangleq (U, V_1, V_2)$. Clearly, $p_{U',X,Y} = p_{U'} p_X p_{Y|U',X}$ and $H(Y|U', X) = 0$. Moreover, we have

$$\begin{aligned}
H(Z|U) &\geq H(Z|U') \\
&\geq H(Y|U'),
\end{aligned}$$

where the last inequality is due to the fact that Y is determined by (Z, U'). Therefore,

$$D_{\text{cr}}(p_X, p_Y, R) \geq \inf_{p_{U'XY} \in W(p_X, p_Y)} \mathbb{E}[d(X, Y)]$$

$$\text{s.t.} \quad H(Y|U') \leq R.$$

This completes the proof of (12).

Note that each realization of U is associated with a deterministic function from $\mathcal{X}$ to $\mathcal{Y}$. As a consequence, the problem boils down to optimizing the probablity distribution defined over this collection of functions. For the finite alphabet case, there are totally $|\mathcal{Y}|^{|\mathcal{X}|}$ such functions. In fact, a simple application of the support lemma shows that only $|\mathcal{Y}|+1$ functions need to be assigned with a positive probability. □

A.3 Linear Program Formulation for Common Randomness

In the finite alphabet case, we can formulate Theorem 3 as follows:

$$\begin{aligned} D_{\text{cr}}(p_X, p_Y, R) = \min_{p_U} & \sum_{u\in\mathcal{U}} p_U(u)\mathbb{E}[d(X, f_u(X))] \\ \text{s.t.} \quad & \sum_{u\in\mathcal{U}} p_U(u)H(f_u(X)) \leq R, \\ & \sum_{u\in\mathcal{U}} p_U(u)\mathbb{P}(f_u(X)=y) = p_Y(y), \quad y\in\mathcal{Y}, \end{aligned}$$

where p_U is defined over $\mathcal{U} \triangleq \{1, 2, \cdots, |\mathcal{Y}|^{|\mathcal{X}|}\}$, and $\{f_u : u \in \mathcal{U}\}$ is the set of all distinct functions from $\mathcal{X}$ to $\mathcal{Y}$. By the support lemma (Appendix C on page 631 of El Gamal & Kim (2011)), only $|\mathcal{Y}|+1$ functions need to be assigned with a positive probability, although it remains to determine whether it is practical to find these functions.

A.4 Binary Case

Proof of Theorem 4.

With common randomness.

There are totally 4 distinct functions from $\{0,1\}$ to $\{0,1\}$:

$$f_1(x) = x, \quad f_2(x) = 1-x, \quad f_3(x) = 0, \quad f_4(x) = 1, \quad x \in \{0,1\}.$$

Therefore, we have

$$\begin{aligned} &\sum_{u\in\mathcal{U}} p_U(u)H(f_u(X)) = H_b(q_X)(p_U(1)+p_U(2)), \\ &\sum_{u\in\mathcal{U}} p_U(u)\mathbb{E}[d_H(X, f_u(X))] = p_U(2) + q_X p_U(3) + (1-q_X)p_U(4), \\ &\sum_{u\in\mathcal{U}} p_U(u)\mathbb{P}(f_u(X)=1) = p_U(1)q_X + (1-q_X)p_U(2) + p_U(4). \end{aligned}$$

In light of Theorem 3,

$$\begin{aligned} R_{\text{cr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), D) = \min_{p_U(1),\cdots,p_U(4)} & H_b(q_X)(p_U(1)+p_U(2)) \\ \text{s.t.} \quad & p_U(2) + q_X p_U(3) + (1-q_X)p_U(4) \leq D, && (24) \\ & q_X p_U(1) + (1-q_X)p_U(2) + p_U(4) = q_Y, && (25) \\ & p_U(1)+p_U(2)+p_U(3)+p_U(4) = 1, && (26) \\ & p_U(1), p_U(2), p_U(3), p_U(4) \geq 0. && (27) \end{aligned}$$

Note that

$$\begin{aligned} & p_U(2) + q_X p_U(3) + (1-q_X)p_U(4) \\ & = p_U(2) + q_X p_U(3) + (1-q_X)(1-p_U(1)-p_U(2)-p_U(3)) \end{aligned} \tag{28}$$

$$= -(1-q_X)p_U(1) + q_X p_U(2) + (2q_X - 1)p_U(3) + 1 - q_X, \tag{29}$$

where (28) is due to (26). Moreover, it follows by (25) and (26) that

$$p_U(3) = -(1-q_X)p_U(1) - q_X p_U(2) + 1 - q_Y. \tag{30}$$

Substituting (30) into (29) and invoking the fact that $p_U(2) \geq 0$ gives

$$\begin{aligned}
&p_U(2) + q_X p_U(3) + (1-q_X)p_U(4) \\
&= -2(1-q_X)q_X(p_U(1)+p_U(2)) + 4(1-q_X)q_X p_U(2) + D_{\max}^{(B)} \\
&\geq -2(1-q_X)q_X(p_U(1)+p_U(2)) + D_{\max}^{(B)},
\end{aligned}$$

which, together with (24), implies

$$p_U(1) + p_U(2) \geq \frac{1}{2(1-q_X)q_X}(D_{\max}^{(B)} - D).$$

As a consequence, we must have

$$R_{\mathrm{cr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), D) \geq \frac{H_b(q_X)}{2(1-q_X)q_X}(D_{\max}^{(B)} - D).$$

One can readily verify that this lower bound is tight as it is attained by p_U^* with

$$\begin{aligned}
p_U^*(1) &= \frac{1}{2(1-q_X)q_X}(D_{\max}^{(B)} - D), \\
p_U^*(2) &= 0, \\
p_U^*(3) &= -\frac{1}{2q_X}(D_{\max}^{(B)} - D) + 1 - q_Y, \\
p_U^*(4) &= -\frac{1}{2(1-q_X)}(D_{\max}^{(B)} - D) + q_Y,
\end{aligned}$$

which satisfies (24)–(27) for $D \in [D_{\min}^{(B)}, D_{\max}^{(B)}]$. The expression of $D_{\mathrm{cr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), R)$ can be obtained by taking the inverse of $R_{\mathrm{cr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), D)$.

Without common randomness.

We will rely on some results which will come after this proof.

Note that Hamming distortion coincides with MSE distortion when $\mathcal{X} = \mathcal{Y} = \{0, 1\}$. So Theorem 1, Lemma 1, and Lemma 2 are applicable here. In particular, in light of Lemmas 1 and 2, for any $R \geq 0$ and $\epsilon > 0$, there exists a joint distribution $p_{X\hat{X}\hat{Y}Y}$ compatible with the given marginal distributions p_X and p_Y such that $\hat{X}$ and $\hat{Y}$ are deterministically related finite-support random variables with $H(\hat{X}) = H(\hat{Y}) \leq R$ and $X \leftrightarrow \hat{X} \leftrightarrow \hat{Y} \leftrightarrow Y$ form a Markov chain; moreover, $\hat{X} = \mathbb{E}[X|\hat{X}]$, $\hat{Y} = \mathbb{E}[Y|\hat{Y}]$, and

$$\mathbb{E}[\|X - \hat{X}\|^2] + \mathbb{E}[\|Y - \hat{Y}\|] + \mathbb{E}[\|\hat{X} - \hat{Y}\|^2] \leq D_{\mathrm{ncr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), R) + \epsilon. \tag{31}$$

Without loss of generality, we assume $\hat{X}$ and $\hat{Y}$ take value from $\{\hat{x}_i\}_{i=1}^N$ and $\{\hat{y}_i\}_{i=1}^N$, respectively, and $\hat{Y} = \psi(\hat{X})$, where ψ is a bijection from $\{\hat{x}_i\}_{i=1}^N$ to $\{\hat{y}_i\}_{i=1}^N$ with $\hat{y}_i = \psi(\hat{x}_i)$, $i = 1, \cdots, N$. Let $\theta_i \triangleq p_{\hat{X}}(\hat{x}_i)$, or equivalently, $\theta_i \triangleq p_{\hat{Y}}(\hat{y}_i)$, $i = 1, \cdots, N$. Note that $\hat{X} = \mathbb{E}[X|\hat{X}]$ and $\hat{Y} = \mathbb{E}[Y|\hat{Y}]$ if and only if $p_{X|\hat{X}}(1|\hat{x}_i) = \hat{x}_i$ and $p_{Y|\hat{Y}}(1|\hat{y}_i) = \hat{y}_i$ for $\theta_i > 0$, $i = 1, \cdots, N$. So the constraints $\sum_{i=1}^N p_{\hat{X}}(\hat{x}_i)p_{X|\hat{X}}(1|\hat{x}_i) = q_X$ and $\sum_{i=1}^N p_{\hat{Y}}(\hat{y}_i)p_{Y|\hat{Y}}(1|\hat{y}_i) = q_Y$ can be written equivalently as $\sum_{i=1}^N \theta_i \hat{x}_i = q_X$ and $\sum_{i=1}^N \theta_i \hat{y}_i = q_Y$. Moreover, it is easy to verify that

$$\begin{aligned}
&\mathbb{E}[\|X - \hat{X}\|^2] + \mathbb{E}[\|Y - \hat{Y}\|^2] + \mathbb{E}[\|\hat{X} - \hat{Y}\|^2] \\
&= \sum_{i=1}^N \theta_i(1-\hat{x}_i)\hat{x}_i + \sum_{i=1}^N \theta_i(1-\hat{y}_i)\hat{y}_i + \sum_{i=1}^n \theta_i(\hat{x}_i - \hat{y}_i)^2
\end{aligned}$$

$$= \sum_{i=1}^{N} \theta_i(\hat{x}_i + \hat{y}_i - 2\hat{x}_i\hat{y}_i).$$

In light of Theorem 1 and (31), the following optimization problem (P) yields an upper bound on $D_{\mathrm{ncr}}(p_X, p_Y, R)$ with a gap at most ϵ:

$$\min_{(\theta_i,\hat{x}_i,\hat{y}_i,)_{i=1}^N} \sum_{i=1}^{N} \theta_i(\hat{x}_i + \hat{y}_i - 2\hat{x}_i\hat{y}_i) \qquad \text{(P)}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} \theta_i \log \frac{1}{\theta_i} \le R, \quad \sum_{i=1}^{N} \theta_i = 1, \quad \sum_{i=1}^{N} \theta_i\hat{x}_i = q_X, \quad \sum_{i=1}^{N} \theta_i\hat{y}_i = q_Y,$$
$$\theta_i \ge 0, \quad \hat{x}_i \in [0,1], \quad \hat{y}_i \in [0,1], \quad i = 1, \cdots, N.$$

Given $(\theta_i, \hat{y}_i)_i^N$, (P) degenerates to a linear programming problem with respect $(\hat{x}_i)_{i=1}^N$ over hyper-rectangle $[0,1]^N$ subject to the constraint $\sum_{i=1}^N \theta_i\hat{x}_i = q_X$, for which the minimum is attained at a point on an edge of $[0,1]^N$. Therefore, it suffices to consider $(\hat{x}_i)_{i=1}^N$ with at most one element different from 0 and 1. By a similar argument, it can be shown that there is no loss of optimality in assuming that at most one of $\hat{y}_i$, $i = 1, \cdots, N$, takes value other than 0 or 1. Due to the merge of different elements, the one-to-one relationship might not be preserved. Nevertheless, by Lemma 2, we just need to consider deterministically related $\hat{X}$ and $\hat{Y}$ with support size at most 3. Applying the linear programming argument to (P) with $N = 3$ shows that, at the cost of potentially compromising the one-to-one relationship, at most one element in the support of $\hat{X}$ as well as the support of $\hat{Y}$ need to take value different from 0 and 1. In the case that the bijection is lost, $\hat{X}$ or $\hat{Y}$ must have a reduced support size. One can restore the bijection by invoking Lemma 2, then use the linear programming argument to assign extreme values to all but at most one element in the support. Following this line of reasoning, we can conclude that the attention can be restricted to deterministically related $\hat{X}$ and $\hat{Y}$ with support size at most 3 and at most one element in the support different from 0 and 1. Moreover, the following configurations can be excluded.

1. Support size $= 3$ and the existence of pairs $(\hat{x}, \hat{y})$ and $(\hat{x}', \hat{y}')$ for some $\hat{x} > \hat{x}'$ and $\hat{y} < \hat{y}'$ ($(\hat{x}, \hat{y})$ is said to be a pair if $\hat{X} = \hat{x} \Leftrightarrow \hat{Y} = \hat{y}$): Since

$$\begin{aligned}
&(\hat{x} - \hat{y})^2 + (\hat{x}' - \hat{y}')^2 - (\hat{x} - \hat{y}')^2 - (\hat{x}' - \hat{y})^2 \\
&= -2\hat{x}\hat{y} - 2\hat{x}'\hat{y}' + 2\hat{x}\hat{y}' + 2\hat{x}'\hat{y} \\
&= 2(\hat{x} - \hat{x}')(\hat{y}' - \hat{y}) \\
&> 0,
\end{aligned}$$

   it follows that $\mathbb{E}[\|\hat{X} - \hat{Y}\|^2]$ can be strictly reduced by moving the same amount of probability from $\{\hat{X} = \hat{x}, \hat{Y} = \hat{y}\}$ to $\{\hat{X} = \hat{x}, \hat{Y} = \hat{y}'\}$ and from $\{\hat{X} = \hat{x}', \hat{Y} = \hat{y}'\}$ to $\{\hat{X} = \hat{x}', \hat{Y} = \hat{y}\}$. This modification does not affect $p_{\hat{X}}$ and $p_{\hat{Y}}$, and consequently $H(\hat{X})$, $H(\hat{Y})$, $\mathbb{E}[\|X - \hat{X}\|^2]$, $\mathbb{E}[\|Y - \hat{Y}\|^2]$ remain the same. So the distortion-rate performance of this configuration is strictly suboptimal.

2. Support size $= 2$ and the existence of pairs $(\hat{x}, \hat{y})$ and $(\hat{x}', \hat{y}')$ for some $\hat{x} > \hat{x}'$ and $\hat{y} < \hat{y}'$: Same as configuration 1).

3. Support size $= 2$ and existence of pairs $(\hat{x}, 1)$ and $(0, \hat{y})$ for some $\hat{x} \in (0,1)$ and $\hat{y} \in (0,1)$: It follows by $\mathbb{E}[X|\hat{X}] = \hat{X}$ and $\mathbb{E}[Y|\hat{Y}] = \hat{Y}$ that

$$\begin{aligned}
p_{\hat{X}\hat{Y}}(\hat{x}, 1) &= \frac{q_X}{\hat{x}}, \\
p_{\hat{X}\hat{Y}}(0, \hat{y}) &= 1 - \frac{q_X}{\hat{x}}, \\
\hat{y} &= 1 - \frac{1 - q_Y}{1 - \frac{q_X}{\hat{x}}}.
\end{aligned}$$

Clearly, $H(\hat{X}) = H(\hat{Y}) = H_b(\frac{q_X}{\hat{x}})$. Since $\hat{x} \in (0,1)$ and $\hat{y} \in (0,1)$, we must have $q_X < \frac{q_X}{\hat{x}} < q_Y$, which implies $H(\hat{X}) = H(\hat{Y}) > \min\{H(X), H(Y)\}$. Furthermore, it can be verified that

$$\begin{aligned}
&\mathbb{E}[\|X - \hat{X}\|^2] + \mathbb{E}[\|Y - \hat{Y}\|^2] + \mathbb{E}[\|\hat{X} - \hat{Y}\|^2] \\
&= \frac{q_X}{\hat{x}}\hat{x}(1-\hat{x}) + \left(1 - \frac{q_X}{\hat{x}}\right)\hat{y}(1-\hat{y}) + \frac{q_X}{\hat{x}}(1-\hat{x})^2 + \left(1 - \frac{q_X}{\hat{x}}\right)\hat{y}^2 \\
&= q_Y - q_X.
\end{aligned}$$

However, this end-to-end distortion is obviously achievable when $R = \min\{H(X), H(Y)\}$. So the rate-distortion performance of this configuration is strictly suboptimal.

4. Support size $= 2$ and the existence of pairs $(\hat{x}, 0)$ and $(1, \hat{y})$ for some $\hat{x} \in (0,1)$ and $\hat{y} \in (0,1)$: Same as configuration 3).

In view of the excluded configurations, we are left with the case where $p_{\hat{X}\hat{Y}}$ assigns all probabilities to $\{\hat{X} = 0, \hat{Y} = 0\}$, $\{\hat{X} = 1, \hat{Y} = 1\}$, and $\{\hat{X} = \hat{x}, \hat{Y} = \hat{y}\}$ for some $\hat{x} \in [0,1]$ and $\hat{y} \in [0,1]$. So it suffices to consider the $N = 3$ version of (P) with $\hat{x}_1 = \hat{y}_1 = 0$, $\hat{x}_3 = \hat{y}_3 = 1$, $\hat{x}_2 = \hat{x}$, and $\hat{y}_2 = \hat{y}$. The constraints $\sum_{i=1}^3 \theta_i = 1$, $\sum_{i=1}^3 \theta_i \hat{x}_i = q_X$, and $\sum_{i=1}^3 \theta_i \hat{y}_i = q_Y$ imply

$$\begin{aligned}
\theta_1 &= 1 - q_X - (1-\hat{x})\theta, \\
\theta_3 &= q_X - \hat{x}\theta, \\
\hat{y} &= \frac{q_Y - q_X}{\theta} + \hat{x}.
\end{aligned}$$

In this way, we get a simplified optimization problem (P'):

$$\begin{aligned}
\min_{\theta, \hat{x}} \; & 2\hat{x}(1-\hat{x})\theta + (1-2\hat{x})(q_Y - q_X) \qquad \text{(P')} \\
\text{s.t.} \quad & (1 - q_X - (1-\hat{x})\theta)\log\frac{1}{1 - q_X - (1-\hat{x})\theta} + \theta\log\frac{1}{\theta} + (q_X - \hat{x}\theta)\log\frac{1}{q_X - \hat{x}\theta} \le R, \\
& \hat{x} \in [0,1], \quad \theta \in [0,1], \quad (1-\hat{x})\theta \in [q_Y - q_X, 1 - q_X], \quad \hat{x}\theta \in [q_X - q_Y, q_X].
\end{aligned}$$

Note that (P') does not depend on ϵ and consequently yields the exact characterization of $D_{\text{ncr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), R)$. Therefore, $R_{\text{ncr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), D)$ is characterized by the following optimization problem (P''):

$$\begin{aligned}
\min_{\theta, \hat{x}} \; & (1 - q_X - (1-\hat{x})\theta)\log\frac{1}{1 - q_X - (1-\hat{x})\theta} + \theta\log\frac{1}{\theta} + (q_X - \hat{x}\theta)\log\frac{1}{q_X - \hat{x}\theta} \qquad \text{(P'')} \\
\text{s.t.} \quad & 2\hat{x}(1-\hat{x})\theta + (1-2\hat{x})(q_Y - q_X) \le D, \\
& \hat{x} \in [0,1], \quad \theta \in [0,1], \quad (1-\hat{x})\theta \in [q_Y - q_X, 1 - q_X], \quad \hat{x}\theta \in [q_X - q_Y, q_X].
\end{aligned}$$

Given $\hat{x}$, the objective function of (P'') is concave in θ and consequently its minimum is attained at an endpoint of $[\underline{\theta}, \overline{\theta}]$, where

$$\begin{aligned}
\underline{\theta} &\triangleq \max\left\{0, \frac{q_Y - q_X}{1-\hat{x}}, \frac{q_X - q_Y}{\hat{x}}\right\}, \\
\overline{\theta} &\triangleq \min\left\{1, \frac{1 - q_X}{1-\hat{x}}, \frac{q_X}{\hat{x}}, \frac{D - (1-2\hat{x})(q_Y - q_X)}{2\hat{x}(1-\hat{x})}\right\}.
\end{aligned}$$

Without loss of generality, we assume $q_Y \ge q_X$ and $q_X + q_Y \le 1$. The following statements can be easily verified.

1. For $\hat{x} \in [0, \frac{D + q_X - q_Y}{2(1-q_Y)}]$,

$$\begin{aligned}
\underline{\theta} &= \frac{q_Y - q_X}{1-\hat{x}}, \\
\overline{\theta} &= \frac{1 - q_X}{1-\hat{x}}.
\end{aligned}$$

2. For $\hat{x} \in (\frac{D+q_X-q_Y}{2(1-q_Y)}, \frac{q_X+q_Y-D}{2q_Y}]$,

$$\underline{\theta} = \frac{q_Y - q_X}{1-\hat{x}},$$
$$\overline{\theta} = \frac{D-(1-2\hat{x})(q_Y-q_X)}{2\hat{x}(1-\hat{x})}.$$

3. For $\hat{x} \in (\frac{q_X+q_Y-D}{2q_Y}, \frac{q_X}{q_Y}]$,

$$\underline{\theta} = \frac{q_Y - q_X}{1-\hat{x}},$$
$$\overline{\theta} = \frac{q_X}{\hat{x}}.$$

4. For $\hat{x} > \frac{q_X}{q_Y}$, $[\underline{\theta}, \overline{\theta}]$ is empty.

Note that when $\hat{x} \in [0, \frac{q_X}{q_Y}]$ and $\theta = \underline{\theta}$,

$$\begin{aligned}
&(1-q_X-(1-\hat{x})\theta)\log\frac{1}{1-q_X-(1-\hat{x})\theta} + \theta\log\frac{1}{\theta} + (q_X-\hat{x}\theta)\log\frac{1}{q_X-\hat{x}\theta}\\
&= (1-q_Y)\log\frac{1}{1-q_Y} + \underline{\theta}\log\frac{1}{\underline{\theta}} + (q_X-\hat{x}\underline{\theta})\log\frac{1}{q_X-\hat{x}\underline{\theta}}\\
&\geq H_b(q_Y)\\
&= H(Y).
\end{aligned}$$

So it suffices to consider the case $\hat{x} \in [0, \frac{q_X}{q_Y}]$ and $\theta = \overline{\theta}$, for which (P") is reduced to the following form:

$$\min_{\hat{x}\in[0,\frac{q_X}{q_Y}]} \eta(\hat{x}),$$

where

$$\eta(\hat{x}) \triangleq \begin{cases}
\frac{1-q_X}{1-\hat{x}}\log\frac{1-\hat{x}}{1-q_X} + \frac{q_X-\hat{x}}{1-\hat{x}}\log\frac{1-\hat{x}}{q_X-\hat{x}}, & \hat{x} \in [0, \frac{D+q_X-q_Y}{2(1-q_Y)}],\\
\frac{2\hat{x}-q_X+(1-2\hat{x})q_Y-D}{2\hat{x}}\log\frac{2\hat{x}}{2\hat{x}-q_X+(1-2\hat{x})q_Y-D} & \\
+\frac{D-(1-2\hat{x})(q_Y-q_X)}{2\hat{x}(1-\hat{x})}\log\frac{2\hat{x}(1-\hat{x})}{D-(1-2\hat{x})(q_Y-q_X)} & \\
+\frac{q_X+(1-2\hat{x})q_Y-D}{2(1-\hat{x})}\log\frac{2(1-\hat{x})}{q_X+(1-2\hat{x})q_Y-D}, & \hat{x} \in (\frac{D+q_X-q_Y}{2(1-q_Y)}, \frac{q_X+q_Y-D}{2q_Y}],\\
\frac{\hat{x}-q_X}{\hat{x}}\log\frac{\hat{x}}{\hat{x}-q_X} + \frac{q_X}{\hat{x}}\log\frac{\hat{x}}{q_X}, & \hat{x} \in (\frac{q_X+q_Y-D}{2q_Y}, \frac{q_X}{q_Y}].
\end{cases}$$

Note that $\eta(\hat{x})$ is a continuous function over $[0, \frac{q_X}{q_Y}]$. Moreover, $\eta(\hat{x})$ decreases monotonically from $H_b(q_X)$ to $H_b(\frac{D_{\max}^{(B)}-D}{2-q_X-q_Y-D})$ as $\hat{x}$ varies from 0 to $\frac{D+q_X-q_Y}{2(1-q_Y)}$. Since $\eta(\hat{x})$ is a concave function of $\frac{q_X}{\hat{x}}$ for $\hat{x} \in [\frac{q_X+q_Y-D}{2q_Y}, \frac{q_X}{q_Y}]$, it follows that

$$\min_{\hat{x}\in[\frac{q_X+q_Y-D}{2q_Y},\frac{q_X}{q_Y}]} \eta(\hat{x}) = \min\left\{\eta(\frac{q_X+q_Y-D}{2q_Y}), \eta(\frac{q_X}{q_Y})\right\} = \min\left\{H_b(\frac{2q_Xq_Y}{q_X+q_Y-D}), H_b(q_Y)\right\}.$$

So we have

$$\begin{aligned}
&\min_{\hat{x}\in[0,\frac{D+q_X-q_Y}{2(1-q_Y)}]\cup[\frac{q_X+q_Y-D}{2q_Y},\frac{q_X}{q_Y}]} \eta(\hat{x})\\
&= \min\left\{H_b(\frac{D_{\max}^{(B)}-D}{2-q_X-q_Y-D}), H_b(\frac{2q_Xq_Y}{q_X+q_Y-D}), H_b(q_Y)\right\}\\
&= \min\left\{H_b(\frac{D_{\max}^{(B)}-D}{2-q_X-q_Y-D}), H_b(\frac{2q_Xq_Y}{q_X+q_Y-D})\right\} \qquad (32)
\end{aligned}$$

$$
\begin{aligned}
&= \min\left\{\eta(\frac{D+q_X-q_Y}{2(1-q_Y)}), \eta(\frac{q_X+q_Y-D}{2q_Y})\right\} \\
&\geq \min_{\hat{x}\in[\frac{D+q_X-q_Y}{2(1-q_Y)}, \frac{q_X+q_Y-D}{2q_Y}]} \eta(\hat{x}),
\end{aligned}
$$

where (32) is due to the fact that $H_b(q_Y) \geq H_b(q_X) \geq H_b(\frac{D_{\max}^{(B)}-D}{2-q_X-q_Y-D})$ (with the first inequality being a consequence of $q_X \leq q_Y$ and $q_X + q_Y \leq 1$). So the problem boils down to solving

$$
\min_{\hat{x}\in[\frac{D+q_X-q_Y}{2(1-q_Y)}, \frac{q_X+q_Y-D}{2q_Y}]} \eta(\hat{x}).
$$

It can be verified that the minimum is attained at $\hat{x} = \frac{D+q_X-q_Y}{2(1-q_Y)}$. This completes the proof of (13). A graphical illustration of the entropy-constrained optimal transport plan for the binary case can be found in Figure 6.

On the tightness of the upper bound.

Note that

$$
\begin{aligned}
D_{\mathrm{mse}}(\mathcal{B}(q_X), R) &= \frac{1}{2} D_{\mathrm{ncr}}(\mathcal{B}(q_X), \mathcal{B}(q_X), R) \\
&= \begin{cases} -\frac{(1-q_X)^2}{1-H_b^{-1}(R)} + 1 - q_X, & q_X \leq \frac{1}{2}, \\ -\frac{q_X^2}{1-H_b^{-1}(R)} + q_X, & q_X > \frac{1}{2}, \end{cases} \quad R \in [0, H_b(q_X)], \\
D_{\mathrm{mse}}(\mathcal{B}(q_Y), R) &= \frac{1}{2} D_{\mathrm{ncr}}(\mathcal{B}(q_Y), \mathcal{B}(q_Y), R) \\
&= \begin{cases} -\frac{(1-q_Y)^2}{1-H_b^{-1}(R)} + 1 - q_Y, & q_Y \leq \frac{1}{2}, \\ -\frac{q_Y^2}{1-H_b^{-1}(R)} + q_Y, & q_Y > \frac{1}{2}, \end{cases} \quad R \in [0, H_b(q_Y)].
\end{aligned}
$$

Moreover, we have

$$
\begin{aligned}
&\begin{cases} p_{\hat{X}^*}(\frac{q_X - H_b^{-1}(R)}{1-H_b^{-1}(R)}) = 1 - H_b^{-1}(R), \\ p_{\hat{X}^*}(1) = H_b^{-1}(R), \end{cases} \quad R \in [0, H_b(q_X)], q_X \leq \frac{1}{2}, \\
&\begin{cases} p_{\hat{X}^*}(0) = H_b^{-1}(R), \\ p_{\hat{X}^*}(\frac{q_X}{1-H_b^{-1}(R)}) = 1 - H_b^{-1}(R), \end{cases} \quad R \in [0, H_b(q_X)], q_X \geq \frac{1}{2}, \\
&\begin{cases} p_{\hat{Y}^*}(\frac{q_Y - H_b^{-1}(R)}{1-H_b^{-1}(R)}) = 1 - H_b^{-1}(R), \\ p_{\hat{Y}^*}(1) = H_b^{-1}(R), \end{cases} \quad R \in [0, H_b(q_Y)], q_Y \leq \frac{1}{2}, \\
&\begin{cases} p_{\hat{Y}^*}(0) = H_b^{-1}(R), \\ p_{\hat{Y}^*}(\frac{q_Y}{1-H_b^{-1}(R)}) = 1 - H_b^{-1}(R), \end{cases} \quad R \in [0, H_b(q_Y)], q_Y \geq \frac{1}{2}.
\end{aligned}
$$

So

$$
\begin{aligned}
&W_2^2(p_{\hat{X}^*}, p_{\hat{Y}^*}) \\
&= \begin{cases} \frac{(q_X-q_Y)^2}{1-H_b^{-1}(R)}, & q_X, q_Y \leq \frac{1}{2} \text{ or } q_X, q_Y \geq \frac{1}{2}, \\ \frac{(q_Y-q_X+H_b^{-1}(R))^2}{1-H_b^{-1}(R)} + H_b^{-1}(R) - \frac{2q_Y(1-q_X)H_b^{-1}(R)}{(1-H_b^{-1}(R))^2}, & q_X < \frac{1}{2}, q_Y > \frac{1}{2}, \\ \frac{(q_X-q_Y+H_b^{-1}(R))^2}{1-H_b^{-1}(R)} + H_b^{-1}(R) - \frac{2q_X(1-q_Y)H_b^{-1}(R)}{(1-H_b^{-1}(R))^2}, & q_X > \frac{1}{2}, q_Y < \frac{1}{2}. \end{cases}
\end{aligned}
$$

Based on the above expressions, one can easily evaluate the upper bound $\overline{D}_{\mathrm{ncr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), R)$ and the lower bound $\underline{D}_{\mathrm{ncr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), R)$. In particular, it can be verified that when $q_X, q_Y \leq \frac{1}{2}$ or $q_X, q_Y \geq \frac{1}{2}$,

$$
\overline{D}_{\mathrm{ncr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), R) = D_{\mathrm{ncr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), R), \quad R \in [0, \min\{H_b(q_X), H_b(q_Y)\}],
$$

i.e., there is no penalty for using optimal quantizer and dequantizer in the conventional rate-distortion sense. This completes the proof of (14).

Connections with the conventional distortion-rate functions.

It is easy to verify that

$$\begin{aligned}
D_{\mathrm{cr}}(\mathcal{B}(q_X), R) &\triangleq \min_{q_Y \in [0,1]} D_{\mathrm{cr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), R) \\
&= \begin{cases} q_X \left(1 - \frac{R}{H_b(q_X)}\right), & R \in [0, H_b(q_X)], q_X \le \frac{1}{2}, \\ (1-q_X)\left(1 - \frac{R}{H_b(q_X)}\right), & R \in [0, H_b(q_X)], q_X > \frac{1}{2}, \end{cases} \\
D_{\mathrm{ncr}}(\mathcal{B}(q_X), R) &\triangleq \min_{q_Y \in [0,1]} D_{\mathrm{ncr}}(\mathcal{B}(q_X), \mathcal{B}(q_Y), R) \\
&= \begin{cases} q_X - H_b^{-1}(R), & R \in [0, H_b(q_X)], q_X \le \frac{1}{2}, \\ 1 - q_X - H_b^{-1}(R), & R \in [0, H_b(q_X)], q_X > \frac{1}{2}, \end{cases}
\end{aligned}$$

which are respectively the conventional one-shot distortion-distortion function (or equivalently, one-shot distortion-rate-perception function with an inactive perception constraint) for $\mathcal{B}(q_X)$ with and without common randomness. In general, $D_{\mathrm{ncr}}(\mathcal{B}(q_X), R)$ is different from $D_{\mathrm{mse}}(\mathcal{B}(q_X), R)$ (the former is strictly above the latter). The reason is as follows: even though the output distribution constraint is removed in the definition of $D_{\mathrm{ncr}}(\mathcal{B}(q_X), R)$, the output alphabet remains to be $\{0, 1\}$; in contrast, for $D_{\mathrm{mse}}(\mathcal{B}(q_X), R)$, the output alphabet is relaxed to $\mathbb{R}$.

A.5 Auxiliary Results

Lemma 1 (Finite Support Approximation)**.** For any $R \ge 0$ and $\epsilon > 0$, there exists a joint distribution $p_{X\hat{X}\hat{Y}Y}$ compatible with the given marginal distributions p_X and p_Y such that $X \leftrightarrow \hat{X} \leftrightarrow \hat{Y} \leftrightarrow Y$ form a Markov chain and $\hat{X}$ is a finite-support random variable with $H(\hat{X}) \le R$ (or $\hat{Y}$ is a finite-support random variable with $H(\hat{Y}) \le R$); moreover, $\mathbb{E}[X|\hat{X}] = \hat{X}$, $\mathbb{E}[Y|\hat{Y}] = \hat{Y}$, and

$$\mathbb{E}[\|X - \hat{X}\|^2] + \mathbb{E}[\|Y - \hat{Y}\|] + \mathbb{E}[\|\hat{X} - \hat{Y}\|^2] \le D_{\mathrm{ncr}}(p_X, p_Y, R) + \epsilon.$$

Proof. In light of Theorem 1, we can find $p_{X\hat{X}\hat{Y}Y}$ such that $X \leftrightarrow \hat{X} \leftrightarrow \hat{Y} \leftrightarrow Y$ form a Markov chain, $H(\hat{X}) \le R$, $\mathbb{E}[X|\hat{X}] = \hat{X}$, $\mathbb{E}[Y|\hat{Y}]$, and

$$\mathbb{E}[\|X - \hat{X}\|^2] + \mathbb{E}[\|Y - \hat{Y}\|] + \mathbb{E}[\|\hat{X} - \hat{Y}\|^2] \le D_{\mathrm{ncr}}(p_X, p_Y, R) + \frac{\epsilon}{2}. \tag{33}$$

The proof is complete if $\hat{X}$ is a finite-support random variable. So it suffices to consider the case where $\hat{X}$ takes value from some countably infinite set $\{\hat{x}_i\}_{i=1}^{\infty}$. Since $\mathbb{E}[\|X\|^2] < \infty$ and $\mathbb{E}[\|Y\|^2] < \infty$, it follows that there exists a positive integer N such that

$$\mathbb{P}\{\hat{X} \in \{\hat{x}_i\}_{i=N}^{\infty}\}\mathbb{E}[\|X\|^2 + \|Y\|^2 | \hat{X} \in \{\hat{x}_i\}_{i=N}^{\infty}] \le \frac{\epsilon}{4}.$$

Let $\hat{X}' \triangleq \hat{X}$ if $\hat{X} \in \{\hat{x}_i\}_{i=1}^{N-1}$ and $\hat{X}' \triangleq \mathbb{E}[X|\hat{X} \in \{\hat{x}_i\}_{i=N}^{\infty}]$ if $\hat{X} \in \{\hat{x}_i\}_{i=N}^{\infty}$. Note that

$$\begin{aligned}
&\mathbb{E}[\|X - \hat{X}'\|^2] + \mathbb{E}[\|\hat{X}' - \hat{Y}\|^2] \\
&= \mathbb{P}\{\hat{X} \in \{\hat{x}_i\}_{i=1}^{N-1}\}\mathbb{E}[\|X - \hat{X}'\|^2 + \|\hat{X}' - \hat{Y}\|^2 | \hat{X} \in \{\hat{x}_i\}_{i=1}^{N-1}] \\
&\quad + \mathbb{P}\{\hat{X} \in \{\hat{x}_i\}_{i=N}^{\infty}\}\mathbb{E}[\|X - \hat{X}'\|^2 + \|\hat{X}' - \hat{Y}\|^2 | \hat{X} \in \{\hat{x}_i\}_{i=N}^{\infty}] \\
&= \mathbb{P}\{\hat{X} \in \{\hat{x}_i\}_{i=1}^{N-1}\}\mathbb{E}[\|X - \hat{X}\|^2 + \|\hat{X} - \hat{Y}\|^2 | \hat{X} \in \{\hat{x}_i\}_{i=1}^{N-1}] \\
&\quad + \mathbb{P}\{\hat{X} \in \{\hat{x}_i\}_{i=N}^{\infty}\}\mathbb{E}[\|X - \hat{X}'\|^2 + \|\hat{X}' - \hat{Y}\|^2 | \hat{X} \in \{\hat{x}_i\}_{i=N}^{\infty}] \\
&\le \mathbb{P}\{\hat{X} \in \{\hat{x}_i\}_{i=1}^{N-1}\}\mathbb{E}[\|X - \hat{X}\|^2 + \|\hat{X} - \hat{Y}\|^2 | \hat{X} \in \{\hat{x}_i\}_{i=1}^{N-1}] \\
&\quad + \mathbb{P}\{\hat{X} \in \{\hat{x}_i\}_{i=N}^{\infty}\}\mathbb{E}[\|X - \hat{X}'\|^2 + 2\|\hat{X}'\|^2 + 2\|\hat{Y}\|^2 | \hat{X} \in \{\hat{x}_i\}_{i=N}^{\infty}]
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{P}\{\hat{X} \in \{\hat{x}_i\}_{i=1}^{N-1}\}\mathbb{E}[\|X-\hat{X}\|^2 + \|\hat{X}-\hat{Y}\|^2 | \hat{X} \in \{\hat{x}_i\}_{i=1}^{N-1}] \\
&\quad + \mathbb{P}\{\hat{X} \in \{\hat{x}_i\}_{i=N}^{\infty}\}\mathbb{E}[\|X\|^2 + \|\hat{X}'\|^2 + 2\|\hat{Y}\|^2 | \hat{X} \in \{\hat{x}_i\}_{i=N}^{\infty}] \\
&\leq \mathbb{E}[\|X-\hat{X}\|^2 + \|\hat{X}-\hat{Y}\|^2] + 2\mathbb{P}\{\hat{X} \in \{\hat{x}_i\}_{i=N}^{\infty}\}\mathbb{E}[\|X\|^2 + \|Y\|^2 | \hat{X} \in \{\hat{x}_i\}_{i=N}^{\infty}] \\
&\leq \mathbb{E}[\|X-\hat{X}\|^2] + \mathbb{E}[\|\hat{X}-\hat{Y}\|^2] + \frac{\epsilon}{2}. \qquad (34)
\end{aligned}
$$

Now define a new joint distribution $p_{X\hat{X}''\hat{Y}''Y}$ such that $p_{X\hat{X}''} = p_{X\hat{X}'}$, $p_{\hat{X}''\hat{Y}''} = p_{\hat{X}'\hat{Y}}$, $p_{Y\hat{Y}''} = p_{Y\hat{Y}}$, and $X \leftrightarrow \hat{X}'' \leftrightarrow \hat{Y}'' \leftrightarrow Y$ form a Markov chain. It is clear that $\hat{X}''$ is a finite-support random variable with $H(\hat{X}'') = H(\hat{X}') \leq H(\hat{X}) \leq R$, $\mathbb{E}[X|\hat{X}''] = \hat{X}''$, $\mathbb{E}[Y|\hat{Y}''] = \hat{Y}''$, and

$$
\begin{aligned}
&\mathbb{E}[\|X-\hat{X}''\|^2] + \mathbb{E}[\|Y-\hat{Y}''\|] + \mathbb{E}[\|\hat{X}''-\hat{Y}''\|^2] \\
&\leq \mathbb{E}[\|X-\hat{X}'\|^2] + \mathbb{E}[\|Y-\hat{Y}\|] + \mathbb{E}[\|\hat{X}'-\hat{Y}\|^2] \\
&\leq \mathbb{E}[\|X-\hat{X}\|^2] + \mathbb{E}[\|Y-\hat{Y}\|] + \mathbb{E}[\|\hat{X}-\hat{Y}\|^2] + \frac{\epsilon}{2} \qquad (35) \\
&\leq D_{\mathrm{ncr}}(p_X, p_Y, R) + \epsilon, \qquad (36)
\end{aligned}
$$

where (35) and (36) are due to (33) and (34), respectively. This completes the proof of Lemma 1. □

Lemma 2 (Deterministic on Finite Support)**.** Let $X \leftrightarrow \hat{X} \leftrightarrow \hat{Y} \leftrightarrow Y$ be a Markov chain with $\mathbb{E}[X|\hat{X}] = \hat{X}$, $\mathbb{E}[Y|\hat{Y}] = \hat{Y}$, and assume that $\hat{X}$ (or $\hat{Y}$) is a finite-support random variable. There exist deterministically related random variables $\hat{X}'$ and $\hat{Y}'$, with the support size no greater than that of $\hat{X}$ and $H(\hat{X}') = H(\hat{Y}') \leq H(\hat{X})$ (or $H(\hat{X}') = H(\hat{Y}') \leq H(\hat{Y})$), such that $X \leftrightarrow \hat{X}' \leftrightarrow \hat{Y}' \leftrightarrow Y$ form a Markov chain, $\mathbb{E}[X|\hat{X}'] = \hat{X}'$, $\mathbb{E}[Y|\hat{Y}'] = \hat{Y}'$, and

$$
\begin{aligned}
&\mathbb{E}[\|X-\hat{X}'\|^2] + \mathbb{E}[\|Y-\hat{Y}'\|] + \mathbb{E}[\|\hat{X}'-\hat{Y}'\|^2] \\
&= \mathbb{E}[\|X-\hat{X}\|^2] + \mathbb{E}[\|Y-\hat{Y}\|] + \mathbb{E}[\|\hat{X}-\hat{Y}\|^2].
\end{aligned}
$$

Proof. Let $\tilde{Y} \triangleq \mathbb{E}[Y|\hat{X}]$. Since $\tilde{Y} \leftrightarrow \hat{Y} \leftrightarrow Y$ form a Markov chain and $\mathbb{E}[Y|\hat{Y}] = \hat{Y}$, we have $\tilde{Y} = \mathbb{E}[\hat{Y}|\hat{X}]$. Construct a new joint distribution $p_{X\hat{X}'\hat{Y}'Y}$ such that $p_{X\hat{X}'} = p_{X\hat{X}}$, $p_{\hat{X}'\hat{Y}'} = p_{\hat{X}\tilde{Y}}$, $p_{Y\hat{Y}'} = p_{Y\tilde{Y}}$, and $X \leftrightarrow \hat{X}' \leftrightarrow \hat{Y}' \leftrightarrow Y$ form a Markov chain. It is clear that $\mathbb{E}[X|\hat{X}'] = \hat{X}'$, $\mathbb{E}[Y|\hat{Y}'] = \hat{Y}'$, and

$$
\begin{aligned}
&\mathbb{E}[\|X-\hat{X}'\|^2] + \mathbb{E}[\|Y-\hat{Y}'\|^2] + \mathbb{E}[\|\hat{X}'-\hat{Y}'\|^2] \\
&= \mathbb{E}[\|X-\hat{X}\|^2] + \mathbb{E}[\|Y-\tilde{Y}\|^2] + \mathbb{E}[\|\hat{X}-\tilde{Y}\|^2] \\
&= \mathbb{E}[\|X-\hat{X}\|^2] + \mathbb{E}[\|Y-\hat{Y}\|^2] + \mathbb{E}[\|\hat{Y}-\tilde{Y}\|^2] + \mathbb{E}[\|\hat{X}-\tilde{Y}\|^2] \\
&= \mathbb{E}[\|X-\hat{X}\|^2] + \mathbb{E}[\|Y-\hat{Y}\|^2] + \mathbb{E}[\|\hat{X}-\hat{Y}\|^2],
\end{aligned}
$$

where the last equality is due to $\tilde{Y} = \mathbb{E}[\hat{Y}|\hat{X}]$. If the function that maps $\hat{X}'$ to $\hat{Y}'$ (or equivalently, maps $\hat{X}$ to $\tilde{Y}$) is invertible, then $\hat{X}$, $\hat{X}'$, $\hat{Y}'$ have the same support size and $H(\hat{X}) = H(X') = H(Y')$, which completes the proof. Otherwise, the support size of $\hat{Y}'$ must be strictly smaller than that of $\hat{X}'$ (which is the same as that of $\hat{X}$) and $H(\hat{Y}') < H(\hat{X}') = H(\hat{X})$. We can alternately reduce the support sizes of $\hat{X}'$ and $\hat{Y}'$ using this argument until they are deterministically related (and consequently have the same support size and the same entropy). This can be accomplished in a finite number of steps because the reduction in support size cannot continue forever. □

□

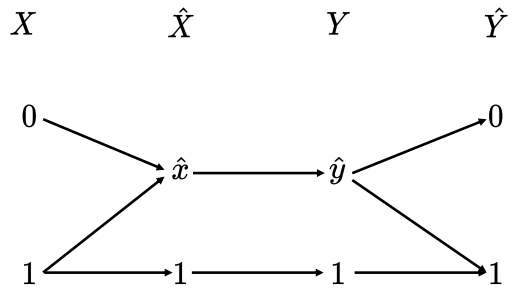


Figure 6: Illustration of the entropy-constrained optimal transport plan for the binary case (assuming $q_X + q_Y \leq 1$), where $p_{\hat{X}}(\hat{x}) = p_{\hat{Y}}(\hat{y}) = 1 - H_b^{-1}(R)$ with $\hat{x} = \frac{q_X - H_b^{-1}(R)}{1 - H_b^{-1}(R)}$ and $\hat{y} = \frac{q_Y - H_b^{-1}(R)}{1 - H_b^{-1}(R)}$. It is interesting to note that the quantizer $p_{\hat{X}|X}$ does not depend on p_Y while the dequantizer $p_{Y|\hat{Y}}$ does not depend on p_X. So they are decoupled in a certain sense. Moreover, $p_{\hat{X}|X}$ and $p_{Y|\hat{Y}}$ coincide respectively with optimal quantizer $p_{\hat{X}^*|X}$ and dequantizer $p_{Y|\hat{Y}^*}$ in the conventional rate-distortion sense when $q_X, q_Y \leq 1/2$.

B Experimental Results

B.1 Dataset

Image super-resolution is conducted on MNIST (LeCun et al., 1998). For synthesizing low-resolution images, we perform bilinear downsampling on the original image from 28×28 to 7×7. The samples in Figure 5(b) show that the low resolution digits are blurry and some of them are hard to recognize. Image denoising is conducted on SVHN (Netzer et al., 2011). In our experiments, we synthesize the noisy image with additive Gaussian noise. The standard deviation is set to 20.

B.2 Universal Quantization

Let $\mathcal{C}$ be our codebook for quantization. Recall that the encoder uses a $\tanh$ activation so its output lies in $(-1, 1)^d$. Given dimension d and L quantization levels per dimension as parameters, $\mathcal{C}$ will consist of L uniformly spaced intervals across all d dimensions. The upper bound of model rate is given by $d \log(L)$. With this codebook, universal quantization (Ziv, 1985; Gray & Stockham, 1993; Theis & Agustsson, 2021) is implemented as follows. We assume the sender and receiver have access to the same $u \sim U[-1/(L-1), +1/(L-1)]^d$. The sender computes

$$z = \arg\min_{c \in \mathcal{C}} \|f(x) + u - c\|$$

and sends z to the receiver. The receiver decodes the image by passing $z - u$ through the decoder. This is also known as a subtractive dither in literature (Gray & Stockham, 1993). For super-resolution and image denoising, the interval L is respectively fixed at 4 and 8.

B.3 Training Details

To induce distributional shift, we use the Wasserstein GAN for our experiments. We alternate between training the encoder/decoder f, g, and the critic h. By Kantorovich-Rubinstein duality (Villani, 2009), the critic is used to approximate

$$W_1(p_Y, p_{\tilde{Y}}) = \sup_{\|\nabla h\| \leq 1} \mathbb{E}[h(Y)] - \mathbb{E}[h(\tilde{Y})], \tag{37}$$

where $\tilde{Y} = g(Q(f(X) + U) - U)$ for U as in Appendix B.2. The Lipschitz constraint is implemented with a gradient penalty (Gulrajani et al., 2017) in practice.

Super-resolution. The training for end-to-end network lasts for 50 epochs. λ in (17) is fixed at $1e-3$ across all rates. The learning rate initialized to be 0.0001 and is decayed by a factor of

5 after 30 epochs. The Adam (Kingma & Ba, 2014) optimizer is used. Table 3 illustrates the detailed training setting. For the helper two-branch network at a specific rate constraint, we load the pre-trained encoder weight of the corresponding end-to-end network, as well as two randomly initialized decoders g_1, g_2. Note that only theese two decoders are trainable. During training, we use the Adam optimizer with the learning rate initialized at 0.0001. There are a total of 100 epochs until the convergence of the two decoders. The learning rate is decayed once at 50 epochs by a factor of 5. Detailed training settings are shown in Table 4.

Image Denoising. The experiments for image denoising share many settings with image super-resolution. Tables 3 and 4 can be reused to reproduce the experiments on image denoising. Here, we list the difference between them. For denoising, the end-to-end model is trained for 100 epochs with λ fixed at $3e-3$ across all rates. The learning rate is decayed by a factor of 5 after 40 epochs. The two-branch model is trained for total 200 epochs and we decay the learning rate at 100 epochs by a factor of 5.

Table 3: Hyperparameters used for training end-to-end model in Fig. 4(a). α is the learning rate, (β_1, β_2) are the parameters for Adam, and λ_{GP} is the gradient penalty coefficient.

	α	β_1	β_2	λ_{GP}
Encoder	10^{-4}	0.5	0.999	-
Decoder	10^{-4}	0.5	0.999	-
Critic	10^{-4}	0.5	0.999	10

Table 4: Hyperparameters used for training two-branch model in Fig. 4(b). α is the learning rate, (β_1, β_2) are the parameters for Adam, and λ_{GP} is the gradient penalty coefficient.

	α	β_1	β_2	λ_{GP}
Encoder	0	-	-	-
Decoder-1	10^{-4}	0.5	0.999	-
Decoder-2	10^{-4}	0.5	0.999	-
Critic	10^{-4}	0.5	0.999	10

B.4 Detailed Results in Figure. 5

In Fig. 5(a)(c), we have provided a comparison between the case with or without common randomness in the form of a scatter chart. Here, we present detailed quantities of each point in Fig. 5(a)(c). Table 5 shows the number of each dot for super-resolution experiments, and Table 6 present the value of each dot for image denoising. From the two tables, we can further see the utility of common randomness quantitatively.

Table 5: The detailed number of MSE distortion loss in Fig. 5(a).

Super-resolution **with** Common Randomness

Rate	4	6	8	10	12	14	16	18
MSE	0.0515	0.0457	0.0420	0.0394	0.0372	0.0353	0.0339	0.0324
Rate	20	22	24	26	28	30	32	-
MSE	0.0313	0.0300	0.0297	0.0285	0.0280	0.0277	0.0269	-

Super-resolution **without** Common Randomness

Rate	4	6	8	10	12	14	16	18
MSE	0.0558	0.0506	0.0463	0.0435	0.0415	0.0396	0.0383	0.0367
Rate	20	22	24	26	28	30	32	-
MSE	0.0351	0.0337	0.0331	0.0328	0.0315	0.0308	0.0300	-

Table 6: The detailed number of MSE distortion loss in Fig. 5(c).

Image Denoising **with** Common Randomness

Rate	12	18	24	30	36	42	48	54	60
MSE	0.0219	0.0195	0.0175	0.01634	0.0154	0.0147	0.0138	0.0135	0.0129
Rate	66	72	78	84	90	96	102	108	114
MSE	0.0126	0.0123	0.0118	0.0117	0.0115	0.0112	0.0109	0.0107	0.0106

Image Denoising **without** Common Randomness

Rate	12	18	24	30	36	42	48	54	60
MSE	0.0230	0.0208	0.0189	0.0175	0.0165	0.0157	0.0151	0.0145	0.014
Rate	66	72	78	84	90	96	102	108	114
MSE	0.0137	0.0133	0.0130	0.0127	0.0123	0.0120	0.0118	0.0116	0.0114

B.5 Comparison with Ground Truth

In order to illustrate the rate-distortion trade-offs, we report the MSE distortion that is measured between *degraded input images* and decoder outputs in Figure 5. Since the input and output distributions are different, we do not expect MSE $\rightarrow 0$ as the rate increases. The MSE distortion between degraded input and restored output is still able to reveal how much content information of the input is preserved in output (lower is better).

We now additionally show the MSE distortion between *ground truth* and decoder outputs in Tables 7 and 8. Concretely, we measure MSE distortion $\mathbb{E}[\|Y - \tilde{Y}\|^2]$, where Y is ground truth and $\tilde{Y}$ is the network output. Note that *for training, the ground truth is only used in an unsupervised fashion with unpaired noisy images*, and here the ground truth-noisy image pairs are only used for test time evaluation. Note also that the MSE distortion is correspondingly lower if common randomness is adopted.

Table 7: Illustration of MSE distortion between network outputs and ground truth for super-resolution.

Super-resolution **with** Common Randomness

Rate	4	8	12	16	20	24	28	32
MSE	0.0582	0.0464	0.0408	0.0380	0.0360	0.0353	0.0348	0.0343

Super-resolution **without** Common Randomness

Rate	4	8	12	16	20	24	28	32
MSE	0.0646	0.0492	0.0426	0.0390	0.0375	0.0362	0.0353	0.0345

Table 8: Illustration of MSE distortion between network outputs and ground truth for image denoising.

Image Denoising **with** Common Randomness

Rate	12	24	36	48	60	72	84	96	108
MSE	0.0157	0.0114	0.0092	0.0077	0.0069	0.0062	0.0056	0.0052	0.0047

Image Denoising **without** Common Randomness

Rate	12	24	36	48	60	72	84	96	108
MSE	0.0169	0.0128	0.0104	0.0090	0.0080	0.0072	0.0066	0.0060	0.0057

B.5.1 BREAKDOWN OF THE TABLE 1

It is worth reminding that the each number in Table 1 is the total distortion 20. Here, we provide a breakdown of total distortion in each of the three term for joint training, i.e, $\mathbb{E}[\|X - \tilde{Y}_1\|^2]$, $\mathbb{E}[\|\tilde{Y} - \tilde{Y}_2\|^2]$ and $\mathbb{E}[\|\tilde{Y}_1 - \tilde{Y}_2\|^2]$.

Table 9: Breakdown of the Table 1. At any rate, it can be observed that the total losses of our approximation system (L_2) are very close to that of end-to-end learning system under the setting without common randomness ($\mathbb{E}[\|X - \tilde{Y}\|^2]$).

Image Super-resolution Using Joint Training

Rate	$\mathbb{E}[\|X - \tilde{Y}_1\|^2]$	$\mathbb{E}[\|\tilde{Y} - \tilde{Y}_2\|^2]$	$\mathbb{E}[\|\tilde{Y}_1 - \tilde{Y}_2\|^2]$	L_2	$\mathbb{E}[\|X - \tilde{Y}\|^2]$
4	0.0355	0.0227	0.0004	**0.0586**	**0.0558**
10	0.0223	0.0222	0.0008	**0.0453**	**0.0435**
20	0.0136	0.0191	0.00013	**0.0349**	**0.0351**
30	0.00122	0.0172	0.0015	**0.0309**	**0.0308**

Image Denoising Using Joint Training

Rate	$\mathbb{E}[\|X - \tilde{Y}_1\|^2]$	$\mathbb{E}[\|\tilde{Y} - \tilde{Y}_2\|^2]$	$\mathbb{E}[\|\tilde{Y}_1 - \tilde{Y}_2\|^2]$	L_2	$\mathbb{E}[\|X - \tilde{Y}\|^2]$
12	0.01911	0.00497	0.00020	**0.02428**	**0.02302**
30	0.01458	0.00435	0.00026	**0.01919**	**0.01746**
60	0.01168	0.00378	0.00032	**0.01578**	**0.01401**
90	0.01035	0.00323	0.00028	**0.01386**	**0.01229**

B.6 NETWORK ARCHITECTURE

Super-resolution. The detailed network structure for end-to-end model and two-branch model are respectively presented in Table 10 and Table 11. The last linear layer of the encoder controls the number of output symbols.

Denoising. The detailed network structure for end-to-end model and two-branch model are respectively presented in Table 12 and Table 13. The last linear layer of the encoder controls the number of output symbols.

Table 10: Model architectures of end-to-end network used in super-resolution.

Encoder	Decoder	Critic
Input	Input	Input
Conv2D, l-ReLU	Linear, BatchNorm1D, l-ReLU	Conv2D, l-ReLU
Conv2D, l-ReLU	Linear, BatchNorm1D, l-ReLU	Conv2D, l-ReLU
Flatten	Unflatten	Conv2D, l-ReLU
Linear, l-ReLU	ConvT2D, BatchNorm2D, l-ReLU	Linear
Linear, l-ReLU	ConvT2D, BatchNorm2D, l-ReLU	
Linear, Tanh	ConvT2D, BatchNorm2D, Sigmoid	
Quantizer		

Table 11: Model architectures of two-branch network used in super-resolution.

Encoder	Decoder1 and 2
Input	Input
Conv2D, l-ReLU	Linear, BatchNorm1D, l-ReLU
Conv2D, l-ReLU	Linear, BatchNorm1D, l-ReLU
Flatten	Unflatten
Linear, l-ReLU	ConvT2D, BatchNorm2D, l-ReLU
Linear, l-ReLU	ConvT2D, BatchNorm2D, l-ReLU
Linear, Tanh	ConvT2D, Sigmoid
Quantizer	

Table 12: Model architectures of end-to-end network used in image denoising.

Encoder	Decoder	Critic
Input	Input	Input
Conv2D, l-ReLU	Linear, BatchNorm1D, l-ReLU	Conv2D, l-ReLU
Conv2D, l-ReLU	Linear, BatchNorm1D, l-ReLU	Conv2D, l-ReLU
Conv2D, l-ReLU	Unflatten	Conv2D, l-ReLU
Flatten	ConvT2D, BatchNorm2D, l-ReLU	Linear
Linear, Tanh	ConvT2D, BatchNorm2D, l-ReLU	
Quantizer	ConvT2D, BatchNorm2D, l-ReLU	
	ConvT2D, BatchNorm2D, Sigmoid	

Table 13: Model architectures of two-branch network used in image denoising. ResBlock is formed using two Conv2D and skip connection.

Encoder	Decoder1 and 2
Input	Input
Conv2D, l-ReLU	Linear, BatchNorm1D, l-ReLU
Conv2D, l-ReLU	Linear, BatchNorm1D, l-ReLU
Conv2D, l-ReLU	Unflatten
Flatten	ConvT2D, l-ReLU
Linear, Tanh	ResBlock, ConvT2D, l-ReLU
Quantizer	ResBlock, ConvT2D, l-ReLU
	ResBlock, ConvT2D, Sigmoid