Rate-Distortion-Perception Tradeoff Based on the Conditional Perception Measure

Sadaf Salehkalaibar, Buu Phan, Ashish Khisti and Wei Yu
Electrical and Computer Engineering Department
University of Toronto
Toronto, Canada
e-mail: sadafs@ece.utoronto.ca, troung.phan@mail.utoronto.ca, {akhisti,weiyu}@ece.utoronto.ca

***Abstract*—In this paper, we study the rate-distortion-perception tradeoff generalizing the classical rate-distortion theory by adding a perception constraint to generate visually pleasing reconstructions. The perception metric measures the divergence between the distributions of the input and the reconstruction when both distributions are conditioned on the encoder's output. This metric, originally introduced by Mentzer et al. for the video compression setting, is called as *conditional perception measure*. We characterize the rate-distortion-perception tradeoff for a general source. In the Gaussian setting, we show that jointly Gaussian reconstructions are indeed optimal. Interestingly, to achieve a perceptually perfect reconstruction, comparing to the minimum mean square error (MMSE) reconstruction, we only need extra 0.5 bits/sample for the compression rate.**

I. Introduction

The classical rate-distortion theory [1] aims at optimizing the compression rate of a system subject to a distortion constraint on the reconstruction. However, through experiments, it has been observed that satisfying a low distortion does not necessarily lead to visually pleasing reconstructions. To address this problem, a perception metric has been introduced that generalizes the classical structure to a new paradigm called as *rate-distortion-perception* framework [2]. Unlike distortion, the perception metric does not involve a reference.

The perceptual quality of reconstructions has been considered in image and video compression settings in several previous works [3]–[8]. It has been observed that satisfying a high perceptual quality comes at the cost of increased distortion. For example, in the image compression setting, the reconstruction corresponding to the optimal solution of the classical rate-distortion theory (without perception constraint) may represent a blurry image where the edges are not sharp. This is because the distortion measure is typically the mean square error (MSE), but a minimum MSE (MMSE) reconstruction does not directly maximize the perception quality. In a previous work [9], it has been shown that for a given encoder, by slightly increasing the distortion (by no more than a factor of two), a reconstruction with sharper edges can be obtained by adding a perception criterion.

The perception measure considered in [2] and [9] is defined to be the divergence between the marginal distributions of the input and reconstruction. Recently, a different perception measure has been considered for the video compression setting [10], [11]. This measure corresponds to the divergence between the conditional distributions of the input and reconstruction conditioned on the output of the encoder. Through experiments, the conditional perception measure has been shown to correspond to a significant improvement in terms of achieving a higher perceptual quality.

In this work, we study the conditional perception measure of [10] from a theoretical perspective. In the image compression setting, if the output of the encoder is the MMSE representation, the conditional perception measure has an interesting interpretation. That is, conditioning on the MMSE representation helps the decoder improve a possibly blurry reconstruction by adjusting the fine details of the output image to match the conditional distributions. This would enhance the perceptual quality of the reconstructed images, because conditioning on the MMSE representation constrains the decoder not to deviate too much from the general content (even blurry) of the image.

For a general source, this paper characterizes the rate-distortion-perception (RDP) tradeoff for the conditional perception measure. We further evaluate the RDP characterization for a Gaussian source and show that the jointly Gaussian distribution is indeed optimal for reconstruction. Interestingly, the optimal operation that transforms the MMSE reconstruction to another one that satisfies the perfect perceptual constraint is to introduce an artificial noise. This is different from the optimal transform of [9] for the marginal-based perception measure where it is shown that scaling converts the MMSE reconstruction to the one with perfect perceptual quality.

For the case of perfect perceptual quality, the RD tradeoff has an interesting interpretation. It shows that in order to achieve a reconstruction with a perfect perceptual quality, as compared to the MMSE reconstruction, we only need an extra 0.5 bits/sample in the compression rate. For the fixed encoder setting, we show that similar to [9], the distortion of perceptually perfect reconstructions is at most two times the minimum distortion for the conditional perception measure.

Remarks: We follow the notation of [12]. For the distortion measure, we use the MSE loss and for the perception metric, we use the Wasserstein-2 distance where for two given probability distributions P_X and P_Y over (resp.) $\mathcal{X}$ and $\mathcal{Y}$, it is defined as follows

$$W_2^2(P_X, P_Y) := \inf \mathbb{E}[\|X - Y\|^2], \tag{1}$$

where the infimum is taken over all joint probability distributions with the marginals P_X and P_Y.

II. System Model

Assume that we have a sequence of source observations denoted by X^n which are independently and identically distributed (i.i.d.) according to a given distribution P_X. It is compressed into a message $M \in \mathcal{M}$ by an encoding function defined as follows

$$f : \mathcal{X}^n \to \mathcal{M}, \tag{2}$$

where $M = f(X^n)$. The decoder then generates a reconstruction $\hat{X}^n$ using a possibly stochastic function g given as follows

$$g \colon \mathcal{M} \to \hat{\mathcal{X}}^n, \tag{3}$$

where $\hat{X}^n = g(M)$. Define the following distribution based on the encoding and decoding functions

$$\begin{aligned} &P_{MX^n\hat{X}^n}(m, x^n, \hat{x}^n) := \\ &\qquad P_{X^n}(x^n)\mathbb{1}\{m = f(x^n)\}\mathbb{1}\{\hat{x}^n = g(m)\}, \\ &\qquad\qquad m \in \mathcal{M}, x^n \in \mathcal{X}^n, \hat{x}^n \in \hat{\mathcal{X}}^n. \end{aligned} \tag{4}$$

and let $\ell(M)$ denote the length of the message M.

Definition 1: For a given (D, P) pair, a rate R is said to be achievable if there exist encoding and decoding functions such that

$$\frac{1}{n}\mathbb{E}[\ell(M)] \leq R, \tag{5}$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(X_i - \hat{X}_i)^2] \leq D, \tag{6}$$

$$\max_m W_2^2(P_{\hat{X}^n|M=m}, P_{X^n|M=m}) \leq P. \tag{7}$$

Notice that the perception constraint in (7) conditions on the encoder's output. For the case of $P = 0$, this constraint simplifies to the following

$$P_{\hat{X}_i|M} = P_{X_i|M}, \qquad i = 1, \ldots, n, \tag{8}$$

which is the conditional perception measure studied in [10], [11].

The function $R(D, P)$ denotes the infimum of all achievable rates R and is called the *Rate-Distortion-Perception (RDP)* function for conditional perception measure.

III. Main Results

A. Asymptotic Setting

In this section, we first establish the RDP function for a general source. Then, we specialize the setting to a Gaussian source and show that the jointly Gaussian reconstruction is indeed optimal. For the output to satisfy the perfect perceptual quality ($P = 0$), we discuss that we only need an extra 0.5 bits/sample in compression rate R, as compared to the MMSE reconstruction.

The following theorem provides the RDP function for a general source.

Theorem 1: Consider a discrete memoryless source $X \in \mathcal{X}$ distributed according to P_X. For a given (D, P) pair, the RDP function is given by the following

$$\begin{aligned} R(D, P) = &\min_{P_{U|X}} I(U; X), \\ \text{s.t.}: \ &\exists \hat{X} \colon X \to U \to \hat{X}, \\ &\max_u W_2^2(P_{\hat{X}|U=u}, P_{X|U=u}) \leq P, \\ &\mathbb{E}[(X - \hat{X})^2] \leq D. \end{aligned} \tag{9}$$

Proof: See Section IV-A. ■

Now, assume that the source is Gaussian, i.e., $X \sim \mathcal{N}(0, \sigma^2)$ for some positive σ^2. The RDP function of the Gaussian source for the conditional perception measure is given in the following theorem.

Theorem 2: Suppose that X is a Gaussian source with zero-mean and variance σ^2. We have

$$R(D, P) = \begin{cases} \frac{1}{2}\log\frac{2\sigma^2}{D + \sqrt{P(2D - P)}} & P \leq D, \\ \max\{0, \frac{1}{2}\log\frac{\sigma^2}{D}\} & D > P. \end{cases} \tag{10}$$

Proof: See Section IV-B. ■

In the following, we discuss the RDP function of Theorem 2. For achieving a perfect perceptual quality where $P = 0$, Theorem 2 suggests that the required compression rate, denoted by $R^0(D)$, is given by the following

$$R^0(D) := R(D, 0) = \frac{1}{2}\log\frac{2\sigma^2}{D}. \tag{11}$$

Moreover, when P is large enough, i.e., $P = D$, the perception constraint is not active and Theorem 2 reduces to the classical rate-distortion tradeoff [1] where the MMSE reconstruction is optimal and the rate, denoted by $R^\infty(D)$, is given by the following

$$R^\infty(D) := R(D, D) = \frac{1}{2}\log\frac{\sigma^2}{D}. \tag{12}$$

Notice that

$$R^0(D) - R^\infty(D) = 0.5, \tag{13}$$

which implies that for achieving a perfect perceptual quality, we just need 0.5 extra bits/sample comparing to the rate of MMSE solution.

Furthermore, for a given rate R, if we denote the MMSE reconstruction by $\hat{X}^{\text{MMSE}}$, we can write [1],

$$X = \hat{X}^{\text{MMSE}} + Z^{\text{MMSE}}, \tag{14}$$

where $Z^{\text{MMSE}} \sim \mathcal{N}(0, \sigma^2 2^{-2R})$ is independent of $\hat{X}^{\text{MMSE}}$ and the minimum distortion is given by

$$D_{\min} := \sigma^2 2^{-2R}. \tag{15}$$

On the other side, it will be shown in the proof of achievability at Section IV-A that the reconstruction which satisfies the perfect perceptual quality, denoted by $\hat{X}^0$, can be written as follows

$$\hat{X}^0 = \hat{X}^{\text{MMSE}} + Z^0, \tag{16}$$

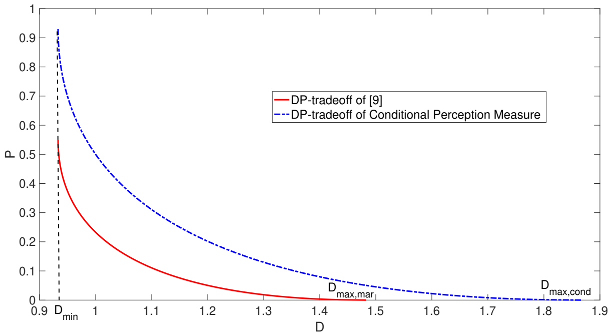

Fig. 1. Distortion-Perception (DP) Tradeoff for a fixed rate $R = 0.05$ bits/sample.

where $Z^0 \sim \mathcal{N}(0, \sigma^2 2^{-2R})$ is independent of $\hat{X}^{\text{MMSE}}$ and the corresponding distortion is given by

$$D_{\text{max,cond}} := 2\sigma^2 2^{-2R}. \tag{17}$$

From (16), we observe that $\hat{X}^0$ can be obtained from $\hat{X}^{\text{MMSE}}$ by simply adding a Gaussian noise of variance $\sigma^2 2^{-2R}$. This transform is different from the one (scaling) proposed in [9] for the marginal-based perception metric as will be discussed in the following.

Next, we compare the RDP function of Theorem 2 with that of [9] where the perception constraint of (7) is replaced by the following

$$W_2^2(P_{\hat{X}^n}, P_{X^n}) \le P, \tag{18}$$

which measures the distance between the marginal distributions of $P_{\hat{X}^n}$ and P_{X^n}. Theorem 1 in [9] states that the RDP function corresponding to the perception measure in (18), denoted by $R^{\text{mar}}(D,P)$, is given by the following

$$R^{\text{mar}}(D,P) = \begin{cases} \frac{1}{2}\log \frac{\sigma^2}{\sigma^2 - \left(\frac{\sigma^2+(\sigma-\sqrt{P})^2-D}{2(\sigma-\sqrt{P})}\right)^2}; & \sqrt{P} \le \sigma - \sqrt{|\sigma^2 - D|}, \\ \max\{0, \frac{1}{2}\log\frac{\sigma^2}{D}\}; & \sqrt{P} > \sigma - \sqrt{|\sigma^2 - D|}. \end{cases} \tag{19}$$

For a fixed compression rate $R = 0.05$, the tradeoff between distortion and perception is shown in Fig. 1. The dashed curve represents the DP-tradeoff for the conditional perception measure while the other curve shows the tradeoff of the marginal-based perception measure which is characterized in [9] and given in (19). For large enough P, both curves yield the same distortion $D_{\min}$ which corresponds to the MMSE reconstruction. Further, we can show that

$$D_{\text{max,mar}} \le D_{\text{max,cond}} \le 2D_{\text{max,mar}}, \tag{20}$$

which means that $D_{\text{max,cond}}$ is slightly larger than $D_{\text{max,mar}}$. However, the conditional perception measure is a more restricted constraint comparing to the marginal-based perception metric. So, the former is more capable of correcting errors in the reconstruction which yields a higher perceptual quality.

Notice that the setting which was discussed in this section is asymptotic where n symbols are encoded and the analysis is studied as $n \to \infty$. In the following section, we discuss the one-shot setting where one symbol is encoded at a time.

B. One-Shot Setting

The one-shot setup is introduced in the following definition. We assume that a shared key $K \in \mathcal{K}$ is available between the encoder and decoder.

Definition 2: Consider the following encoding and decoding functions

$$f^{\text{1-shot}}: \mathcal{X} \times \mathcal{K} \to \mathcal{M}, \tag{21}$$

$$g^{\text{1-shot}}: \mathcal{M} \times \mathcal{K} \to \hat{\mathcal{X}}, \tag{22}$$

where the message $M \in \mathcal{M}$ is encoded to X by the function $X = f^{\text{1-shot}}(M,K)$ and is decoded to $\hat{X}$ by the function $\hat{X} = g^{\text{1-shot}}(M,K)$. For a given (D,P) pair, we say that rate R is achievable if there exist encoding and decoding functions such that

$$\mathbb{E}[\ell(M)] \le R, \tag{23}$$

$$\mathbb{E}[(X - \hat{X})^2] \le D, \tag{24}$$

$$\max_m W_2^2(P_{\hat{X}|M=m}, P_{X|M=m}) \le P. \tag{25}$$

The infimum of all achievable rates is denoted by $R^{\text{1-shot}}(D,P)$ and is called as the *One-Shot RDP Function.*

Using the strong functional representation lemma [13], we can show that

$$R(D,P) \le R^{\text{1-shot}}(D,P) \le R(D,P) + \log(R(D,P)+1) + 5. \tag{26}$$

In the following, we fix the encoder and let the decoder adapt to different points on the DP-tradeoff curve. Furthermore, we consider the perfect perceptual quality where $P = 0$. To that end, we introduce the universal representation in the following.

Definition 3: Let U_r be a representation of X generated by a random transform $P_{U_r|X}$. The distortion set $\Phi_{D^0}(P_{U_r|X})$ denotes the set of all distortions D such that there exists $P_{\hat{X}|U_r X}$ where $X \to U_r \to \hat{X}$ forms a Markov chain and

$$\mathbb{E}[\|X - \hat{X}\|^2] \le D, \tag{27}$$

$$P_{U_r X} = P_{U_r \hat{X}}. \tag{28}$$

Let the MMSE reconstruction be as follows

$$\tilde{X} := \mathbb{E}[X|U_r]. \tag{29}$$

The following theorem states that the distortion of reconstruction with perfect perceptual quality is two times the minimum distortion.

Theorem 3: The set $\Phi_{D^0}(P_{U_r|X})$ is characterized as follows

$$\Phi_{D^0}(P_{U_r|X}) = \{D \colon D \ge 2\mathbb{E}[\|X - \tilde{X}\|^2]\}. \tag{30}$$

Proof: See Section IV-C. ■

The above result shows that $D_{\text{max,cond}}$ shown in Fig. 1 is $2D_{\min}$.

IV. Proofs

A. Proof of Theorem 1

Achievability: Fix a conditional distribution $P_{U|X}$ that attains (9). Thus, according to constraints of optimization (9), there exists a conditional distribution $P_{\hat{X}|XU}$ such that $X \to U \to \hat{X}$ and

$$\max_u W_2^2(P_{\hat{X}|U=u}, P_{X|U=u}) \leq P, \tag{31}$$

$$\mathbb{E}[(X-\hat{X})^2] \leq D. \tag{32}$$

Codebook Generation: Generate 2^{nR} codewords $u^n(m), m \in [1 : 2^{nR}]$ where each sample is drawn i.i.d. according to distribution P_U.

Encoding: The encoder finds a codeword $u^n(m)$ jointly typical with x^n and sends the index m to the decoder.

Decoding: Upon reception of the message m, the decoder first finds the codeword $u^n(m)$. It then generates $\hat{x}^n$ such that each sample is drawn i.i.d. according to the conditional distribution $P_{\hat{X}|U}$.

Analysis: From the covering lemma [1], the encoding is successful if

$$R \geq I(X;U). \tag{33}$$

The analyses of distortion and perception constraints easily follow from (57) and (56) and the fact that the codeword $u^n(m)$ and sequence $\hat{x}^n$ are generated in an i.i.d. manner.

Converse: First, we state the following lemma which will be used later in the proof.

Lemma 1: For given distributions $P_{X_1X_2}$ and $P_{Y_1Y_2}$ over (resp) $\mathcal{X}_1 \times \mathcal{X}_2$ and $\mathcal{Y}_1 \times \mathcal{Y}_2$, we have

$$W_2^2(P_{X_1X_2}, P_{Y_1Y_2}) \geq W_2^2(P_{X_1}, P_{Y_1}) + W_2^2(P_{X_2}, P_{Y_2}). \tag{34}$$

Proof: Consider the following set of inequalities:

$$\begin{aligned}
&W_2^2(P_{X_1X_2}, P_{Y_1Y_2})\\
&= \inf_{\substack{\tilde{P}_{X_1X_2Y_1Y_2}:\\ \tilde{P}_{X_1X_2}=P_{X_1X_2}\\ \tilde{P}_{Y_1Y_2}=P_{Y_1Y_2}}} \mathbb{E}[\|X_1-Y_1\|^2] + \mathbb{E}[\|X_2-Y_2\|^2] && (35)\\
&\geq \inf_{\substack{\tilde{P}_{X_1X_2Y_1Y_2}:\\ \tilde{P}_{X_1X_2}=P_{X_1X_2}\\ \tilde{P}_{Y_1Y_2}=P_{Y_1Y_2}}} \mathbb{E}[\|X_1-Y_1\|^2] + \inf_{\substack{\tilde{P}_{X_1X_2Y_1Y_2}:\\ \tilde{P}_{X_1X_2}=P_{X_1X_2}\\ \tilde{P}_{Y_1Y_2}=P_{Y_1Y_2}}} \mathbb{E}[\|X_2-Y_2\|^2] && (36)\\
&= \inf_{\substack{\tilde{P}_{X_1Y_1}:\\ \tilde{P}_{X_1}=P_{X_1}\\ \tilde{P}_{Y_1}=P_{Y_1}}} \mathbb{E}[\|X_1-Y_1\|^2] + \inf_{\substack{\tilde{P}_{X_2Y_2}:\\ \tilde{P}_{X_2}=P_{X_2}\\ \tilde{P}_{Y_2}=P_{Y_2}}} \mathbb{E}[\|X_2-Y_2\|^2] && (37)\\
&= W_2^2(P_{X_1}, P_{Y_1}) + W_2^2(P_{X_2}, P_{Y_2}), && (38)
\end{aligned}$$

where (36) follows because the infimum of sums is larger than (or equal to) the sum of infimums. ■

Now, we continue with the proof of the converse. Define

$$U_i := M. \tag{39}$$

Consider the rate constraint as follows

$$\begin{aligned}
nR \geq E[\ell(M)] &\geq H(M) && (40)\\
&\geq I(M;X^n) && (41)\\
&= \sum_{i=1}^n I(M;X_i|X^{i-1}) && (42)\\
&= \sum_{i=1}^n I(M, X^{i-1};X_i) && (43)\\
&\geq \sum_{i=1}^n I(M;X_i) && (44)\\
&= \sum_{i=1}^n I(U_i;X_i), && (45)
\end{aligned}$$

where (43) follows because the source is memoryless. Next, consider the perception constraint for each $m \in \mathcal{M}$ as the following

$$\begin{aligned}
nP &\geq W_2^2(P_{\hat{X}^n|M=m}, P_{X^n|M=m})\\
&\geq \sum_{i=1}^n W_2^2(P_{\hat{X}_i|M=m}, P_{X_i|M=m}) && (46)\\
&= \sum_{i=1}^n W_2^2(P_{\hat{X}_i|U_i=u}, P_{X_i|U_i=u}), \qquad u \in \mathcal{U} && (47)
\end{aligned}$$

where (46) follows from Lemma 1.

Now, define a time-sharing random variable $Q \in \{1, \dots, n\}$ independent of (X^n, U^n) and notice that

$$\begin{aligned}
&I(U_Q, Q; X_Q) = \frac{1}{n}\sum_{i=1}^n I(U_i;X_i) \leq R, && (48)\\
&W_2^2(P_{\hat{X}_Q|U_Q=u}, P_{X_Q|U_Q=u})\\
&\qquad = W_2^2(\frac{1}{n}\sum_{i=1}^n P_{\hat{X}_i|U_i=u}, \frac{1}{n}\sum_{i=1}^n P_{X_i|U_i=u}) && (49)\\
&\qquad \leq \frac{1}{n}\sum_{i=1}^n W_2^2(P_{\hat{X}_i|U_i=u}, P_{X_i|U_i=u}) && (50)\\
&\qquad \leq P, && (51)\\
&\mathbb{E}[(X_Q-\hat{X}_Q)^2] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(X_i-\hat{X}_i)^2] \leq D, && (52)
\end{aligned}$$

where (50) follows from the convexity of Wasserstein-2 distance [14, Proposition 3.1.6]. Now, let $U = (U_Q, Q)$, $X = X_Q$ and $\hat{X} = \hat{X}_Q$. This concludes the proof.

B. Proof of Theorem 2

Achievability: We choose the auxiliary random variable U such that

$$X = U + Z, \tag{53}$$

where $Z \sim \mathcal{N}(0, \nu^2)$ is independent from U for some positive ν^2. Now, let $\hat{X}$ be as follows

$$\hat{X} = U + \hat{Z}, \tag{54}$$

where $\hat{Z} \sim \mathcal{N}(0, \hat{\nu}^2)$ is independent from U for some positive $\hat{\nu}^2$. The mutual information term is simplified as follows

$$I(U;X) = \frac{1}{2} \log \frac{\sigma^2}{\nu^2}. \tag{55}$$

The distortion term can be written as the following

$$\mathbb{E}[\|X - \hat{X}\|^2] = \nu^2 + \hat{\nu}^2 \leq D. \tag{56}$$

Finally, the perception constraint is given as follows

$$W_2^2(P_{\hat{X}|U=u}, P_{X|U=u}) = (\hat{\nu} - \nu)^2 \leq P, \qquad \forall u \in \mathcal{U}. \tag{57}$$

Considering (55), (56) and (57) together yields the following achievable rate

$$R \geq \frac{1}{2} \log \frac{2\sigma^2}{D + \sqrt{P(2D - P)}}. \tag{58}$$

Converse: Let $P_{U^*|X}$ be the optimal solution of (9). So, there exists $P_{\hat{X}^*|U^*}$ such that

$$W_2^2(P_{\hat{X}^*|U^*=u}, P_{X|U^*=u}) \leq P, \qquad \forall u \in \mathcal{U}^*, \tag{59}$$

and

$$\mathbb{E}[\|X - \hat{X}^*\|^2] \leq D. \tag{60}$$

Then, define the following auxiliary random variable

$$\hat{X}(U^*) := \mathbb{E}[X|U^*]. \tag{61}$$

Notice that $X \to U^* \to \hat{X}(U^*)$ forms a Markov chain, so we have

$$I(X; \hat{X}(U^*)) \leq I(X; U^*) \leq R. \tag{62}$$

Next, consider the distortion constraint as follows

$$\begin{aligned} D &\geq \mathbb{E}[\|X - \hat{X}^*\|^2] && (63)\\ &= \mathbb{E}[\|X - \hat{X}(U^*) + \hat{X}(U^*) - \hat{X}^*\|^2] && (64)\\ &= \mathbb{E}[\|X - \hat{X}(U^*)\|^2] + \mathbb{E}[\|\hat{X}(U^*) - \hat{X}^*\|^2], && (65) \end{aligned}$$

where (65) follows because

- $X - \hat{X}(U^*)$ is the MSE error, so it is uncorrelated with the data which is U^*;
- $\hat{X}(U^*)$ is a function of U^*;
- $\hat{X}^*$ is a stochastic function of U^* so that it can be written as $\hat{X}^* = h(U^*, Z^*)$ where Z^* is independent of U^* and $X - \hat{X}(U^*)$;
- Thus, given the above three facts, $X - \hat{X}(U^*)$ and $\hat{X}(U^*) - \hat{X}^*$ are uncorrelated;

Now, consider the perception constraint for each $u \in \mathcal{U}$ as follows

$$\begin{aligned} P &\geq W_2^2(P_{\hat{X}^*|U^*=u}, P_{X|U^*=u}) \\ &= \inf_{\substack{\tilde{P}_{\hat{X}^* X|U^*=u}: \\ \tilde{P}_{\hat{X}^*|U^*=u} = P_{\hat{X}^*|U^*=u} \\ \tilde{P}_{X|U^*=u} = P_{X|U^*=u}}} \mathbb{E}_{\tilde{P}}[\|\hat{X}^* - X\|^2 | U^* = u] && (66)\\ &= \inf_{\substack{\tilde{P}_{(\hat{X}^* - \hat{X}(u))(X - \hat{X}(u))|U^*=u}: \\ \tilde{P}_{\hat{X}^* - \hat{X}(u)|U^*=u} = P_{\hat{X}^* - \hat{X}(u)|U^*=u} \\ \tilde{P}_{X - \hat{X}(u)|U^*=u} = P_{X - \hat{X}(u)|U^*=u}}} \mathbb{E}_{\tilde{P}}[\|(\hat{X}^* - \hat{X}(u)) - (X - \hat{X}(u))\|^2 | U^* = u] && (67)\\ &\geq \Big(\sqrt{\mathbb{E}[\|X - \hat{X}(u)\|^2 | U^* = u]} \\ &\qquad - \sqrt{\mathbb{E}[\|\hat{X}^* - \hat{X}(u)\|^2 | U^* = u]}\Big)^2, && (68) \end{aligned}$$

where

- (67) follows because conditioned on $U^* = u$, the solution of the optimization program does change by the following change of the variables $\hat{X}^* \to (\hat{X}^* - \hat{X}(u))$ and $X \to (X - \hat{X}(u))$ since $\hat{X}(u)$ is a function of u;
- (68) follows from Cauchy-Schwarz's inequality where for each distribution $\tilde{P}$, we have

$$\begin{aligned} &\mathbb{E}_{\tilde{P}}[(\hat{X}^* - \hat{X}(u))(X - \hat{X}(u)) | U^* = u] \\ &\quad \leq \sqrt{\mathbb{E}_{\tilde{P}}[\|\hat{X}^* - \hat{X}(u)\|^2 | U^* = u]} \cdot \\ &\qquad \sqrt{\mathbb{E}_{\tilde{P}}[\|X - \hat{X}(u)\|^2 | U^* = u]} && (69)\\ &\quad = \sqrt{\mathbb{E}[\|\hat{X}^* - \hat{X}(u)\|^2 | U^* = u]} \cdot \\ &\qquad \sqrt{\mathbb{E}[\|X - \hat{X}(u)\|^2 | U^* = u]}. && (70) \end{aligned}$$

Now, notice that without loss of optimality, we can assume that for each $u \in \mathcal{U}$,

$$\mathbb{E}[\|X - \hat{X}(u)\|^2 | U^* = u] \geq \mathbb{E}[\|\hat{X}^* - \hat{X}(u)\|^2 | U^* = u]. \tag{71}$$

This is justified in the following. The inequality (68) implies that

$$\Bigg| \sqrt{\mathbb{E}[\|X - \hat{X}(u)\|^2 | U^* = u]} - \sqrt{\mathbb{E}[\|\hat{X}^* - \hat{X}(u)\|^2 | U^* = u]} \Bigg| \leq \sqrt{P}. \tag{72}$$

Now, if the condition (71) is violated for some u', then we have

$$\mathbb{E}[\|\hat{X}^* - \hat{X}(u')\|^2 | U^* = u'] > \mathbb{E}[\|X - \hat{X}(u')\|^2 | U^* = u']. \tag{73}$$

Let $\hat{X}'$ be another reconstruction such that $X \to U^* \to \hat{X}'$, and

$$\begin{aligned} &P_{\hat{X}'|U^*=u'} = P_{X|U^*=u'}, && (74)\\ &P_{\hat{X}'|U^*=u''} = P_{\hat{X}^*|U^*=u''}, \qquad \forall u'' \neq u' \in \mathcal{U}. && (75) \end{aligned}$$

Clearly, $\hat{X}'$ satisfies the inequality (72). It also results in a smaller distortion than $\hat{X}^*$ which is justified as follows

$$\begin{aligned}
&\mathbb{E}[\|X-\hat{X}'\|^2] \\
&\quad= \mathbb{E}[\|X-\hat{X}(U^*)\|^2] + \mathbb{E}[\|\hat{X}'-\hat{X}(U^*)\|^2] \qquad (76)\\
&\quad= P_{U^*}(u')\mathbb{E}[\|X-\hat{X}(u')\|^2|U^*=u'] \\
&\qquad+ \sum_{u''\neq u'} P_{U^*}(u'')\mathbb{E}[\|X-\hat{X}(u'')\|^2|U^*=u''] \\
&\qquad+P_{U^*}(u')\mathbb{E}[\|\hat{X}'-\hat{X}(u')\|^2|U^*=u'] \\
&\qquad+ \sum_{u''\neq u'} P_{U^*}(u'')\mathbb{E}[\|\hat{X}'-\hat{X}(u'')\|^2|U^*=u''] \qquad (77)\\
&\quad\le P_{U^*}(u')\mathbb{E}[\|\hat{X}^*-\hat{X}(u')\|^2|U^*=u'] \\
&\qquad+ \sum_{u''\neq u'} P_{U^*}(u'')\mathbb{E}[\|X-\hat{X}(u'')\|^2|U^*=u''] \\
&\qquad+P_{U^*}(u')\mathbb{E}[\|\hat{X}'-\hat{X}(u')\|^2|U^*=u'] \\
&\qquad+ \sum_{u''\neq u'} P_{U^*}(u'')\mathbb{E}[\|\hat{X}'-\hat{X}(u'')\|^2|U^*=u''] \qquad (78)\\
&\quad= P_{U^*}(u')\mathbb{E}[\|\hat{X}^*-\hat{X}(u')\|^2|U^*=u'] \\
&\qquad+ \sum_{u''\neq u'} P_{U^*}(u'')\mathbb{E}[\|X-\hat{X}(u'')\|^2|U^*=u''] \\
&\qquad+P_{U^*}(u')\mathbb{E}[\|\hat{X}'-\hat{X}(u')\|^2|U^*=u'] \\
&\qquad+ \sum_{u''\neq u'} P_{U^*}(u'')\mathbb{E}[\|\hat{X}^*-\hat{X}(u'')\|^2|U^*=u''] \qquad (79)\\
&\quad= P_{U^*}(u')\mathbb{E}[\|\hat{X}^*-\hat{X}(u')\|^2|U^*=u'] \\
&\qquad+ \sum_{u''\neq u'} P_{U^*}(u'')\mathbb{E}[\|X-\hat{X}(u'')\|^2|U^*=u''] \\
&\qquad+P_{U^*}(u')\mathbb{E}[\|X-\hat{X}(u')\|^2|U^*=u'] \\
&\qquad+ \sum_{u''\neq u'} P_{U^*}(u'')\mathbb{E}[\|\hat{X}^*-\hat{X}(u'')\|^2|U^*=u''] \qquad (80)\\
&\quad= \mathbb{E}[\|\hat{X}^*-\hat{X}(U^*)\|^2] + \mathbb{E}[\|X-\hat{X}(U^*)\|^2] \qquad (81)\\
&\quad= \mathbb{E}[\|X-\hat{X}^*\|^2], \qquad (82)
\end{aligned}$$

where

- (78) follows from inequality (73);
- (79) follows from (75);
- (80) follows from (74);
- (81) follows from re-arranging the terms.

However, this contradicts the fact that $\hat{X}^*$ is the optimal solution. Thus, without loss of optimality, we can assume (71) in the rest of the proof.

Next, considering (68) and (71), we have the following

$$\begin{aligned}
&\mathbb{E}[\|\hat{X}^*-\hat{X}(u)\|^2|U^*=u] \ge \\
&\qquad\Big(\mathbb{E}[\|X-\hat{X}(u)\|^2|U^*=u]-\sqrt{P}\Big)^2. \qquad (83)
\end{aligned}$$

Now, taking an average of the above expression over all $u \in \mathcal{U}$, we get

$$\begin{aligned}
&\mathbb{E}[\|\hat{X}^*-\hat{X}(U^*)\|^2] \\
&\quad\ge \sum_u P_{U^*}(u)\Big(\sqrt{\mathbb{E}[\|X-\hat{X}(u)\|^2|U^*=u]}-\sqrt{P}\Big)^2 \qquad (84)\\
&\quad= \mathbb{E}[\|X-\hat{X}(U^*)\|^2] \\
&\qquad\quad -2\sqrt{P}\sum_u P_{U^*}(u)\sqrt{\mathbb{E}[\|X-\hat{X}(u)\|^2|U^*=u]}+P \\
&\hspace{20em}(85)\\
&\quad\ge \mathbb{E}[\|X-\hat{X}(U^*)\|^2]-2\sqrt{P}\sqrt{\mathbb{E}[\|X-\hat{X}(U^*)\|^2]}+P \\
&\hspace{20em}(86)\\
&\quad= \Big(\sqrt{\mathbb{E}[\|X-\hat{X}(U^*)\|^2]}-\sqrt{P}\Big)^2, \qquad (87)
\end{aligned}$$

where (86) follows because the mapping $x \mapsto \sqrt{x}$ is a concave function. Now, combining (87) with (65) gives the following inequality

$$D \ge \mathbb{E}[\|X-\hat{X}(U^*)\|^2] + \Big(\sqrt{\mathbb{E}[\|X-\hat{X}(U^*)\|^2]}-\sqrt{P}\Big)^2, \tag{88}$$

which is equivalent to the following

$$\mathbb{E}[\|X-\hat{X}(U^*)\|^2] \le \frac{\sqrt{P}+\sqrt{2D-P}}{2}. \tag{89}$$

Now, notice that the rate constraint in (62) implies the following set of inequalities

$$\begin{aligned}
R &\ge I(X;\hat{X}(U^*)) && (90)\\
&= H(X)-H(X|\hat{X}(U^*)) && (91)\\
&\ge H(X)-H(X-\hat{X}(U^*)|\hat{X}(U^*)) && (92)\\
&\ge H(X)-H(X-\hat{X}(U^*)) && (93)\\
&\ge H(X)-\frac{1}{2}\log 2\pi e(\mathbb{E}[\|X-\hat{X}(U^*)\|^2]) && (94)\\
&\ge H(X)-\frac{1}{2}\log 2\pi e\frac{2D+2\sqrt{P(2D-P)}}{4} && (95)\\
&= \frac{1}{2}\log\frac{2\sigma^2}{D+\sqrt{P(2D-P)}}, && (96)
\end{aligned}$$

where (93) follows because conditioning does not increase the entropy; (94) follows because $2^{H(X-\hat{X}(U^*))} \ge 2\pi e\mathbb{E}[\|X-\hat{X}(U^*)\|^2]$; (95) follows from (89). This concludes the proof.

C. Proof of Theorem 3

For a given representation U_r, the distortion set $\Phi_{D^0}(P_{U_r|X})$ is the set of all distortions D such that the conditions in Definition 3 are satisfied. Now, consider the distortion constraint as follows

$$\begin{aligned}
D \ge \mathbb{E}[\|X-\hat{X}\|^2] &= \mathbb{E}[\|X-\tilde{X}+\tilde{X}-\hat{X}\|^2] && (97)\\
&= \mathbb{E}[\|X-\tilde{X}\|^2]+\mathbb{E}[\|\tilde{X}-\hat{X}\|^2] && (98)\\
&= 2\mathbb{E}[\|X-\tilde{X}\|^2], && (99)
\end{aligned}$$

where (98) follows because

- $X-\tilde{X}$ is the MMSE;
- $\tilde{X}-\hat{X}$ is a function of data U_r and a noise independent of $X-\tilde{X}$; so, $X-\tilde{X}$ and $\tilde{X}-\hat{X}$ are uncorrelated;

(99) follows because

- $P_{\hat{X}|U_r} = P_{X|U_r}$;
- $\tilde{X}$ which is the MMSE reconstruction and it is a function of U_r.

V. Conclusion

In this paper, we studied the RDP tradeoff based on conditional perception measure. We showed that in the Gaussian setting, the optimal operation that converts the MMSE reconstruction to the one satisfying the perfect perceptual quality is inserting corrections by an artificial noise.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd Ed*. Wiley, 2006.

[2] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6228–6237.

[3] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. Van Gool, "Generative adversarial networks for extreme learned image compression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 221–231.

[4] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *5th International Conference on Learning Representations*, 2017.

[5] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," in *5th International Conference on Learning Representations*, 2017.

[6] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. V. Gool, "Conditional probability models for deep image compression," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] A. Golinski, R. Pourreza, Y. Yang, G. Sautiere, and T. S. Cohen, "Feedback recurrent autoencoder for video compression," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[8] R. Yang, Y. Yang, J. Marino, and S. Mandt, "Hierarchical autoregressive modeling for neural video compression," 2020. [Online]. Available: https://arxiv.org/pdf/2010.10258.pdf

[9] G. Zhang, J. Qian, J. Chen, and A. Khisti, "Universal rate-distortion-perception representations for lossy compression," in *Advances in Neural Information Processing Systems*, 2021, pp. 11 517–11 529.

[10] F. Mentzer, E. Agustsson, J. Ballé, D. Minnen, N. Johnston, and G. Toderici, "Neural video compression using gans for detail synthesis and propagation," in *European Conference on Computer Vision*, 2022.

[11] F. Mentzer, G. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," in *Advances in Neural Information Processing Systems*, 2020.

[12] A. El Gamal and Y. H. Kim, *Network Information Theory*. Cambridge University Press, 2011.

[13] C. T. Li and A. El Gamal, "Strong functional representation lemma and applications to coding theorems," *IEEE Trans. on Info. Theory*, vol. 64, no. 11, pp. 6967–6978, 2018.

[14] V. M. Panaretos and Y. Zemel, *An invitation to statistics in Wasserstein space*. Springer, 2020.