A CODING THEOREM FOR THE RATE–DISTORTION–PERCEPTION FUNCTION

Lucas Theis
Google Research
theis@google.com

Aaron B. Wagner
Cornell University
wagner@cornell.edu

ABSTRACT

The *rate–distortion–perception function* (RDPF; Blau and Michaeli, 2019) has emerged as a useful tool for thinking about realism and distortion of reconstructions in lossy compression. Unlike the rate–distortion function, however, it is unknown whether encoders and decoders exist that achieve the rate suggested by the RDPF. Building on results by Li and El Gamal (2018), we show that the RDPF can indeed be achieved using stochastic, variable-length codes. For this class of codes, we also prove that the RDPF lower-bounds the achievable rate.

1 INTRODUCTION

Lossy compression seeks to represent data with as few bits as possible while also reconstructing the data as closely as possible. In most applications of compression, however, low bit-rates and low distortion of the input are not the only desiderata. Another criterion often considered is the realism or "perceptual quality" of reconstructions. While distortion is typically measured using semimetrics, realism is more readily expressed in terms of a divergence between the source and reconstruction ensembles. When this divergence is zero, reconstructions are indistinguishable from real data and therefore maximally realistic, even though the distortion between source-reconstruction pairs may be large.

Approximately optimizing divergences using adversarial networks (Goodfellow et al., 2014) has been key to many approaches in generative modeling. In the context of compression, the joint optimization of distortions and divergences has shown particular promise for photo-realistic reconstructions at low bit-rates (e.g., Rippel and Bourdev, 2017; Ledig et al., 2017; Santurkar et al., 2018; Agustsson et al., 2019). Blau and Michaeli (2018) provided a theoretical motivation for this approach by showing that in general there is a tension between divergences and distortions, suggesting that optimizing distortion alone is not sufficient to achieve realism.

Blau and Michaeli (2019) further formalized the idea of a trade-off between different desiderata and introduced the *(information) rate–distortion–perception function (RDPF)*. This function gives the hypothetical rate of an optimal code under constraints on distortion and realism. However, unlike the conventional rate–distortion function (Shannon, 1948; 1959), it is not yet clear whether the rate suggested by the RDPF is operationally achievable, and under what conditions.

We show that the RDPF indeed describes the limit of compression under suitable constraints on distortion and realism. In Section 2 we will define the most important concepts needed for our results in Section 3.

2 PRELIMINARIES

In practice, the data source $\mathbf{X}$ is generally modeled as a real-valued (though possibly discrete) random variable $\mathbf{X} : \Omega \to \mathbb{R}^M$. Our results apply to sources in more abstract spaces, however. Logarithms are to the base 2 and entropy and mutual information are measured in bits.

Definition 1. For any distortion d, divergence D, and a random variable $\mathbf{X}$, the *(information) rate–distortion–perception function* (RDPF) is defined as (Blau and Michaeli, 2019)

$$R(\theta_d, \theta_D) = \inf_{P_{\hat{\mathbf{X}}|\mathbf{X}}} I[\mathbf{X}, \hat{\mathbf{X}}] \quad \text{s.t.} \quad \mathbb{E}[d(\mathbf{X}, \hat{\mathbf{X}})] \leq \theta_d \quad \text{and} \quad D[P_{\mathbf{X}}, P_{\hat{\mathbf{X}}}] \leq \theta_D. \tag{1}$$

More generally, we can define a rate function for an arbitrary set of constraints on $P_{\mathbf{X},\hat{\mathbf{X}}}$ as follows.

Definition 2. For a source $\mathbf{X} \sim P_{\mathbf{X}}$ and a set of real-valued functions D_i of joint distributions $P_{\mathbf{X},\hat{\mathbf{X}}}$, the *(information) rate function* (IRF) is defined as

$$R(\boldsymbol{\theta}) = \inf_{P_{\hat{\mathbf{X}}|\mathbf{X}}} I[\mathbf{X}, \hat{\mathbf{X}}] \quad \text{s.t.} \quad \forall i : D_i[P_{\mathbf{X},\hat{\mathbf{X}}}] \leq \theta_i. \tag{2}$$

The RDPF is a special case of the IRF where we chose

$$D_1[P_{\mathbf{X},\hat{\mathbf{X}}}] = \mathbb{E}[d(\mathbf{X}, \hat{\mathbf{X}})] \quad \text{and} \quad D_2[P_{\mathbf{X},\hat{\mathbf{X}}}] = D[P_{\mathbf{X}}, P_{\hat{\mathbf{X}}}]. \tag{3}$$

The generalized constraint in Equation 2 was anticipated by Shannon (1948). We prove our results for the IRF, from which results on the RDPF immediately follow. The ready unification of distortion and perception suggests that these concepts are less distinct than it might first appear. Note that rate-distortion theorems with multiple, competing constraints have been considered in other contexts (Kakavand and El Gamal, 2006; Csiszár and Körner, 1981, Exercise 2.2.14).

So far we have only defined informational rate functions which may or may not have any connection to the operational compression problem. The hope is, of course, that the value of these functions represents the rate of the best codes whose reconstructions satisfy the given constraints. That is, we hope that a code exists which *achieves* the given rate and, conversely, that no code can do better.

Definition 3. For an arbitrary set $\mathcal{X}$, we define a *(stochastic) encoder* as any function contained in

$$\mathcal{F} = \{f : \mathcal{X} \times \mathbb{R} \to \mathbb{N}_0\}. \tag{4}$$

Similarly, we define *(stochastic) decoders* as elements of

$$\mathcal{G} = \{g : \mathbb{N}_0 \times \mathbb{R} \to \mathcal{X}\}. \tag{5}$$

A *(stochastic) code* is an element of $\mathcal{F} \times \mathcal{G}$.

Note that a stochastic encoder expects data and an additional source of randomness as input. Without loss of generality, we can assume that this randomness is represented by a single real number (or an infinite number of bits). Similar to the encoder, the decoder receives additional random bits as input. The encoder and decoder may receive the same random bits or they may each have their own source of randomness, and this has implications for the efficiency of a code (e.g., Cuff, 2008). Throughout the paper, we assume that the encoder and decoder receive the same bits.

Next we define our notion of achievability. Following the conventions in information theory, we distinguish between one-shot achievability and (asymptotic) achievability.

Definition 4. For a source $\mathbf{X} \sim P_{\mathbf{X}}$ and a given set of constraints, we say that a rate R is *one-shot achievable* if an encoder $f \in \mathcal{F}$, a decoder $g \in \mathcal{G}$, and a random variable U exist with

$$K = f(\mathbf{X}, U) \quad \text{and} \quad \hat{\mathbf{X}} = g(K, U) \tag{6}$$

such that the joint distribution $P_{\mathbf{X},\hat{\mathbf{X}}}$ satisfies the constraints and the conditional entropy of K is not more than R, $H[K \mid U] \leq R$.

This definition is justified by the fact that a variable-rate prefix-free entropy code always exists whose coding cost is within 1 bit of the entropy, $H[K \mid U]$. Note that an entropy coder may use the variable U to assign bits to K since U is known to both the encoder and the decoder.

In one-shot achievability, the source $\mathbf{X}$ is not assumed to have any discernible structure. This is in contrast to (asymptotic) achievability, for which the source is assumed to be a long i.i.d. sequence.

Definition 5. For a source $\mathbf{X} \sim P_{\mathbf{X}}$ and a given set of constraints, we say that a rate R is *(asymptotically) achievable* if there exists a sequence of encoders $f_N : \mathcal{X}^N \times \mathbb{R} \to \mathbb{N}_0$ and decoders $g_N : \mathbb{N}_0 \times \mathbb{R} \to \mathcal{X}^N$ with

$$K_N = f(\mathbf{X}^N, U) \quad \text{and} \quad \hat{\mathbf{X}}^N = g(K_N, U) \tag{7}$$

such that each joint distribution $P_{\mathbf{X}_n, \hat{\mathbf{X}}_n}$ $(n = 1, \dots, N)$ satisfies the constraints and

$$\lim_{N \to \infty} H[K_N \mid U]/N \leq R. \tag{8}$$

An important detail is that this definition requires that the reconstruction of each individual data point satisfies the constraints, not just when averaged over the N points.

3 Achievability of the IRF (and RDPF)

We first prove the one-shot achievability of an upper bound on the IRF.

Theorem 1. *Let an arbitrary source $\mathbf{X} \sim P_{\mathbf{X}}$ and constraints $D_i[P_{\mathbf{X},\hat{\mathbf{X}}}] \leq \theta_i$ be given. If*

$$R > R(\boldsymbol{\theta}) + \log(R(\boldsymbol{\theta}) + 1) + 4, \tag{9}$$

then R is one-shot achievable.

Proof. This result is a direct consequence of the properties of the *Poisson functional representation* introduced by Li and El Gamal (2018). By the definition of the IRF, there exists a $P_{\hat{\mathbf{X}}|\mathbf{X}}$ such that

$$\forall i : D_i[P_{\mathbf{X},\hat{\mathbf{X}}}] \leq \theta_i \quad \text{and} \quad I[\mathbf{X}, \hat{\mathbf{X}}] \leq R(\boldsymbol{\theta}) + \varepsilon \tag{10}$$

for any $\varepsilon > 0$. Using a shared source of randomness, an encoder and decoder are both able to generate a draw from the following random variables for $i \in \mathbb{N}$,

$$\hat{\mathbf{X}}_i \sim P_{\hat{\mathbf{X}}}, \quad S_i \sim \text{Exp}(1), \quad T_i = \textstyle\sum_{j=1}^{i} S_i. \tag{11}$$

Here, S_i are exponentially distributed so that T_i are the epochs of a Poisson process. For an input $\mathbf{x}$, the encoder selects one of the candidates $\hat{\mathbf{X}}_i$ as follows,

$$K = \operatorname*{argmin}_{i \in \mathbb{N}} T_i \frac{dP_{\hat{\mathbf{X}}}}{dP_{\hat{\mathbf{X}}|\mathbf{X}}(\,\cdot \mid \mathbf{x})}(\hat{\mathbf{X}}_i). \tag{12}$$

After receiving K, the decoder reconstructs $\hat{\mathbf{X}}_K$. That is,

$$f(\mathbf{x}, U) = K \quad \text{and} \quad g(K, U) = \hat{\mathbf{X}}_K, \tag{13}$$

where U is a random variable representing the shared source of randomness used to construct the random variables in Equation 11. Li and El Gamal (2018) proved that $\hat{\mathbf{X}}_K \sim P_{\hat{\mathbf{X}}|\mathbf{X}}$, that is, the code is communicating a sample from the conditional distribution. It was further shown that

$$H[K] < I[\mathbf{X}, \hat{\mathbf{X}}] + \log(I[\mathbf{X}, \hat{\mathbf{X}}] + 1) + 4, \tag{14}$$

that is, the indices' entropy is not much more than the mutual information. Hence,

$$H[K \mid U] \leq H[K] < R(\boldsymbol{\theta}) + \log(R(\boldsymbol{\theta}) + \varepsilon) + 4 + \varepsilon < R, \tag{15}$$

if we choose ε small enough. □

Next, we show that the IRF is a lower bound on the one-shot achievable rate. This extends a similar result by Blau and Michaeli (2019, Appendix C) from deterministic to stochastic codes.

Theorem 2. *Let an arbitrary source $\mathbf{X} \sim P_{\mathbf{X}}$ and constraints $D_i[P_{\mathbf{X},\hat{\mathbf{X}}}] \leq \theta_i$ be given. If $R < \infty$ is one-shot achievable, then $R \geq R(\boldsymbol{\theta})$.*

Proof. Applying properties of mutual information and discrete entropy, we have

$$\begin{aligned} H[f(\mathbf{X},U) \mid U] &\ge I[\mathbf{X}, f(\mathbf{X},U) \mid U] \quad (16) \\ &\ge I[\mathbf{X}, \hat{\mathbf{X}} \mid U] \quad (17) \\ &= I[\mathbf{X}, \hat{\mathbf{X}} \mid U] + I[\mathbf{X}, U] \quad (18) \\ &= I[\mathbf{X}, (\hat{\mathbf{X}}, U)] \ge I[\mathbf{X}, \hat{\mathbf{X}}] \ge R(\boldsymbol{\theta}). \quad \square \end{aligned}$$

We are now ready to prove the IRF (and hence RDPF) coding theorem.

Theorem 3. *Let an arbitrary source $\mathbf{X} \sim P_{\mathbf{X}}$ and constraints $D_i[P_{\mathbf{X},\hat{\mathbf{X}}}] \le \theta_i$ be given. Then $R < \infty$ is achievable if and only if $R \ge R(\boldsymbol{\theta})$.*

Proof. We first show that if $R(\boldsymbol{\theta}) < \infty$, then $R(\boldsymbol{\theta})$ is itself achievable. By the definition of the IRF, for each $N > 0$ there exists a (potentially different) $P_{\hat{\mathbf{X}}|\mathbf{X}}$ such that

$$I[\mathbf{X}, \hat{\mathbf{X}}] < R(\boldsymbol{\theta}) + \frac{1}{N} \tag{19}$$

and $P_{\mathbf{X},\hat{\mathbf{X}}}$ satisfies the constraints. Following the same approach as in Theorem 1, we can construct a code communicating a sample from

$$P_{\hat{\mathbf{X}}^N|\mathbf{X}^N} = \textstyle\prod_{n=1}^N P_{\hat{\mathbf{X}}|\mathbf{X}}, \tag{20}$$

where the right-hand side indicates a product measure, that is, $\hat{\mathbf{X}}_n$ will only depend on $\mathbf{X}_n$. Each individual data point and reconstruction $(\mathbf{X}_n, \hat{\mathbf{X}}_n)$ follows the marginal distribution $P_{\mathbf{X},\hat{\mathbf{X}}}$ and thus satisfies the constraints. We encode the data points jointly into some K_N so that its average entropy is now at most

$$\begin{aligned} \frac{H[K_N]}{N} &< \frac{1}{N}(I[\mathbf{X}^N, \hat{\mathbf{X}}^N] + \log(I[\mathbf{X}^N, \hat{\mathbf{X}}^N] + 1) + 4) \quad (21) \\ &= I[\mathbf{X}, \hat{\mathbf{X}}] + \frac{1}{N}\log(N I[\mathbf{X}, \hat{\mathbf{X}}] + 1) + \frac{4}{N} \quad (22) \\ &< R(\boldsymbol{\theta}) + \frac{1}{N}\log(N R(\boldsymbol{\theta}) + 2) + \frac{5}{N}. \quad (23) \end{aligned}$$

This converges to $R(\boldsymbol{\theta})$ in the limit of large N, proving the former's achievability. Turning to the converse, for each N, let $R_N(\boldsymbol{\theta})$ denote the IRF of $\mathbf{X}^N$ with constraints $D_i[P_{\mathbf{X}_n,\hat{\mathbf{X}}_n}] \le \theta_i \;\forall\, i, n$. If R is achievable then there exists a sequence of codes (f_N, g_N) satisfying these constraints and

$$\lim_{N\to\infty} H[K_N \mid U]/N \le R, \tag{24}$$

where $K_N = f_N(\mathbf{X}_N, U)$. But by Theorem 2 we must have $H[K_N \mid U] \ge R_N(\boldsymbol{\theta})$ and so

$$R \ge \lim_{N\to\infty} H[K_N \mid U]/N \ge \lim_{N\to\infty} R_N(\boldsymbol{\theta})/N. \tag{25}$$

It thus suffices to show that $R_N(\boldsymbol{\theta}) \ge N R(\boldsymbol{\theta})$. But this follows from the chain rule for mutual information, the data processing theorem, and the independence of X_n from X_m $(m \ne n)$:

$$\begin{aligned} I[\mathbf{X}^N, \hat{\mathbf{X}}^N] &= \textstyle\sum_{n=1}^N I[\mathbf{X}_n, \hat{\mathbf{X}}^N \mid \mathbf{X}_1, \ldots, \mathbf{X}_{n-1}] \\ &= \textstyle\sum_{n=1}^N I[\mathbf{X}_n, \hat{\mathbf{X}}^N \mid \mathbf{X}_1, \ldots, \mathbf{X}_{n-1}] + \sum_{n=1}^N I[\mathbf{X}_n, (\mathbf{X}_1, \ldots, \mathbf{X}_{n-1})] \\ &= \textstyle\sum_{n=1}^N I[\mathbf{X}_n, (\hat{\mathbf{X}}^N, \mathbf{X}_1, \ldots, \mathbf{X}_{n-1})] \\ &\ge \textstyle\sum_{n=1}^N I[\mathbf{X}_n, \hat{\mathbf{X}}_n] \ge \sum_{n=1}^N R(\boldsymbol{\theta}) = N R(\boldsymbol{\theta}). \quad \square \end{aligned}$$

4 Discussion

By building on the results of Li and El Gamal (2018) we proved a complete coding theorem for the RDPF (and more generally, any IRF). This clarifies the practical relevance of the RDPF. Note that our asymptotic formulation in Definition 5 requires the constraints to hold for each n, not just on average. This stronger constraint is satisfied by the construction of Li and El Gamal (2018) and eliminated the need for the assumption, used by Blau and Michaeli (2019), that the divergence measure is convex. On the other hand, unlike Blau and Michaeli (2019), our formulation allowed for variable-length coding and common randomness between the encoder and decoder. The case of fixed-rate and/or deterministic codes is also of interest but is left for future research.

Acknowledgments

We would like to thank Johannes Ballé, Eirikur Agustsson and Ashok Popat for helpful discussions and feedback on the manuscript. The second author was supported by the US National Science Foundation under grants CCF-1617673, CCF-1934985, and CCF-2008266 and the US Army Research Office under grant W911NF-18-1-0426.

References

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/CVF International Conference on Computer Vision (ICCV)*, pages 221–231, 2019.

Y. Blau and T. Michaeli. The perception-distortion tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018.

Y. Blau and T. Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, 2019.

I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Akadémiai Kiaodó, Budapest, 1981.

P. Cuff. Communication requirements for generating correlated random variables. In *2008 IEEE International Symposium on Information Theory*, pages 1393–1397, 2008.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, 2014.

H. Kakavand and A. El Gamal. Source description cost. In *2006 IEEE International Symposium on Information Theory*, pages 272–276, 2006.

C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017.

C. T. Li and A. El Gamal. Strong Functional Representation Lemma and Applications to Coding Theorems. *IEEE Transactions on Information Theory*, 64(11):6967–6978, 2018. doi: 10.1109/TIT.2018.2865570.

O. Rippel and L. Bourdev. Real-time adaptive image compression. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2922–2930, 2017.

S. Santurkar, D. Budden, and N. Shavit. Generative compression. In *2018 Picture Coding Symposium (PCS)*, pages 258–262, 2018.

C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27: 379–423, 1948.

C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Natl. Conv. Record*, pages 142–163, 1959.