TOWARDS EMPIRICAL SANDWICH BOUNDS ON THE RATE-DISTORTION FUNCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Rate-distortion (R-D) function, a key quantity in information theory, characterizes the fundamental limit of how much a data source can be compressed subject to a fidelity criterion, by *any* lossy compression algorithm. As researchers push for ever-improving compression performance, establishing the R-D function of any given data source is not only of scientific interest, but also sheds light on the room for possible improvement of compression algorithms. Previous work on this problem relied on distributional assumptions on the data source (Gibson, 2017) or only applied to low-dimensional discrete data. By contrast, this paper makes the first attempt at an algorithm for sandwiching the R-D function of arbitrary sources requiring only data samples. We verify the tightness of our bounds on Gaussian and banana-shaped sources, and demonstrate the scalability of our upper bound on natural images. Our results indicate room for improving the compression performance of state-of-the-art methods by one PSNR at various bitrates.

1 INTRODUCTION

From storing astronomical images captured by the Hubble telescope, to delivering familiar faces and voices over video chats, data compression, i.e., communicating the "same" information but with less bits, is commonplace and indispensable to our digital life, and even arguably lies at the heart of intelligence (Mahoney, 2009). While for lossless compression, there exist practical algorithms that can compress any discrete data arbitrarily close to the information theory limit (Ziv & Lempel, 1977; Witten et al., 1987), no such universal algorithm has been found for *lossy* data compression (Berger & Gibson, 1998), and significant research efforts have dedicated to lossy compression algorithms for various data. Recently, deep learning has shown promise for learning lossy compressors from raw data examples, with continually improving compression performance often matching or exceeding traditionally engineered methods (Minnen et al., 2018; Agustsson et al., 2020; Yang et al., 2020a).

However, there are fundamental limits to the performance of any lossy compression algorithm, due to the inevitable trade-off between *rate*, the average number of bits needed to represent the data, and the *distortion* incurred by lossy representations. This trade-off is formally described by the rate-distortion (R-D) function, for a given *source* (i.e., the data distribution of interest; referred to as such in information theory) and distortion metric. The R-D function characterizes the best theoretically achievable rate-distortion performance by any compression algorithm, which can be seen as a lossy-compression counterpart and generalization of Shannon's entropy for lossless compression.

Despite its fundamental importance, the R-D function is generally unknown analytically, and establishing it for general data sources, especially real world data, is a difficult problem (Gibson, 2017). The default method for computing R-D functions, the Blahut-Arimoto algorithm (Blahut, 1972; Arimoto, 1972), only works for discrete data with a known probability mass function and has a complexity exponential in the data dimensionality. Applying it to an unknown data source requires discretization (if it is continuous) and estimating the source probabilities by a histogram, both of which introduce errors and are computationally infeasible beyond a couple of dimensions. Previous work characterizing the R-D function of images and videos (Hayes et al., 1970; Gibson, 2017) all required a statistical model of the source; thus the result is only as meaningful as the source model.

In this work, we make progress on this problem by introducing new tools for bounding the R-D function of a *general* (i.e., discrete, absolutely continuous, or neither), *unknown* memoryless source. Our contributions are as follows:

1. We apply machine learning techniques to obtain new algorithms for producing upper and lower bounds on the R-D function of an unknown source, requiring only i.i.d. data samples.
2. Our upper bound draws from the deep generative modeling toolbox, and is closely related to a type of β-VAEs in learned data compression (Ballé et al., 2017); we clarify how these models optimize a model-independent upper bound on the source rate-distortion function.
3. We theoretically derive a general lower bound on the R-D function that can in principle be optimized by gradient ascent. Facing the difficulty of the problem involving global optimization, we restrict to a (mean-)squared error distortion and obtain a practical algorithm.
4. We show experimentally that our upper bound algorithm can converge to the *exact* analytical R-D function on randomly sampled Gaussian sources. On a banana-shaped source and its high-dimensional projections, we obtain tight sandwich bounds. As far as we know, this is the first successful attempt at computationally characterizing the R-D function of such a non-trivial source in high dimensions, where the Blahut-Arimoto algorithm is infeasible.
5. We establish new upper bounds on the R-D function of natural images. Based on a ResNet-VAE architecture, our best upper bound indicates possible room for improving the compression performance of state-of-the-art methods by one PSNR at various bitrates.

We begin by introducing needed concepts of rate-distortion theory in Section. 2, then describe our upper and lower bound algorithms in Section. 3 and Section. 4, respectively. We review related work in Section. 5, report experimental results in Section. 6, and conclude in Section. 7.

2 Background

Rate-distortion (R-D) theory deals with the fundamental trade-off between the average number of bits per sample (*rate*) used to represent a data source X and the *distortion* incurred by the lossy representation $\hat{X}$. It asks the following question about the limit of lossy compression: for a given data source and a distortion metric (a.k.a., a fidelity criterion), what is the minimum number of bits (per sample) needed to represent the source at a tolerable level of distortion, regardless of the computation complexity of the compression procedure? The answer is given by the rate-distortion function $R(D)$. To introduce it, let the source and its reproduction take values in the sets $\mathcal{X}$ and $\hat{\mathcal{X}}$, conventionally called the *source* and *reproduction alphabets*, respectively. We define the data source formally by a random variable $X \in \mathcal{X}$ following a (usually unknown) distribution P_X, and assume a distortion metric $\rho : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$ has been given, such as the squared error $\rho(x, \hat{x}) = \|x - \hat{x}\|^2$. The rate-distortion function is then defined by the following constrained optimization problem,

$$R(D) = \inf_{Q_{\hat{X}|X}:\, \mathbb{E}[\rho(X,\hat{X})] \le D} I(X; \hat{X}), \tag{1}$$

where we consider all random transforms $Q_{\hat{X}|X}$ whose expected distortion is within the given threshold $D \ge 0$, and minimize the mutual information between the source X and its reproduced representation $\hat{X}$ [1]. Shannon's lossy source coding theorems (Shannon, 1948; 1959) gave operational significance to the above mathematical definition of $R(D)$, as the minimum achievable rate with which any lossy compression algorithm can code i.i.d. data samples at a distortion level within D.

The R-D function thus gives the tightest lower bound on the rate-distortion performance of any lossy compression algorithm, and can inform the design and analysis of such algorithms. If the operational rate-distortion performance of an algorithm lies high above the source $R(D)$-curve $(D, R(D))$, then further performance improvement may be expected; otherwise, its rate-distortion performance is already close to theoretically optimal, and we may focus our attention on improving other aspects of the algorithm. Unfortunately, the R-D function does not have an analytical expression in general.

3 Upper Bound Algorithm

For a known discrete source, the Blahut-Arimoto (BA) algorithm converges to $R(D)$ from above by fixed-point equations. For a general unknown source, it is not clear how this can be done. We

[1] Both the expected distortion and mutual information terms are defined w.r.t. the joint distribution $P_X Q_{\hat{X}|X}$. We give formal measure theoretic definitions of various quantities in the appendix.

propose to solve the underlying variaitonal problem approximately by gradient descent. In exchange for generality and scalability, we lose the optimality guarantee of the BA algorithm, only arriving at a stochastic upper bound of $R(D)$ in general. By R-D theory (Cover & Thomas, 2006), every (distortion, rate) pair lying above the $R(D)$-curve is in principle realizable by a (possibly expensive) compression algorithm; an upper bound on $R(D)$, therefore, reveals what kind of (improved) R-D performance is theoretically possible, without suggesting how it can be practically achieved.

Variational Formulation. We adopt the same unconstrained variational objective as the Blahut-Arimoto (BA) algorithm (Blahut, 1972; Arimoto, 1972), in its most general form,

$$\mathcal{L}(Q_{\hat{X}|X}, Q_{\hat{X}}, \lambda) := \mathbb{E}_{x\sim P_X}[KL(Q_{\hat{X}|X=x}\|Q_{\hat{X}})] + \lambda \mathbb{E}_{P_X Q_{\hat{X}|X}}[\rho(X, \hat{X})], \tag{2}$$

where $Q_{\hat{X}}$ is an arbitrary probability measure on $\hat{\mathcal{X}}$ and $KL(\cdot\|\cdot)$ denotes the Kullback-Leibler (KL) divergence. This objective can be seen as a Lagrangian relaxation of the constrained problem defining $R(D)$, where the first (*rate*) term is a variational upper bound on the mutual information $I(X;\hat{X})$, and the second (*distortion*) term enforces the distortion tolerance constraint in Eq. 1. For each fixed $\lambda > 0$, a global minimizer of $\mathcal{L}$ yields a point $(\mathcal{R}, \mathcal{D})$ on the $R(D)$ curve (Csiszar, 1974), where $\mathcal{R}$ and $\mathcal{D}$ are simply the two terms of $\mathcal{L}$ evaluated at the optimal $(Q_{\hat{X}|X}, Q_{\hat{X}})$. Repeating this minimization for various λ then traces out the $R(D)$ curve. Based on this connection, the BA algorithm carries out the minimization by coordinate descent on $\mathcal{L}$ via fixed-point equations, each time setting $Q_{\hat{X}|X}$ to be optimal w.r.t. $Q_{\hat{X}}$ and vice versa; the sequence of alternating distributions can be shown to converge and yield a point on the $R(D)$ curve (Csiszar, 1974). Unfortunately, the BA algorithm only applies when $\mathcal{X}$ and $\hat{\mathcal{X}}$ are finite, and the source distribution P_X known (or estimated from data samples) in the form of a vector of probabilities $P_X(x)$ of every state $x \in \mathcal{X}$. Moreover, the algorithm requires storage and running time exponential in the data dimensionality, since it operates on exhaustive tabular representations of $P_X, \rho, Q_{\hat{X}|X}$, and $Q_{\hat{X}}$. The algorithm therefore quickly becomes infeasible on data with more than a couple of dimensions, not to mention high-dimension data such as natural images. The fixed-point equations of the BA algorithm are known in general settings (Rezaei et al., 2006), but when $\mathcal{X}$ or $\hat{\mathcal{X}}$ is infinite (such as in the continuous case), it is not clear how to carry them out or exhaustively represent the measures $Q_{\hat{X}}$ and $Q_{\hat{X}|X=x}$ for each $x \in \mathcal{X}$.

Proposed Method. To avoid these difficulties, we propose to apply (stochastic) gradient descent on $\mathcal{L}$ w.r.t. flexibly parameterized variational distributions $Q_{\hat{X}|X}$ and $Q_{\hat{X}}$. The distributions can be members from any variational family, as long as $Q_{\hat{X}|X=x}$ is absolutely continuous w.r.t. $Q_{\hat{X}}$ for (P_X-almost) all $x \in \mathcal{X}$. This technical condition ensures their KL divergence is well defined. This is easily satisfied, when $\hat{\mathcal{X}}$ is discrete, by requiring the support of $Q_{\hat{X}|X=x}$ to be contained in that of $Q_{\hat{X}}$ for all $x \in \mathcal{X}$; and in the continuous case, by representing both measures in terms of probability density functions (e.g., normalizing flows (Kobyzev et al., 2021)). In this work we represent $Q_{\hat{X}}$ and $Q_{\hat{X}|X}$ by parametric distributions, and predict the parameters of each $Q_{\hat{X}|X=x}$ by an *encoder* neural network $\phi(x)$ as in amortized inference (Kingma & Welling, 2014); we note it is also possible to represent $Q_{\hat{X}|X}$ non-parametrically (Liu & Wang, 2016). Given a dataset of i.i.d. X samples, we optimize the parameters of the variational distributions by SGD on $\mathcal{L}$; at convergence, the estimates of rate and distortion terms of $\mathcal{L}$ yields a point that lies on a stochastic R-D upper bound $R_U(D)$.

The variational objective $\mathcal{L}$ (Eq. 2) closely resembles the negative ELBO (NELBO) objective of a β-VAE (Higgins et al., 2017), if we regard the reproduction alphabet $\hat{\mathcal{X}}$ as the "latent space". The connection is immediate when $\hat{\mathcal{X}}$ is continuous, so that a squared error ρ specifies the density of a Gaussian likelihood $p(x|\hat{x}) \propto \exp(-\|x - \hat{x}|^2)$. However, unlike in data compression, where $\hat{\mathcal{X}}$ is determined by the application (and often equal to $\mathcal{X}$ for a full-reference distortion), the latent space in a (β-)VAE typically has a lower dimension than $\mathcal{X}$; a *decoder* network is then used to parameterize a likelihood model in the data space. To capture this setup, we introduce a new, arbitrary latent space $\mathcal{Z}$ on which we define variational distributions $Q_{Z|X}, Q_Z$, and a (possibly stochastic) decoder function $\omega : \mathcal{Z} \to \hat{\mathcal{X}}$. This results in an extended objective (with $\mathcal{L}$ being the special case of an identity ω),

$$J(Q_{Z|X}, Q_Z, \omega, \lambda) := \mathbb{E}_{x\sim P_X}[KL(Q_{Z|X=x}\|Q_Z)] + \lambda \mathbb{E}_{P_X Q_{Z|X}}[\rho(X, \omega(Z))]. \tag{3}$$

How does this relate to the original rate-distortion problem? We note that the same results from rate-distortion theory apply, once we identify a new distortion function $\rho_\omega(x, z) := \rho(x, \omega(z))$ and

treat $\mathcal{Z}$ as the reproduction alphabet. Then for each fixed decoder, we may define a ω-dependent rate-distortion function, $R_\omega(D) := \inf_{Q_{Z|X}:\mathbb{E}[\rho_\omega(X,Z)]\leq D} I(X;Z)$. The minimum of J w.r.t. the variational distributions produces a point on the $R_\omega(D)$ curve. Moreover, as a consequence of the data processing inequality $I(X;Z) \geq I(X;\omega(Z))$, we can prove (in Appendix A.3) that $R_\omega(D) \geq R(D)$ for any ω, with equality for bijective ω . Therefore, we can minimize the NELBO-like objective Eq. 3 w.r.t. parameters of $(Q_{Z|X}, Q_Z, \omega)$ similar to training a β-VAE, knowing that we are optimizing an upper bound on the information R-D function of the data source. This can be seen as a generalization to the lossless case (with a countable $\mathcal{X}$), where minimizing the NELBO minimizes an upper bound on the Shannon entropy of the source (Frey & Hinton, 1997), the limit of lossless compression.

The tightness of our bound depends on the choice of variational distributions. The freedom to define them over any suitable latent space $\mathcal{Z}$ can simplify the modeling task (of which there are many tools (Salakhutdinov, 2015; Kobyzev et al., 2021)). e.g., we can work with densities on a continuous $\mathcal{Z}$, even if $\hat{\mathcal{X}}$ is high-dimensional and discrete. We can also treat Z as the concatenation of sub-vectors $[Z_1, Z_2, ..., Z_L]$, and parameterize Q_Z in terms of simpler component distributions $Q_Z = \prod_{l=1}^{L} Q_{Z_l|Z_{<l}}$ (similarly for $Q_{Z|X}$). We exploit these properties in our Sec. 6.3 experiments.

4 Lower Bound Algorithm

Without knowing the tightness of an R-D upper bound, we could be wasting time and resources trying to improve the R-D performance of a compression algorithm, when it is in fact already close to the theoretical limit. This would be avoided if we could find a matching *lower* bound on $R(D)$. Unfortunately, the problem turns out to be much more difficult computationally. Indeed, every compression algorithm, or every pair of variational distributions $(Q_{\hat{X}}, Q_{\hat{X}|X})$ yields a point above $R(D)$. Conversely, establishing a lower bound requires disproving the existence of *any* compression algorithm that can conceivably operate below the $R(D)$ curve. In this section, we derive an algorithm that can in principle produce arbitrarily tight R-D lower bounds. However, as an indication of its difficulty, the problem requires *globally* maximizing a family of partition functions. By restricting to a continuous reproduction alphabet and a squared error distortion, we make some progress on this problem and demonstrate useful lower bounds on data with low effective dimension (see Sec. 6.2).

Dual characterization of $R(D)$. While upper bounds on $R(D)$ arise naturally out of its definition as a minimization problem, a variational lower would require expressing $R(D)$ through a *maximization* problem. For this, we introduce a "conjugate" function as the optimum of the Lagrangian Eq. 2 ($Q_{\hat{X}}$ is eliminated by replacing the rate upper bound with the exact mutual information $I(X;\hat{X})$):

$$F(\lambda) \quad := \quad \inf_{Q_{\hat{X}|X}} I(X;\hat{X}) \; + \; \lambda \mathbb{E}[\rho(X,\hat{X})]. \quad (4)$$

Geometrically, $F(\lambda)$ is the maximum R-axis intercept of a straight line with slope $-\lambda$, among all such lines that lie below or tangent to $R(D)$; the R-D curve can then be found by taking the upper envelope of lines with slope $-\lambda$ and R-intercept $F(\lambda)$, i.e., $R(D) = \max_{\lambda\geq 0} F(\lambda) - \lambda D$. A key result for our lower bound is the dual characterization of $F(\lambda)$ in terms of maximization (depicted in Fig. 1):

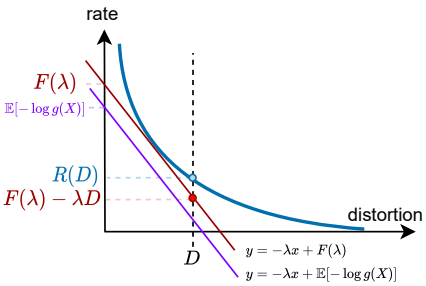

Figure 1: The geometry of the R-D lower bound problem. For a given slope $-\lambda$, we seek to maximize the R-axis intercept, $\mathbb{E}[-\log g(X)]$, over all $g \geq 0$ functions admissible according to Eq. 6.

Theorem 4.1. *(Csiszár, 1974) Under basic conditions (e.g., satisfied by a bounded ρ; see Appendix A.2), it holds that (all expectations below are with respect to the data source r.v. $X \sim P_X$)*

$$F(\lambda) = \max_{g(x)} \{\mathbb{E}[-\log g(X)]\} \quad (5)$$

where the maximization is over $g : \mathcal{X} \to [0, \infty)$ satisfying the constraint

$$\mathbb{E}\left[\frac{\exp(-\lambda\rho(X,\hat{x}))}{g(X)}\right] = \int \frac{\exp(-\lambda\rho(x,\hat{x}))}{g(x)} dP_X(x) \leq 1, \forall \hat{x} \in \hat{\mathcal{X}} \quad (6)$$

In other words, every admissible g yields a lower bound of $R(D)$, via an underestimator of the intercept $\mathbb{E}[-\log g(X)] \leq F(\lambda)$. We give the origin and history of this result in related work Sec. 5.

Proposed Unconstrained Formulation. The constraint in Eq. 6 is concerning – it is a family of possibly infinitely many constraints, one for each $\hat{x}$. To make the problem easier to work with, we propose to eliminate the constraints by the following transformation. Let g be defined in terms of another function $u(x) \geq 0$ and a scalar c depending on u, such that

$$g(x) := cu(x), \quad \text{where } c := \sup_{\hat{x} \in \hat{\mathcal{X}}} \Psi_u(\hat{x}), \quad \text{and } \Psi_u(\hat{x}) := \mathbb{E}\left[\frac{\exp -\lambda\rho(X, \hat{x})}{u(X)}\right].$$

This reparameterization of g is without loss of generality, and can be shown to always satisfy the constraint in Eq. 6. While this form of g bears a superficial resemblance to an energy-based model (LeCun et al., 2006), with $\frac{1}{c}$ resembling a normalizing constant, there is an important difference: $c = \sup_{\hat{x}} \Psi_u(\hat{x})$ is in fact the supremum of a family of "partition functions" $\Psi_u(\hat{x})$ indexed by $\hat{x}$; we thus refer to c as the *sup-partition function*. Although all these quantities have λ-dependence, we omit this from our notation since λ is a fixed input parameter (as in the upper bound algorithm).

Consequently, F is now the result of *unconstrained* maximization over all u functions, and we obtain a lower bound on it by restricting u to a subset of functions with parameters θ (e.g., neural networks),

$$F(\lambda) = \max_{u \geq 0} \{\mathbb{E}[-\log u(X)] - \log \sup_{\hat{x} \in \hat{\mathcal{X}}} \Psi_u(\hat{x})\} \geq \max_{\theta} \{\mathbb{E}[-\log u_\theta(X)] - \log \sup_{\hat{x} \in \hat{\mathcal{X}}} \Psi_\theta(\hat{x})\}$$

For convenience, define the maximization objective by

$$\ell(\theta) := \mathbb{E}[-\log u_\theta(X)] - \log c(\theta), \quad c(\theta) := \sup_{\hat{x} \in \hat{\mathcal{X}}} \Psi_\theta(\hat{x}). \tag{7}$$

Then, it can in principle be maximized by stochastic gradient ascent using samples from P_X. However, computing the sup-partition function $c(\theta)$ poses serious computation challenges: even evaluating $\Psi_\theta(\hat{x})$ for a single $\hat{x}$ value involves a potentially high-dimensional integral w.r.t. P_X; this is only exacerbated by the need to globally optimize w.r.t. $\hat{x}$, an NP-hard problem even in one-dimension.

Proposed Method. To tackle this problem, we propose a sample-based over-estimator of the sup-partition function inspired by the IWAE estimator (Burda et al., 2015). Fix u_θ for now. Noting that $\Psi(\hat{x}) := \mathbb{E}[\psi(X, \hat{x})]$ is an expectation w.r.t. P_X, where $\psi(x, \hat{x}) := \frac{\exp -\lambda\rho(x,\hat{x})}{u(x)}$, one may then naturally consider a plug-in estimator for c, replacing the expectation by a sample estimate of $\Psi(\hat{x})$. Formally, given a sequence of i.i.d. random variables $X_1, X_2, ... \sim P_X$, we define the estimator $C_k := \sup_{\hat{x}} \frac{1}{k} \sum_i \psi(X_i, \hat{x})$ for each $k \geq 1$. We can then prove (see Theorem A.3 and proof in Appendix) that $\mathbb{E}[C_1] \geq \mathbb{E}[C_2] \geq ...c$, i.e., C_k is on average an over-estimator of the sup-partition function c; and like the Importance-Weighted ELBO (Burda et al., 2015), the bias of the estimator decreases monotonically as $k \to \infty$, and that under continuity assumptions, C_k is asymptotically unbiased and converges to c. In light of this, we replace c by $\mathbb{E}[C_k]$ and obtain a k-sample under-estimator of the objective $\ell(\theta)$ (which in turn underestimates $F(\lambda)$):

$$\ell_k(\theta) := \mathbb{E}[-\log u_\theta(X)] - \log \mathbb{E}[C_k]; \quad \text{moreover, } \ell_1(\theta) \leq \ell_2(\theta) \leq ... \leq \ell(\theta).$$

Unfortunately, we still cannot apply stochastic gradient ascent to ℓ_k, as two more difficulties remain. First, C_k is still hard to compute, as it is defined through a global maximization problem. We note that by restricting to a suitable ρ and $\hat{X} = X$, the maximization objective of C_k has the form of a kernel density estimate (KDE). For a squared error distortion, this becomes a Gaussian mixture density, $\frac{1}{k} \sum_i \psi(x_i, \hat{x}) \propto \sum_i \pi_i \exp(-\lambda \|x_i - \hat{x}\|^2)$, with centroids defined by the samples $x_1, ..., x_k$, and mixture weights $\pi_i = (ku(x_i))^{-1}$. The global mode of a Gaussian mixture can generally be found by hill-climbing from each of the k centroids, except in rare artificial examples (Carreira-Perpinan, 2000; 2020); we therefore use this procedure to compute C_k, but note that other methods exist (Lee et al., 2019; Carreira-Perpinan, 2007). The second difficulty is that even if we could estimate $\mathbb{E}[C_k]$ (with samples of C_k computed by global optimization), the objective ℓ_k requires an estimate of its logarithm; a naive application of Jensen's inequality $-\log \mathbb{E}[C_k] \geq \mathbb{E}[-\log C_k]$ results in an *over*-estimator (as does the IWAE estimator), whereas we require a lower bound. Following Poole et al. (2019), we underestimate $-\log(x)$ by its linearization around some parameter $\alpha > 0$, resulting in the following lower bound objective:

$$\tilde{\ell}_k(\theta) := \mathbb{E}[-\log u_\theta(X)] - \mathbb{E}[C_k]/\alpha - \log \alpha + 1. \tag{8}$$

$\ell_k(\theta)$ can finally be estimated by sample averages, and yields a lower bound on the optimal intercept $F(\lambda)$ by the chain of inequalities, $\tilde{\ell}_k(\theta) \leq \ell_k(\theta) \leq \ell(\theta) \leq F(\lambda)$. A trained model u_{θ^*} then yields an R-D lower bound, $R_L(D) = -\lambda D + \tilde{\ell}_k(\theta^*)$. We give the detailed algorithm in Appendix A.4.

5 Related Work

Machine Learning: The past few years have seen significant progress in applying machine learning to data compression. For lossless compression, explicit likelihood models (van den Oord et al., 2016; Hoogeboom et al., 2019) directly lead themselves to entropy coding, and bits-back coding techniques are actively being developed for efficient compression with latent variable models (Townsend et al., 2019; Kingma et al., 2019; Ho et al., 2019; Ruan et al., 2021). In the lossy domain, Theis et al. (2017); Ballé et al. (2017) showed that a particular type of VAE can be trained to perform data compression using the same objective Eq. 3. The variational distributions in such a model have shape restrictions to simulate quantization and entropy coding (Ballé et al., 2017). Our upper bound is directly inspired by this line of work, and suggests that such a model can in principle compute the source R-D function when equipped with sufficiently expressive variational distributions and a "rich enough" latent space (see explanation in Sec. 3). We note however not all compressive autoencoders admit a probabilistic formulation (Theis et al., 2017); recent work has found training with hard quantization to improve compression performance (Minnen & Singh, 2020), and methods have been developed (Agustsson & Theis, 2020; Yang et al., 2020b) to reduce the gap between approximate quantization at training time and hard quantization at test time. Departing from compressive autoencoders, Yang et al. (2020c) and Flamich et al. (2020) applied regular Gaussian VAEs to data compression to exploit the flexiblility of variable-width variational posterior distributions. The REC algorithm from Flamich et al. (2020) can in theory transmit a sample of $Q_{\hat{X}|X}$ with a rate equal to the rate term of the NELBO-like Eq. 3, but comes with non-negligible overhead. Our experiment in Sec. 6.3 points to the theoretically possible gain in image compression performance from this approach. Agustsson & Theis (2020) proved the general difficulty of this approach without assumptions on the form of $Q_{\hat{X}|X}$, and showed that the particular case of a uniform $Q_{\hat{X}|X}$ leads to efficient implementation based on dithered quantization.

Information theory has also broadly influenced unsupervised learning (Alemi et al., 2018; Poole et al., 2019) and representation learning (Tishby et al., 2000). The Information Bottleneck method (Tishby et al., 2000) was directly motivated as a more general R-D theory and borrows from the BA algorithm. Alemi et al. (2018) analyzed the relation between generative modeling and representation learning with a similar R-D Lagrangian to Eq. 2, but used an abstract, model-dependent distortion $\rho(\hat{x}, x) := -\log p(x|\hat{x})$ with an arbitrary $\hat{\mathcal{X}}$ and without considering a data compression task. Recently, Huang et al. (2020) proposed to evaluate decoder-based generative models by computing a restricted version of $R_\omega(D)$ (with $Q_{\hat{x}}$ fixed to a Gaussian); our result in Sec. 3 ($R_\omega(D) \geq R(D)$) gives a principled way to interpret and compare these model-dependent R-D curves.

Information Theory: While the BA algorithm (Blahut, 1972; Arimoto, 1972) solves the problem of computing the $R(D)$ of a known discrete source, no tools currently exist for the general and unknown case. Riegler et al. (2018) share our goal of computing $R(D)$ of a general source, but still require the source being known analytically and supported on a known reference measure. Harrison & Kontoyiannis (2008) consider the same setup as ours of estimating $R(D)$ of an unknown source from samples, but focus on purely theoretical aspects assuming prefect optimization. They prove statistical consistency of such estimators for a general class of alphabets and distortion metrics, assuring that our stochastic bounds on $R(D)$ optimized from data samples, with unlimited computation and samples, can converge to the true $R(D)$. Perhaps closest in spirit to our work is Gibson (2017), who estimate lower bounds on $R(D)$ of speech and video using Gaussian autoregressive models of the source. The correctness of their bounds therefore depend on the modeling assumptions.

A variational lower bound on $R(D)$ was already proposed by Shannon (1959), and later extended (Berger, 1971) to the present version similar to Theorem 4.1. In the finite-alphabet case, the maximization characterization of $R(D)$ directly follows from taking the Lagrange dual of its standard definition in Eq. 1 (which is itself a convex optimization problem); the dual problem can then be solved by convex optimization (Chiang & Boyd, 2004), but faces the same limitations as the BA algorithm. A rigorous proof of Thereom 4.1 for general alphabets is more involved, and is based on repeated applications of information divergence inequalities (Csiszár, 1974; Gray, 2011).

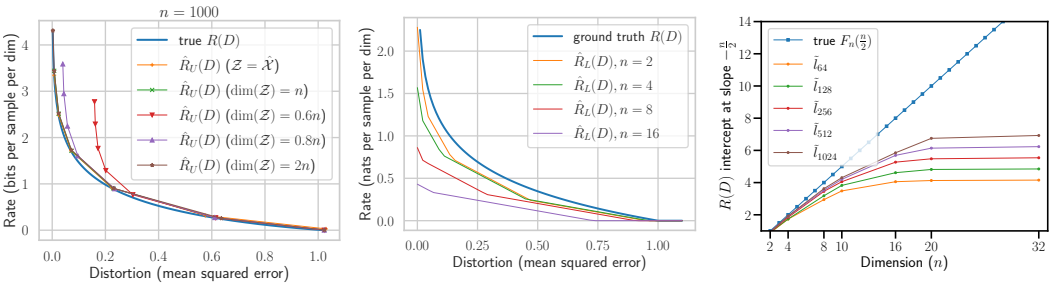

Figure 2: **Left**: R-D upper bounds on a randomly generated 1000-dimensional Gaussian source; **Middle**: R-D lower bounds on standard Gaussian sources with increasing dimensions; **Right**: R-axis intercept estimates at $\lambda = \frac{n}{2}$ from our lower bound algorithm, trained with increasing n and k.

6 Experiments

We experiment on three types of sources. On **random Gaussian sources**, we show that our upper bound algorithm can converge to the *exact* R-D function; our lower bounds, however, become increasingly loose in higher dimensions, and our experiments with varying k offers some insight into this issue. On **banana-shaped sources**, which do not have analytical $R(D)$, we obtain tight sandwich bounds. On the 2-D banana source (Ballé et al., 2021), our bounds lie closely to the R-D function produced by the BA algorithm. We also randomly map the 2-D banana source to higher dimensions (up to $n = 1000$) and still obtain tight sandwich bounds, where the BA algorithm is no longer infeasible. This shows that our lower bound algorithm can scale to higher dimensional data with low *effective* dimension. Finally, we run large-scale experiments on **natural images** to establish bounds on its R-D function. While the effective data dimension is still too high for us to obtain useful lower bounds, we focus on upper bounds. We base our upper bounds on the autoencoder architecture of a state-of-the-art image compression model, as well as a ResNet-VAE. Our upper bounds suggest a possible 1 PSNR improvement in compression quality at various bitrates.

6.1 Gaussian sources

The diagonal Gaussian is one of the only sources with a known analytical R-D function. As a first test of our algorithms, we apply them to randomly generated Gaussians in increasingly high dimensions.

For the upper bound algorithm, we chose $Q_{\hat{X}}$ and $Q_{\hat{X}|X}$ to be factorized Gaussians with learned parameters. We also consider optimizing the variational distributions in a latent space $\mathcal{Z}$ with varying dimensions, using an MLP decoder to map from $\mathcal{Z}$ to $\hat{\mathcal{X}}$ (see Sec. 3). The resulting R-D upper bounds are shown in Fig. 2-Left for a $n = 1000$ dimensional Gaussian, which shows our R-D upper bound (**yellow** curve) accurately recovers the analytical R-D function. The models with decoders and latent spaces show varied performance; the upper bounds become looser in the low-distortion regime when $\dim(\mathcal{Z}) < n$ (**red** and **purple** curves), while remaining tight when $\dim(\mathcal{Z}) \geq n$ (**green**, **brown**). The results are similar to sources in lower-dimension, and up to $n = 10000$ (beyond which we ran out of GPU memory for the larger models). We observe the best performance with a linear (or identity) decoder and a simple linear encoder (for the parameters of $Q_{\hat{X}|X}$); using deeper networks with nonlinear activation required more training iterations for SGD to converge, and often to a poorer upper bound. In fact, for a diagonal Gaussian source, it can be shown analytically that the optional $Q^*_{Y|X=x}$ is a Gaussian whose mean depends linearly on x, and an identity (no) decoder is optimal.

For the lower bound algorithm, we parameterize $\log u$ by a 2-layer MLP, and study the effect of source dimension n and the number of samples k used in our estimator C_k (and objective $\tilde{\ell}_k$). To simplify comparison of R-D results across different source dimensions, here we consider standard Gaussian sources, whose R-D curve does not vary with n if we scale the rate by $\frac{1}{n}$ (i.e., rate per sample per dimension); the results on randomly generated Gaussians are similar. First, we fix $k = 1024$ and examine the resulting lower bounds; as shown in Fig. 2-Middle, the bounds quickly become loose with increasing source dimension. This is likely due to the over-estimation bias of our estimator C_k for the sup-partition function, which causes under-estimation in the objective $\tilde{\ell}_k$. While C_k is defined similarly to an M-estimator (Van der Vaart, 2000), it is hard to analyze its convergence behavior in

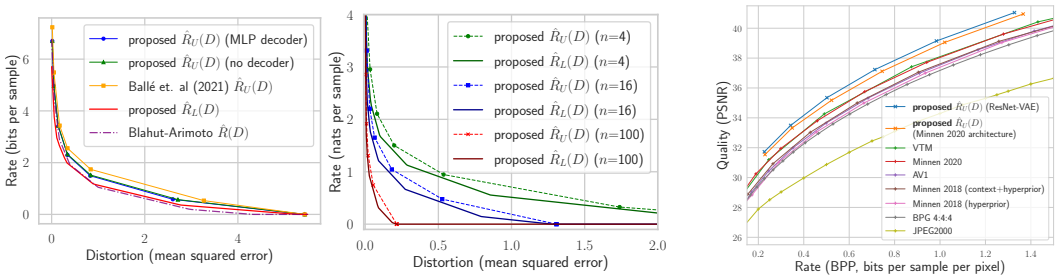

Figure 3: **Left, Middle**: Proposed R-D sandwich bounds applied to the 2-D banana source, and its random projection to higher dimensions, respectively. **Right**: Quality-rate curves of ours and state-of-the-art image compression methods on Kodak (1993), corresponding to R-D upper bounds.

general, as it depends on the function u being learned. Empirically, we observe the bias of C_k is amplified by an increase in the source dimension n or λ, such that an increasingly large k (possibly exponential in n) is required to effectively train a u model from random initialization. To illustrate this, we ran a series of experiments with k ranging from 64 to 1024 on increasingly high dimensional sources, each time setting $\lambda = \frac{n}{2}$. For an n-dimensional Gaussian, the true R-intercept, $F_n(\lambda)$, has an analytical form; in particular, $F_n(\frac{n}{2}) = \frac{n}{2}$. In Fig. 2-Right, we plot the final objective estimate, $\tilde{\ell}_k$, using the converged MLP model, one for each k and n. As we can see, the maximum achievable $\tilde{\ell}_k$ plateaus to the value $\log k$ as we increase n, and for sufficiently high dimension (e.g., $n = 20$ here), doubling k only brings out a constant ($\log 2$) improvement in the objective. This issue is related to the number of samples k needed to reliably estimate the mode of a distribution with a kernel density estimator, and we expand on this in the Appendix.

6.2 Banana-Shaped Sources

The quickly deteriorating lower bound on the previous Gaussian experiment makes our goal of sandwiching the $R(D)$ of a general source seem hopeless. In this experiment, we demonstrate that we can, in fact, obtain tightly matching lower and upper bounds on high dimensional data with sufficiently low *effective* dimension. We borrow the 2D banana-shaped source from Ballé et al. (2021), obtained by a nonlinear transform of a 2D Gaussian. First, we establish sandwich bounds on the original 2D source, and illustrate the effect of different variational distributions on the upper bounds in Fig. 3-Left. Our upper bound model (**blue** curve) uses the same autoencoder architecture as the one in Ballé et al. (2021) (**orange**), but replaces its factorized $Q_{\hat{x}}$ and uniform $Q_{\hat{X}|X}$ with normalizing flows, resulting in a tighter bound. Removing the decoder required deeper flow transforms to achieve the same bound (**green**). Our lower bound (**red**) here agrees with the R-D function of the discretized source from the BA algorithm (**purple**), and hugs the upper bounds in low and high distortion regimes. Next, we map the 2D banana source to $\mathbb{R}^n$ with a randomly sampled $n \times 2$ matrix. Fig. 3-Middle shows that we still obtain tight sandwich bounds in higher dimensions (we verified this for n up to 1000; see figure in Appendix), where the BA algorithm is infeasible to apply. Unlike in the Gaussian experiment, where increasing dimension required (seemingly exponentially) larger k for a reasonable lower bound, here a constant $k = 2048$ worked well for all n we tested. The key difference compared to the Gaussian is that despite the high dimension, the data here posses considerable structure and still lies on a low-dimension manifold [2], a property conjured to hold for real-world data (Goodfellow et al., 2016).

6.3 Natural Images

For signals such as images or video, which can have arbitrarily high spatial dimension, it is not immediately clear how to define i.i.d. samples. But since our focus is on image compression, we follow the common practice of extracting random 256×256-pixel crops from high resolution images as in learned image compression research (Ballé et al., 2017), noting that current methods cannot effectively exploit spatial redundancies larger than this scale (Minnen & Singh, 2020). As a

[2]Indeed, the standard Gaussian attains Shannon's $R(D)$ upper bound for any source with a density under the squared error distortion (Shannon, 1959), and is in this sense the hardest continuous source to compress.

representative dataset of natural images, we used images from the COCO 2017 (Lin et al., 2014) train set that are larger than 512×512, and randomly downsampled them to remove compression artifacts.

Instead of working with variational distributions over the set of pixel values $\hat{\mathcal{X}} = \mathcal{X} = \{0, 1, ..., 255\}^n$, which would require specialized tools for representing and optimizing high-dimensional discrete distributions (Salimans et al., 2017; Hoogeboom et al., 2019; Maddison et al., 2016; Jang et al., 2016), for simplicity we parameterize the variational distributions in Euclidean latent spaces (see Sec. 3). We borrow the convolutional autoencoder architecture of a state-of-the-art image compression model (Minnen & Singh, 2020), and set the variational distributions to be diagonal Gaussians with learned means and variances (we still use the deep factorized hyperprior, but drop the shape restriction). Inspired by Vahdat & Kautz (2020), we also constructed a ResNet-VAE model (Kingma et al., 2016) with 6 layers of latent variables. For the two models, $\dim(\mathcal{Z}) \approx 0.28 \dim(\mathcal{X})$ and $\dim(\mathcal{Z}) \approx 0.66 \dim(\mathcal{X})$, respectively. We trained models with various λ and mean-squared error (MSE) distortion, and evaluated them on the Kodak (1993) and Tecnick (Asuni & Giachetti, 2014) datasets. Following image compression conventions, we report the rate in bits-per-pixel, and the quality (i.e., negative distortion) in PSNR averaged over the images for each (λ, model) pair [3]. The resulting *quality-rate* (Q-R) curves can be interpreted as giving upper bounds on the R-D functions of the image-generating distributions. We plot them in Fig. 3, along with the Q-R performance (in actual bitrate) of various hand-engineered and learned image compression methods (Ballé et al., 2017; Minnen et al., 2018; Minnen & Singh, 2020), for the Kodak dataset (see similar results on Tecnick in Appendix 4). Our β-VAE model based on (Minnen & Singh, 2020) (**orange**) lies on average 0.85 PSNR higher than the state-of-the-art base compressive autoencoder, and contrary to our expectation, we did not obtain drastically improved bounds by using normalizing flow instead of Gaussian $Q_{\hat{X}|X}$ distributions. By using a deeper latent architecture, our new model gives a higher Q-R curve (**blue**). We leave it to future work to investigate which choice of autoencoder architecture, latent space, and variational distributions are most effective, as well as how the R-D performance of such a β-VAE can be realized by an actual compression algorithm (see discussions in Sec. 5).

7 Discussions

In this work, we used machine learning techniques to computationally bound the rate-distortion function of a data source, a key quantity that characterizes the fundamental performance limit of all lossy compression algorithms, but is largely unknown. Departing from prior work in the information theory community (Gibson, 2017; Riegler et al., 2018), our approach applies broadly to general data sources and requires only i.i.d. samples, making it more suitable for real-world application.

Our upper bound method is a gradient descent version of the classic Blahut-Arimoto algorithm, and closely relates to (and extends) variational autoencoders from neural lossy compression research. Our lower bound optimizes a dual formulation of the R-D function, which has been known for some time but seen little application outside of theoretical work. Due to difficulties involving global optimization, our lower bound method currently requires a squared error distortion for tractability in the continuous case, and only yields useful bounds on data sources with a low *effective* dimension.

To properly interpret bounds on the R-D function, we emphasize that the significance of the R-D function is two-fold: 1. for a given distortion tolerance D, no coding procedure can operate with a rate less than $R(D)$, and that 2. this rate is asymptotically achievable by some (potentially expensive) procedure. Therefore, a lower bound makes a universal statement about what kind of rate-distortion performance is "too good to be true". The story is more subtle for the upper bound, due to the asymptotic nature of $R(D)$. The achievability proof relies on a random coding procedure that jointly compresses multiple data samples in increasingly long blocks (Shannon, 1959). When compressing at a finite block length b (e.g., $b = 1$ when compressing individual images), $R(D)$ is generally no longer achievable, due to a rate overhead that scales like $b^{-\frac{1}{2}}$ (Kostina & Verdú, 2012). Extending our work to the case of compression with finite-block lengths could be an impactful future direction.

Finally, it would be interesting to extend our approach to perceptually more relevant distortions (Wang et al., 2003) and non-i.i.d. data.

[3] Technically, to compute an R-D upper bound with MSE ρ, the distortion needs to be evaluated by averaging MSE (instead of PSNR) on samples.

ETHICS AND REPRODUCIBILITY STATEMENTS

As our work deals with theoretical aspects of lossy compression, we are not aware of any resulting ethical implications. Our code and instructions for reproducing the datasets and experimental results can be found in this anonymous repo https://drive.google.com/drive/folders/1uI3q7sVU6T8iF2PoAjNXSAwVHpYLeyHC?usp=sharing.

REFERENCES

E. Agustsson and L. Theis. Universally Quantized Neural Compression. In *Advances in Neural Information Processing Systems 33*, 2020.

Eirikur Agustsson, David Minnen, Nick Johnston, Johannes Balle, Sung Jin Hwang, and George Toderici. Scale-space flow for end-to-end optimized video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8503–8512, 2020.

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken ELBO. In *International Conference on Machine Learning*, pp. 159–168. PMLR, 2018.

Suguru Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

N. Asuni and A. Giachetti. TESTIMAGES: A large-scale archive for testing visual devices and basic image processing algorithms (SAMPLING 1200 RGB set). In *STAG: Smart Tools and Apps for Graphics*, 2014. URL https://sourceforge.net/projects/testimages/files/OLD/OLD_SAMPLING/testimages.zip.

J. Ballé, V. Laparra, and E. P. Simoncelli. End-to-end Optimized Image Compression. In *International Conference on Learning Representations*, 2017.

J. Ballé, P. A. Chou, D. Minnen, S. Singh, N. Johnston, E. Agustsson, S. J. Hwang, and G. Toderici. Nonlinear transform coding. *IEEE Trans. on Special Topics in Signal Processing*, 15, 2021.

T Berger. Rate distortion theory, a mathematical basis for data compression (prentice-hall. *Inc. Englewood Cliffs, New Jersey*, 1971.

Toby Berger and Jerry D Gibson. Lossy source coding. *IEEE Transactions on Information Theory*, 44(6):2693–2723, 1998.

R. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972. doi: 10.1109/TIT.1972.1054855.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Miguel A. Carreira-Perpinan. Mode-finding for mixtures of gaussian distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1318–1323, 2000.

Miguel A. Carreira-Perpinan. Gaussian mean-shift is an em algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):767–776, 2007. doi: 10.1109/TPAMI.2007.1057.

Miguel A. Carreira-Perpinan. How many modes can a Gaussian mixture have, 2020. URL https://faculty.ucmerced.edu/mcarreira-perpinan/research/GMmodes.html.

Mung Chiang and Stephen Boyd. Geometric programming duals of channel capacity and rate distortion. *IEEE Transactions on Information Theory*, 50(2):245–258, 2004.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*, volume 2. John Wiley & Sons, 2006.

I. Csiszar. On the computation of rate-distortion functions (corresp.). *IEEE Transactions on Information Theory*, 20(1):122–124, 1974. doi: 10.1109/TIT.1974.1055146.

Imre Csiszár. On an extremum problem of information theory. *Studia Scientiarum Mathematicarum Hungarica*, 9, 01 1974.

G. Flamich, M. Havasi, and J. M. Hernández-Lobato. Compressing Images by Encoding Their Latent Representations with Relative Entropy Coding, 2020. Advances in Neural Information Processing Systems 34.

Brendan J. Frey and Geoffrey E. Hinton. Efficient stochastic source coding and an application to a bayesian network source model. *The Computer Journal*, 40(2_and_3):157–165, 1997.

Jerry Gibson. Rate distortion functions and rate distortion function lower bounds for real-world sources. *Entropy*, 19(11):604, 2017.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Robert M Gray. *Entropy and information theory*. Springer Science & Business Media, 2011.

Matthew T. Harrison and Ioannis Kontoyiannis. Estimation of the rate–distortion function. *IEEE Transactions on Information Theory*, 54(8):3757–3762, Aug 2008. ISSN 0018-9448. doi: 10.1109/tit.2008.926387. URL http://dx.doi.org/10.1109/TIT.2008.926387.

J. Hayes, A. Habibi, and P. Wintz. Rate-distortion function for a gaussian source model of images (corresp.). *IEEE Transactions on Information Theory*, 16(4):507–509, 1970. doi: 10.1109/TIT.1970.1054496.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *Iclr*, 2(5):6, 2017.

Jonathan Ho, Evan Lohn, and Pieter Abbeel. Compression with flows via local bits-back coding. In *Advances in Neural Information Processing Systems*, pp. 3874–3883, 2019.

Emiel Hoogeboom, Jorn Peters, Rianne van den Berg, and Max Welling. Integer discrete flows and lossless compression. In *Advances in Neural Information Processing Systems*, pp. 12134–12144, 2019.

Sicong Huang, Alireza Makhzani, Yanshuai Cao, and Roger Grosse. Evaluating lossy compression rates of deep generative models. *International Conference on Machine Learning*, 2020.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

D. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pp. 4743–4751, 2016.

Friso H Kingma, Pieter Abbeel, and Jonathan Ho. Bit-swap: Recursive bits-back coding for lossless compression with hierarchical latent variables. *arXiv preprint arXiv:1905.06845*, 2019.

Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, Nov 2021. ISSN 1939-3539. doi: 10.1109/tpami.2020.2992934. URL http://dx.doi.org/10.1109/TPAMI.2020.2992934.

Kodak. PhotoCD PCD0992, 1993. URL http://r0k.us/graphics/kodak/.

Victoria Kostina. When is shannon's lower bound tight at finite blocklength? In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 982–989. IEEE, 2016.

Victoria Kostina and Sergio Verdú. Fixed-length lossy compression in the finite blocklength regime. *IEEE Transactions on Information Theory*, 58(6):3309–3338, 2012.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Jasper C. H. Lee, Jerry Li, Christopher Musco, Jeff M. Phillips, and Wai Ming Tai. Finding the mode of a kernel density estimate, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in Neural Information Processing Systems*, 29, 2016.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Matt Mahoney. Rationale for a large text compression benchmark. *Retrieved (Oct. 1st, 2021) from: http://mattmahoney.net/dc/rationale.html*, 2009.

D. Minnen and S. Singh. Channel-wise autoregressive entropy models for learned image compression. In *IEEE International Conference on Image Processing (ICIP)*, 2020.

D. Minnen, J. Ballé, and G. D. Toderici. Joint Autoregressive and Hierarchical Priors for Learned Image Compression. In *Advances in Neural Information Processing Systems 31*. 2018.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pp. 2338–2347, 2017.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5171–5180. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/poole19a.html.

Farzad Rezaei, NU Ahmed, and Charalambos D Charalambous. Rate distortion theory for general sources with potential application to image compression. *International Journal of Applied Mathematical Sciences*, 3(2):141–165, 2006.

Erwin Riegler, Günther Koliander, and Helmut Bölcskei. Rate-distortion theory for general sets and measures, 2018.

Yangjun Ruan, Karen Ullrich, Daniel Severo, James Townsend, Ashish Khisti, Arnaud Doucet, Alireza Makhzani, and Chris J Maddison. Improving lossless compression rates via monte carlo bits-back coding. *arXiv preprint arXiv:2102.11086*, 2021.

Ruslan Salakhutdinov. Learning deep generative models. *Annual Review of Statistics and Its Application*, 2:361–385, 2015.

T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. PixelCNN++: A pixelcnn implementation with discretized logistic mixture likelihood and other modifications. In *International Conference on Learning Representations*, 2017.

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

CE Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec., March 1959*, 4:142–163, 1959.

L. Theis, W. Shi, A. Cunningham, and F. Huszár. Lossy Image Compression with Compressive Autoencoders. In *International Conference on Learning Representations*, 2017.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

James Townsend, Tom Bird, and David Barber. Practical lossless compression with latent variables using bits back coding. *arXiv preprint arXiv:1901.04866*, 2019.

Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *Neural Information Processing Systems (NeurIPS)*, 2020.

A. van den Oord, N. Kalchbrenner, O. Vinyals, A. Graves L. Espeholt, and K. Kavukcuoglu. Conditional Image Generation with PixelCNN Decoders. In *Advances in Neural Information Processing Systems 29*, pp. 4790–4798, 2016.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pp. 1398–1402 Vol.2, 2003. doi: 10.1109/ACSSC.2003.1292216.

Ian H Witten, Radford M Neal, and John G Cleary. Arithmetic coding for data compression. *Communications of the ACM*, 30(6):520–540, 1987.

Ruihan Yang, Yibo Yang, Joseph Marino, and Stephan Mandt. Hierarchical autoregressive modeling for neural video compression. In *International Conference on Learning Representations*, 2020a.

Yibo Yang, Robert Bamler, and Stephan Mandt. Improving inference for neural image compression. In *Neural Information Processing Systems (NeurIPS), 2020*, 2020b.

Yibo Yang, Robert Bamler, and Stephan Mandt. Variational Bayesian Quantization. In *International Conference on Machine Learning*, 2020c.

Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on information theory*, 23(3):337–343, 1977.

A APPENDIX

A.1 TECHNICAL DEFINITIONS

In this work we consider a the source to be represented by a random variable $X : \Omega \to \mathcal{X}$, i.e., a measurable function on an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and P_X is the image measure of $\mathbb{P}$ under X. We suppose the source and reproduction spaces are standard Borel spaces, $(\mathcal{X}, \mathcal{A}_{\mathcal{X}})$ and $(\hat{\mathcal{X}}, \mathcal{A}_{\hat{\mathcal{X}}})$, equipped with sigma-algebras $\mathcal{A}_{\mathcal{X}}$ and $\mathcal{A}_{\hat{\mathcal{X}}}$, respectively.

Conditional distribution The notation $Q_{\hat{X}|X}$ denotes an arbitrary conditional distribution (also known as a Markov kernel), i.e., it satifies

1. For any $x \in \mathcal{X}$, $Q_{\hat{X}|X=x}(\cdot)$ is a probability measure on $\hat{\mathcal{X}}$;
2. For any measurable set $B \in \mathcal{A}_{\hat{\mathcal{X}}}$, $x \to Q_{\hat{X}|X=x}(B)$ is a measurable function on $\mathcal{X}$.

Induced joint and marginal measures Given a source distribution P_X, each test channel distribution $Q_{\hat{X}|X}$ defines a joint distribution $P_X Q_{\hat{X}|X}$ on the product space $\mathcal{X} \times \hat{\mathcal{X}}$ (equipped with the usual product sigma algebra, $\mathcal{A}_{\mathcal{X}} \times \mathcal{A}_{\hat{\mathcal{X}}}$) as follows:

$$P_X Q_{\hat{X}|X}(E) := \int_{\mathcal{X}} P_X(dx) \int_{\hat{\mathcal{X}}} \mathbf{1}\{(x, \hat{x}) \in E\} Q_{\hat{X}|X=x}(d\hat{x}),$$

for all measurable sets $E \in \mathcal{A}_{\mathcal{X}} \times \mathcal{A}_{\hat{\mathcal{X}}}$. The induced $\hat{x}$-marginal distribution $P_{\hat{X}}$ is then defined by

$$P_{\hat{X}}(B) = \int_{\mathcal{X}} Q_{\hat{X}|X=x}(B) P_X(dx),$$

for all measurable sets $\forall B \in \mathcal{A}_{\hat{\mathcal{X}}}$.

KL Divergence We use the general definition of Kullback-Leibler (KL) divergence between two probability measures P, Q defined on a common measurable space:

$$D(P\|Q) := \begin{cases} \int \log \frac{dP}{dQ} dP, & \text{if } P \ll Q \\ \infty, & \text{otherwise.} \end{cases}$$

$P \ll Q$ denotes that P is absolutely continuous w.r.t. Q (i.e., for all measurable sets E, $Q(E) = 0 \implies P(E) = 0$). $\frac{dP}{dQ}$ denotes the Radon-Nikodym derivative of P w.r.t. Q; for discrete distributions, we can simply take it to be the ratio of probability mass functions; and for continuous distributions, we can simply take it to be the ratio of probability density functions.

Mutual Information Given P_X and $Q_{\hat{X}|X}$, the mutual information $I(X;Y)$ is defined as

$$I(X; \hat{X}) := D(P_X Q_{\hat{X}|X} \| P_X \otimes P_{\hat{X}}) = \mathbb{E}_{x \sim P_X}[D(Q_{\hat{X}|X=x} \| P_{\hat{X}})],$$

where $P_{\hat{X}}$ is the $\hat{x}$-marginal of the joint $P_X Q_{\hat{X}|X}$, $P_X \otimes P_{\hat{X}}$ denotes the usual product measure, and $D(\cdot\|\cdot)$ is the KL divergence.

For the mutual information upper bound, it's easy to show that

$$\mathcal{I}_U(Q_{\hat{X}|X}, Q_{\hat{X}}) := \mathbb{E}_{x \sim P_X}[KL(Q_{\hat{X}|X=x} \| Q_{\hat{X}})] = I(X; \hat{X}) + KL(P_{\hat{X}} \| Q_{\hat{X}}), \tag{9}$$

so the bound is tight when $P_{\hat{X}} = Q_{\hat{X}}$.

Obtaining $R(D)$ through the Lagrangian. For each $\lambda \geq 0$, we define the Lagrangian by incorporating the distortion constraint in the definition of $R(D)$ through a linear penalty:

$$\mathcal{L}(Q_{\hat{X}|X}, \lambda) := I(X; \hat{X}) + \lambda \mathbb{E}_{P_X Q_{\hat{X}|X}}[\rho(X, \hat{X})], \tag{10}$$

and define its infimum w.r.t. $Q_{\hat{X}|X}$ by the function

$$F(\lambda) := \inf_{Q_{\hat{X}|X}} I(X;\hat{X}) + \lambda \mathbb{E}[\rho(X,\hat{X})]. \tag{11}$$

Geometrically, $F(\lambda)$ is the maximum of the R-axis intercepts of straight lines of slope $-\lambda$, such that they have no point above the $R(D)$ curve (Csiszár, 1974).

Define $D_{min} := \inf\{D' : R(D') < \infty\}$. Since $R(D)$ is convex, for each $D > D_{min}$, there exists a $\lambda \geq 0$ such that the line of slope $-\lambda$ through $(D, R(D))$ is tangent to the $R(D)$ curve, i.e.,

$$R(D') + \lambda D' \geq R(D) + \lambda(D) = F(\lambda), \quad \forall D'.$$

When this occurs, we say that λ *is associated to* D.

Consequently, the $R(D)$ curve is then the envelope of lines with slope $-\lambda$ and R-axis intercept $F(\lambda)$. Formally, this can be stated as:

Lemma A.1 (Csiszár (1974, Lemma 1.2); Gray (2011, Lemma 9.7))**.** *For every distortion tolerance $D > D_{min}$, where $D_{min} := \inf\{D' : R(D') < \infty\}$, it holds that*

$$R(D) = \max_{\lambda \geq 0} F(\lambda) - \lambda D \tag{12}$$

We can draw the following conclusions:

1. *For each $D > D_{min}$, the maximum above is attained iff λ is associated to D.*
2. *For a fixed λ, if $Q^*_{Y|X}$ achieves the minimum of $\mathcal{L}(\cdot, \lambda)$, then λ is associated to the point $(\mathcal{I}(Q^*_{Y|X}), \rho(Q^*_{Y|X}))$; i.e., there exists a line with slope $-\lambda$ that is tangent to the $R(D)$ curve at $(\mathcal{I}(Q^*_{Y|X}), \rho(Q^*_{Y|X}))$.*

A.2 Full Version of Theorem 4.1

Theorem A.2. *(We use the following version from (Kostina, 2016).) Suppose that the following basic assumptions are satisfied.*

1. *$R(D)$ is finite for some D, i.e., $D_{min} := \inf\{D : R(D) < \infty\} < \infty$;*
2. *The distortion metric ρ is such that there exists a finite set $E \subset \hat{\mathcal{X}}$ such that*

$$\mathbb{E}[\min_{\hat{x} \in E} \rho(X, \hat{x})] < \infty$$

Then, for each $D > D_{min}$, it holds that

$$R(D) = \max_{g(x), \lambda} \{\mathbb{E}[-\log g(X)] - \lambda D\} \tag{13}$$

where the maximization is over $g(x) \geq 0$ and $\lambda \geq 0$ satisfying the constraint

$$\mathbb{E}\left[\frac{\exp(-\lambda \rho(X, \hat{x}))}{g(X)}\right] = \int \frac{\exp(-\lambda \rho(x, \hat{x}))}{g(x)} dP_X(x) \leq 1, \forall y \in \hat{\mathcal{X}} \tag{14}$$

Note: the basic assumption 2 is trivially satisfied when the distortion ρ is bounded from above; the maximization over $g(x) \geq 0$ can be restricted to only $1 \geq g(x) \geq 0$. Unless stated otherwise, we use $\log$ base e in this work, so the $R(D)$ above is in terms of nats.

Theorem 4.1 in the main text in terms of $F(\lambda)$ is equivalent to the above by fixing a D and λ value. In our earlier attempts of the lower bound algorithm, we used the full R-D duality formulation as in Theorem A.2, producing $R(D)$ lower bounds by fixing a D and optimizing over λ for various D values. However, the algorithm often diverged due to drastically changing λ. We avoid this by using current formulation of fixing λ in the optimization, and produce $R(D)$ lower bound by ranging over λ (instead of D).

A.3 THEORETICAL RESULTS

Proof that $R_\omega(D) \geq R(D)$ In learned compression with VAEs, a latent space $\mathcal{Z}$ is introduced, along with a learned decoder transform $\omega : \mathcal{Z} \to \hat{\mathcal{X}}$. This induces a learned distortion function $\rho_\omega : \mathcal{X} \times \mathcal{Z} \to [0, \infty), \rho_\omega(x, z) = \rho(x, \omega(Z))$, where ρ is the "base" distortion.

Below we show that the rate-distortion function under a learned decoder,

$$R_\omega(D) = \inf_{Q_{Z|X}:\mathbb{E}[\rho_\omega(X,Z)]\leq D} I(X;Z) = \inf_{Q_{Z|X}:\mathbb{E}[\rho(X,\omega(Z))]\leq D} I(X;Z),$$

is an upper bound on the source $R(D)$:

$$R(D) = \inf_{Q_{\hat{X}|X}:\mathbb{E}[\rho(X,\hat{X})]\leq D} I(X;\hat{X}).$$

Moreover, a sufficient condition for $R_\omega(D) = R(D)$ is for ω to be injective, and that the range of ω must cover the support of $Q_{\hat{X}|X=x}$ for P_X-almost all x; e.g., this is satisfied by a bijective ω with $\mathcal{Z} = \hat{\mathcal{X}}$.

Proof. Fix D. Take any admissible conditional distribution $Q_{Z|X}$ that satisfies $\mathbb{E}[\rho(X, \omega(Z))] \leq D$ in the definition of $R_\omega(D)$. Define a new kernel $Q_{\hat{X}|X}$ between $\mathcal{X}$ and $\hat{\mathcal{X}}$ by $Q_{\hat{X}|X=x} := Q_{Z|X=x} \circ \omega^{-1}, \forall x \in \mathcal{X}$, i.e., $Q_{\hat{X}|X=x}$ is the pushforward/image measure of $Q_{Z|X=x}$ induced by ω. Applying data processing inequality to the Markov chain $X \xrightarrow{Q_{Z|X}} Z \xrightarrow{\omega} \hat{X}$, we have $I(X;Z) \geq I(X;\hat{X})$. Moreover, since $Q_{\hat{X}|X}$ is admissible in the definition of $R(D)$, i.e.,

$$\mathbb{E}_{P_X Q_{\hat{X}|X}}[\rho(X,\hat{X})] = \mathbb{E}_{P_X Q_{Z|X}}[\rho(X,\omega(Z))] \leq D$$

we therefore have

$$I(X;Z) \geq I(X;\hat{X}) \geq R(D).$$

Finally, since $I(X;Z) \geq R(D)$ holds for any admissible $Q_{Z|X}$, taking infimum over such $Q_{Z|X}$ gives $R_\omega(D) \geq R(D)$.

To prove $R_\omega(D) = R(D)$ when ω satisfies the sufficiency condition (injective, and $\mathrm{supp}(Q_{\hat{X}|X=x}) \subset \mathrm{range}(\omega)$ for P_X-almost all $x \in \mathcal{X}$), it suffices to show that $R(D) \geq R_\omega(D)$. We use the same argument as before. Take any admissible $Q_{\hat{X}|X}$ in the definition of $R(D)$. We can then construct a $Q_{Z|X}$ by the process $X \xrightarrow{Q_{\hat{X}|X}} Y \xrightarrow{\omega^{-1}} Z$, where the inverse function $\omega^{-1} : y \to z$ is well-defined thanks to ω being bijective (or at least the range of ω covers the support of $Q_{\hat{X}|X}$). Then by DPI we have $I(X;\hat{X}) \geq I(X;Z)$; furthermore we can check $Q_{Z|X}$ is admissible: $\mathbb{E}_{P_X Q_{Z|X}}[\rho(X,\omega(Z))] = \mathbb{E}_{P_X Q_{\hat{X}|X}}[\rho(X,\omega(\omega^{-1}(Y))] = \mathbb{E}_{P_X Q_{\hat{X}|X}}[\rho(X,\hat{X})] \leq D$. So $I(X;\hat{X}) \geq I(X;Z) \geq R_\omega(D)$. Taking infimum over $Q_{\hat{X}|X}$ concludes the proof.

□

Theorem A.3. *Let $X_1, X_2, ... \sim P_X$ be a sequence of i.i.d. random variables. Let $\psi : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}$ be a measurable function. For each k, define the random variable $C_k := \sup_{\hat{x}} \frac{1}{k} \sum_i \psi(X_i, \hat{x})$. Then*

1. *$\mathbb{E}[C_k] = \mathbb{E}_{X_1,...,X_k}[\sup_{\hat{x}} \frac{1}{k} \sum_i \psi(X_i, \hat{x})] \geq \sup_{\hat{x}} \mathbb{E}[\psi(X, \hat{x})] =: c$;*
2. *$\mathbb{E}[C_1] \geq \mathbb{E}[C_2] \geq ... \geq \mathbb{E}[C_k] \geq \mathbb{E}[C_{k+1}] \geq ... \sup_{\hat{x}} \mathbb{E}[\psi(X, \hat{x})] = c$;*
3. *If $\psi(x, \hat{x})$ is bounded and continuous in $\hat{x}$, and if $\hat{\mathcal{X}}$ is compact, then C_k converges to c almost surely (as well as in probability, i.e., $\lim_{k\to\infty} \mathbb{P}(|C_k - c| > \epsilon) = 0, \forall \epsilon > 0$), and $\lim_{k\to\infty} \mathbb{E}[C_k] = c$.*

Proof. We prove each in turn:

1. $\mathbb{E}[C_k] = \mathbb{E}[\sup_{\hat{x}} \frac{1}{k} \sum_i \psi(X_i, \hat{x})] \geq \sup_{\hat{x}} \mathbb{E}[\frac{1}{k} \sum_i \psi(X_i, \hat{x})] = \sup_{\hat{x}} \mathbb{E}[\psi(X, \hat{x})] = c$

2. First, note that $\mathbb{E}[C_1] \geq \mathbb{E}[C_k]$ since

$$\mathbb{E}[C_1] = \mathbb{E}[\sup_{\hat{x}} \psi(X_1, \hat{x})] = \mathbb{E}[\frac{1}{k}\sum_i \sup_{\hat{x}} \psi(X_i, \hat{x})] \geq \mathbb{E}[\sup_{\hat{x}} \frac{1}{k}\sum_i \psi(X_i, \hat{x})] = \mathbb{E}[C_k]$$

We therefore have

$$\begin{aligned}
\mathbb{E}[C_{k+1}] &= \mathbb{E}[\sup_{\hat{x}} \frac{1}{k+1}\sum_{i=1}^{k+1} \psi(X_i, \hat{x})] \\
&= \mathbb{E}[\sup_{\hat{x}} \{\frac{1}{k+1}\sum_{i=1}^{k} \psi(X_i, \hat{x}) + \frac{1}{k+1}\psi(X_{k+1}, \hat{x})\}] \\
&\leq \mathbb{E}[\sup_{\hat{x}} \{\frac{1}{k+1}\sum_{i=1}^{k} \psi(X_i, \hat{x})\} + \sup_{\hat{x}} \{\frac{1}{k+1}\psi(X_{k+1}, \hat{x})\}] \\
&= \frac{k}{k+1}\mathbb{E}[C_k] + \frac{1}{k+1}\mathbb{E}[C_1] \\
&\leq \mathbb{E}[C_k]
\end{aligned}$$

3. The proof for this resembles that of Theorem 1 in (Burda et al., 2015). We use standard arguments from probability theory and real analysis. Fix $\hat{x} \in \hat{\mathcal{X}}$, and consider the random variable $M_k = \frac{1}{k}\sum_{i=1}^{k} \psi(X_i, \hat{x})$. If ψ is bounded, then it follows from the Strong Law of Large Numbers that M_k converges to $\mathbb{E}[M_1] = \mathbb{E}[\psi(X, \hat{x})]$ almost surely; in other words, for every ω outside a set of measure zero,

$$\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} \psi(X_i(\omega)), \hat{x}) = \mathbb{E}[\psi(X(\omega), \hat{x})],$$

Then, for every such ω

$$\lim_{k\to\infty} \sup_{\hat{x}} \frac{1}{k}\sum_{i=1}^{k} \psi(X_i(\omega)), \hat{x}) = \sup_{\hat{x}} \lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} \psi(X_i(\omega)), \hat{x}) = \sup_{\hat{x}} \mathbb{E}[\psi(X(\omega), \hat{x})],$$

where we used the fact that the sequence of continuous functions $s_k(\hat{x}) := \frac{1}{k}\sum_{i=1}^{k} \psi(X_i(\omega)), \hat{x})$ converges pointwise to $s(\hat{x}) := \mathbb{E}[\psi(X(\omega), \hat{x})]$ on a compact set $\hat{\mathcal{X}}$, so s_k converges to s also uniformly, so we are allowed to exchange limit and supremum, i.e., $\lim_{k\to\infty} \sup_{\hat{x}} s_k(\hat{x}) = \sup_{\hat{x}} \lim_{k\to\infty} s_k(\hat{x}) = \sup_{\hat{x}} s(\hat{x})$. But the above equation precisely means that C_k converges to $c = \sup_{\hat{x}} \mathbb{E}[\psi(X, \hat{x})]$ almost surely. Therefore C_k also converges to c in probability, and $\lim_{k\to\infty} \mathbb{E}[C_k] = c$.

□

A.4 Proposed Lower Bound Algorithm

We give an outline of the algorithm in A.4. In our experiments, we only run an approximate version of the global optimization subroutine `compute_`C_k, by running hill-climbing only from the t centroids of the Gaussian mixture that have the t highest mixture weights, for $t << k$. We typically use $t = 10$ with $k = 1024$ in our experiments, which allows us to train efficiently. To report the final R-axis intercept for an R-D lower bound, we carry out the global optimization procedure exhaustively to guarantee correctness.

Algorithm 1: Proposed algorithm for estimating rate-distortion lower bound $R_L(D)$.

Requires: $\lambda > 0$, model u_θ (e.g., a neural network) parameterized by θ, and batch sizes k, m
while θ *not converged* **do**
  `// Estimate` $\mathbb{E}[C_k]$ `by averaging` m `samples of` C_k
  for $j \leftarrow 1$ **to** m **do**
    Draw k data samples, $\{x_1^j, ..., x_k^j\}$
    $\hat{x}^j, C_k^j =$ `compute_`$C_k(\theta, \lambda, \{x_1, ..., x_k\})$
  end
  Set $\mathbb{E}[C_k] = \frac{1}{m}\sum_{j=1}^m C_k^j = \frac{1}{m}\sum_{j=1}^m \frac{1}{k}\sum_{i=1}^k \exp\{-\lambda\rho(x_i^j, \hat{x}^j)\}/u_\theta(x_i^j)$;
  Repeat the same procedure and set $\alpha = \mathbb{E}[C_k]$ using separate m samples of C_k.
  Compute objective $\tilde{\ell}_k(\theta)$ and update θ by gradient ascent .
end
Subroutine `compute_`$C_k(\theta, \lambda, \{x_1, ..., x_k\})$:
  Compute the global optimum of $\phi(\hat{x}) := \frac{1}{k}\sum_{i=1}^k \exp\{-\lambda\rho(x_i, \hat{x})\}/u_\theta(x_i)$,
  $\hat{x}^* = \arg\max \phi(\hat{x})$
  Compute $\hat{C}_k = \phi(\hat{x}^*)$
  Return $(\hat{x}^*, \hat{C}_k)$

A.5 Further Experiment Details

A.5.1 Gaussians

For the randomly generated Gaussians, we sample each dimension of the mean uniformly randomly from $[-0.5, 0.5]$ and variance from $[0, 2]$.

We chose $Q_{\hat{X}}$ and $Q_{\hat{X}|X}$ to be factorized Gaussians; we let the mean and diagonal variance of $Q_{\hat{X}}$ be trainable parameters, and predict the mean and variance of $Q_{\hat{X}|X=x}$ by a one-layer fully connected network (MLP) with $2n$ output units and softplus activation for the n variance components.

We consider three choices of decoder ω: the identity function ("no decoder"), a one-layer network with no activation ("linear decoder"), and a two-layer MLP with leaky ReLU activation ("MLP dcoder"). The resulting R-D curves are shown in Fig 1 (a), (b), for $n = 10$ and 10000. As can be seen, the basic algorithm (without decoder) recovers the ground truth R-D function, as does the version with linear decoder function, independent of the source dimension. The use of a non-linear decoder, however, results in a looser upper bound, and the gap increases with the source dimension. We conjecture this is due to SGD being trapped in local minima, which grow more abundant as the size of the MLP increases. In fact, for the diagonal Gaussian source, it can be shown analytically (Cover & Thomas, 2006) that the optional $Q^*_{Y|X=x}$ is a Gaussian distribution whose mean depends linearly on x, and an identity decoder is optimal.

In another experiment, we investigate the effect of the dimensionality of the latent space on the result. We took the $n = 100$ model from the previous experiment with linear decoder, and modified the dimension of the encoder output, in effect creating a latent space with a different dimension than n. We trained models with various $dim(\mathcal{Z})$, and plot the resulting R-D curves. Figure 1 (c) shows that the R-D upper bound suffers when $dim(\mathcal{Z}) < dim(\hat{\mathcal{X}})$, and is unaffected otherwise.

A.5.2 Banana-shaped Source

For the upper bound algorithm, we base our model architecture on the compressive autoencoder in Ballé et al. (2021), using a 2-dimensional latent space and two-layer MLPs for both the encoder and decoder. We parameterize Q_Z by a MAF (Papamakarios et al., 2017), and $Q_{Z|X}$ by a diagonal Gaussian distribution. The resulting upper bound is shown as the **blue curve** in Fig. **??**, and it can be seen to give a tighter upper bound than the compressive autoencoder (Ballé et al., 2021). As an ablation, we also train a variant model with identity decoder (so $\hat{\mathcal{X}} = \mathcal{Z}$); we found the resulting bound is much looser, and we were only able to match the bound of the original model by increasing

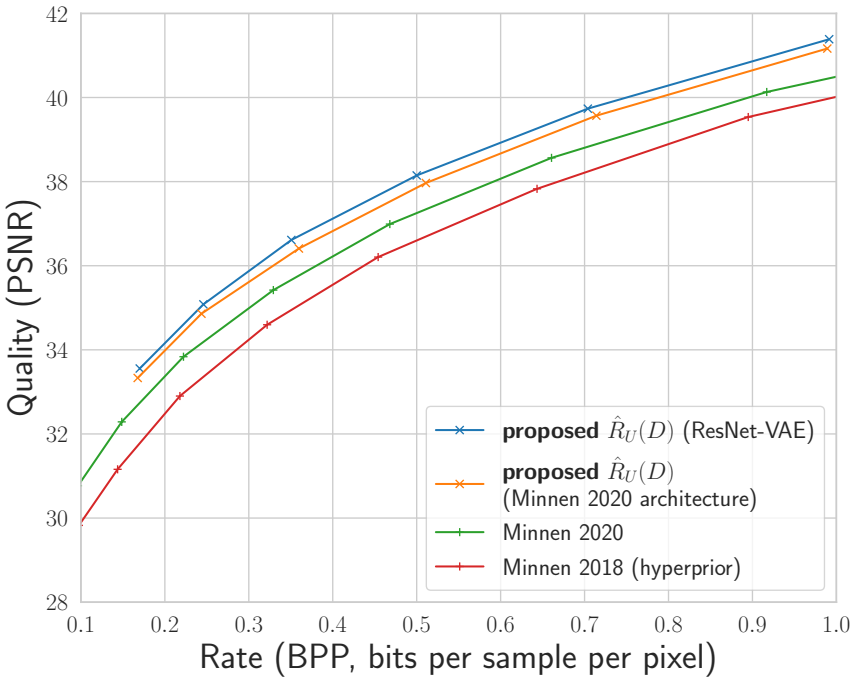

Figure 4: Quality-rate curves on the Tecnick dataset.

the depth of the MAF Q_Z as well as adopting an IAF (Kingma et al., 2016) $Q_{Z|X}$ distribution (plotted in **green curve**).

A.5.3 Natural Images