Evaluating Lossy Compression Rates of Deep Generative Models

Sicong Huang [*1 2 3] **Alireza Makhzani** [*1 2] **Yanshuai Cao** [3] **Roger Grosse** [1 2]

Abstract

The field of deep generative modeling has succeeded in producing astonishingly realistic-seeming images and audio, but quantitative evaluation remains a challenge. Log-likelihood is an appealing metric due to its grounding in statistics and information theory, but it can be challenging to estimate for implicit generative models, and scalar-valued metrics give an incomplete picture of a model's quality. In this work, we propose to use rate distortion (RD) curves to evaluate and compare deep generative models. While estimating RD curves is seemingly even more computationally demanding than log-likelihood estimation, we show that we can approximate the entire RD curve using nearly the same computations as were previously used to achieve a single log-likelihood estimate. We evaluate lossy compression rates of VAEs, GANs, and adversarial autoencoders (AAEs) on the MNIST and CIFAR10 datasets. Measuring the entire RD curve gives a more complete picture than scalar-valued metrics, and we arrive at a number of insights not obtainable from log-likelihoods alone.

1. Introduction

Generative models of images represent one of the most exciting areas of rapid progress of AI (Brock et al., 2019; Karras et al., 2018b;a). However, evaluating the performance of generative models remains a significant challenge. Many of the most successful models, most notably Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), are *implicit generative models* for which computation of log-likelihoods is intractable or even undefined. Evaluation typically focuses on metrics such as the Inception score (Salimans et al., 2016) or the Fréchet Inception Distance (FID) (Heusel et al., 2017), which do not have nearly the same degree of theoretical underpinning as likelihood-based metrics.

Log-likelihoods are one of the most important measures of generative models. Their utility is evidenced by the fact that likelihoods (or equivalent metrics such as perplexity or bits-per-dimension) are reported in nearly all cases where it's convenient to compute them. Unfortunately, computation of log-likelihoods for implicit generative models remains a difficult problem. Furthermore, log-likelihoods have important conceptual limitations. For continuous inputs in the image domain, the metric is often dominated by the fine-grained distribution over pixels rather than the high-level structure. For models with low-dimensional support, one needs to assign an observation model, such as (rather arbitrary) isotropic Gaussian noise (Wu et al., 2016). Lossless compression metrics for GANs often give absurdly large bits-per-dimension (e.g. 10^{14}) which fails to reflect the true performance of the model (Grover et al., 2018; Danihelka et al., 2017). See Theis et al. (2015) for more discussion of limitations of likelihood-based evaluation.

Typically, one is not interested in describing the pixels of an image directly, and it would be sufficient to generate images close to the true data distribution in some metric such as Euclidean distance. For this reason, there has been much interest in Wasserstein distance as a criterion for generative models, since the measure exploits precisely this metric structure (Arjovsky et al., 2017; Gulrajani et al., 2017; Salimans et al., 2018). However, Wasserstein distance remains difficult to approximate, and hence it is not routinely used to evaluate generative models.

We aim to achieve the best of both worlds by measuring *lossy compression* rates of deep generative models. In particular, we aim to estimate the rate distortion function, which measures the number of bits required to match a distribution to within a given distortion. Like Wasserstein distance, it can exploit the metric structure of the observation space, but like log-likelihoods, it connects to the rich literature of probabilistic and information theoretic analysis of generative models. By focusing on different parts of the rate distortion curve, one can achieve different tradeoffs between the description length and the fidelity of recon-

[*]Equal contribution. [1]University of Toronto [2]Vector Institute for Artificial Intelligence [3]Borealis AI. Correspondence to: Alireza Makhzani, Roger Grosse <makhzani, rgrosse@cs.toronto.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

struction — thereby fixing the problem whereby lossless compression focuses on the details at the expense of high-level structure. The lossy compression perspective has the further advantage that the distortion metric need not have a probabilistic interpretation; hence, one is free to use more perceptually valid distortion metrics such as structural similarity (SSIM) (Wang et al., 2004) or distances in a learned feature space (Huang et al., 2018).

Algorithmically, computing rate distortion functions raises similar challenges to estimating log-likelihoods. We show that the rate distortion curve can be computed by finding the normalizing constants of a family of unnormalized probability distributions over the noise variables $\mathbf{z}$. Interestingly, when the distortion metric is squared error, these distributions correspond to the posterior distributions of $\mathbf{z}$ for Gaussian observation models with different variances; hence, the rate distortion analysis generalizes the evaluation of log-likelihoods with Gaussian observation models.

Annealed Importance Sampling (AIS) (Neal, 2001) is currently the most effective general-purpose method for estimating normalizing constants in high dimensions, and was used by Wu et al. (2016) to compare log-likelihoods of a variety of implicit generative models. The algorithm is based on gradually interpolating between a tractable initial distribution and an intractable target distribution. We show that when AIS is used to estimate log-likelihoods under a Gaussian observation model, the sequence of intermediate distributions corresponds to precisely the distributions needed to compute the rate distortion curve. We prove that the AIS estimate of the rate distortion curve is an upper bound on the *entire* rate distortion curve. Furthermore, the tightness of the bound can be validated on simulated data using bidirectional Monte Carlo (BDMC) (Grosse et al., 2015; Wu et al., 2016). Hence, we can approximate the entire rate distortion curve for roughly the same computational cost as a *single log-likelihood estimate*.

We use our rate distortion approximations to study a variety of variational autoencoders (VAEs) (Kingma & Welling, 2013), GANs and adversarial autoencoders (AAE) (Makhzani et al., 2015), and arrive at a number of insights not obtainable from log-likelihoods alone. For instance, we observe that VAEs and GANs have different rate distortion tradeoffs: While VAEs with larger code size can generally achieve better lossless compression rates, their performances drop at lossy compression in the low-rate regime. Conversely, expanding the capacity of GANs appears to bring substantial reductions in distortion at the high-rate regime without any corresponding deterioration in quality in the low-rate regime. We find that increasing the capacity of GANs by increasing the code size (width) has a qualitatively different effect on the rate distortion tradeoffs than increasing the depth. We also find that different GAN variants with the same code size achieve nearly identical RD curves, and that the code size dominates the performance differences between GANs.

2. Background

2.1. Annealed Importance Sampling

Annealed importance sampling (AIS) (Neal, 2001) is a Monte Carlo algorithm based on constructing a sequence of $n+1$ distributions $p_k(\mathbf{z}) = \frac{\tilde{p}_k(\mathbf{z})}{Z_k}$, where $k \in \{0, \ldots, n\}$, between a tractable initial distribution $p_0(\mathbf{z})$ and the intractable target distribution $p_n(\mathbf{z})$. At the k-th state ($0 \leq k \leq n$), the forward distribution q_f and the un-normalized backward distribution $\tilde{q}_b$ are

$$q_f(\mathbf{z}_0, \ldots, \mathbf{z}_k) = p_0(\mathbf{z}_0)\mathcal{T}_0(\mathbf{z}_1|\mathbf{z}_0) \ldots \mathcal{T}_{k-1}(\mathbf{z}_k|\mathbf{z}_{k-1}), \quad (1)$$

$$\tilde{q}_b(\mathbf{z}_0, \ldots, \mathbf{z}_k) = \tilde{p}_k(\mathbf{z}_k)\tilde{\mathcal{T}}_{k-1}(\mathbf{z}_{k-1}|\mathbf{z}_k) \ldots \tilde{\mathcal{T}}_0(\mathbf{z}_0|\mathbf{z}_1), \quad (2)$$

where $\mathcal{T}_k$ is an MCMC kernel that leaves $p_k(\mathbf{z})$ invariant; and $\tilde{\mathcal{T}}_k$ is its reverse kernel. We run M independent AIS chains, numbered $i = 1, \ldots, M$. Let $\mathbf{z}_k^i$ be the k-th state of the i-th chain. The importance weights w_k^i and normalized importance weights $\tilde{w}_k^i$ are

$$w_k^i = \frac{\tilde{q}_b(\mathbf{z}_1^i, \ldots, \mathbf{z}_k^i)}{q_f(\mathbf{z}_1^i, \ldots, \mathbf{z}_k^i)} = \frac{\tilde{p}_1(\mathbf{z}_1^i)}{p_0(\mathbf{z}_1^i)} \frac{\tilde{p}_2(\mathbf{z}_2^i)}{\tilde{p}_1(\mathbf{z}_2^i)} \cdots \frac{\tilde{p}_k(\mathbf{z}_k^i)}{\tilde{p}_{k-1}(\mathbf{z}_k^i)} \quad (3)$$

$$\tilde{w}_k^i = \frac{w_k^i}{\sum_{i=1}^M w_k^i}. \quad (4)$$

At the k-th step, an unbiased estimate of the partition function of $p_k(\mathbf{z})$ can be found using $\hat{Z}_k = \frac{1}{M}\sum_{i=1}^M w_k^i$.

At the k-th step, we define the *AIS distribution* $q_k^{\text{AIS}}(\mathbf{z})$ as the distribution obtained by first sampling $\mathbf{z}_k^1, \ldots, \mathbf{z}_k^M$ from the M parallel chains using the forward distribution $q_f(\mathbf{z}_1^i, \ldots, \mathbf{z}_M^i)$, and then re-sampling these samples based on $\tilde{w}_k^i$. More formally, the AIS distribution is defined as follows:

$$q_k^{\text{AIS}}(\mathbf{z}) = \mathbb{E}_{\prod_{i=1}^M q_f(\mathbf{z}_1^i, \ldots, \mathbf{z}_k^i)}[\sum_{i=1}^M \tilde{w}_k^i \delta(\mathbf{z} - \mathbf{z}_k^i)]. \quad (5)$$

2.2. Bidirectional Monte Carlo.

We know that the AIS log partition function estimate $\log \hat{Z}$ is a *stochastic lower bound* on $\log Z$ (Jensen's inequality). As the result, using the forward AIS distribution as the proposal distribution results in a lower bound on the data log-likelihood. By running AIS in reverse, however, we obtain an upper bound on $\log Z$. However, in order to run the AIS in reverse, we need exact samples from the true posterior, which is only possible on the simulated data. The combination of the AIS lower bound and upper bound on the log partition function is called *bidirectional Monte Carlo* (BDMC), and the gap between these bounds is called the *BDMC gap* (Grosse et al., 2015). We note that

AIS combined with BDMC has been used to estimate log-likelihoods for deep generative models (Wu et al., 2016). In this work, we validate our AIS experiments by using the BDMC gap to measure the accuracy of our partition function estimators.

2.3. Implicit Generative Models

The goal of generative modeling is to learn a model distribution $p(\mathbf{x})$ to approximate the data distribution $p_d(\mathbf{x})$. Implicit generative models define the model distribution $p(\mathbf{x})$ using a latent variable $\mathbf{z}$ with a fixed prior distribution $p(\mathbf{z})$ such as a Gaussian distribution, and a decoder or generator network which computes $\hat{\mathbf{x}} = f(\mathbf{z})$. In some cases (e.g. VAEs, AAEs), the generator explicitly parameterizes a conditional distribution $p(\mathbf{x}|\mathbf{z})$, such as a Gaussian observation model $\mathcal{N}(\mathbf{x}; f(\mathbf{z}), \sigma^2\mathbf{I})$. But in implicit models such as GANs, the generator directly outputs $\hat{\mathbf{x}} = f(\mathbf{z})$. In order to treat VAEs and GANs under a consistent framework, we ignore the Gaussian observation model of VAEs (thereby treating the VAE decoder as an implicit model), and use the squared error distortion of $d(\mathbf{x}, f(\mathbf{z})) = \|\mathbf{x} - f(\mathbf{z})\|_2^2$. However, we note that it is also possible to assume a Gaussian observation model with a fixed σ^2 for GANs, and use the Gaussian negative log-likelihood (NLL) as the distortion measure for both VAEs and GANs: $d(\mathbf{x}, f(\mathbf{z})) = -\log \mathcal{N}(\mathbf{x}; f(\mathbf{z}), \sigma^2\mathbf{I})$. This is equivalent to squared error distortion up to a linear transformation.

2.4. Rate Distortion Theory

Let $\mathbf{x}$ be a random variable that comes from the data distribution $p_d(\mathbf{x})$. Shannon's fundamental compression theorem states that we can compress this random variable losslessly at the rate of $\mathcal{H}(\mathbf{x})$. But if we allow lossy compression, we can compress $\mathbf{x}$ at the rate of R, where $R \leq \mathcal{H}(\mathbf{x})$, using the code $\mathbf{z}$, and have a lossy reconstruction $\hat{\mathbf{x}} = f(\mathbf{z})$ with the distortion of D, given a distortion measure $d(\mathbf{x}, \hat{\mathbf{x}}) = d(\mathbf{x}, f(\mathbf{z}))$. The rate distortion theory (Cover & Thomas, 2012) quantifies the trade-off between the lossy compression rate R and the distortion D. The rate distortion function $\mathcal{R}(D)$ is defined as the minimum number of bits per sample required to achieve lossy compression of the data such that the average distortion measured by the distortion function is less than D. Shannon's rate distortion theorem states that $\mathcal{R}(D)$ equals the minimum of the following optimization problem:

$$\min_{q(\mathbf{z}|\mathbf{x})} \mathcal{I}(\mathbf{z}; \mathbf{x}) \qquad s.t.\ \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[d(\mathbf{x}, f(\mathbf{z}))] \leq D. \tag{6}$$

where the optimization is over the *channel conditional* distribution $q(\mathbf{z}|\mathbf{x})$. Suppose the data-distribution is $p_d(\mathbf{x})$. The channel conditional $q(\mathbf{z}|\mathbf{x})$ induces the joint distribution $q(\mathbf{z}, \mathbf{x}) = p_d(\mathbf{x})q(\mathbf{z}|\mathbf{x})$, which defines the mutual information $\mathcal{I}(\mathbf{z}; \mathbf{x})$. $q(\mathbf{z})$ is the marginal distribution over $\mathbf{z}$ of the joint distribution $q(\mathbf{z}, \mathbf{x})$, and is called the *output marginal* distribution. We can rewrite the optimization of Eq. 6 using the method of Lagrange multipliers as follows:

$$\min_{q(\mathbf{z}|\mathbf{x})} \mathcal{I}(\mathbf{z}; \mathbf{x}) + \beta \mathbb{E}_{q(\mathbf{z},\mathbf{x})}[d(\mathbf{x}, f(\mathbf{z}))]. \tag{7}$$

2.5. Variational Bounds on Mutual Information

We must modify the standard rate distortion formalism slightly in order to match the goals of generative model evaluation. Specifically, we are interested in evaluating lossy compression with coding schemes corresponding to *particular* trained generative models, including the fixed prior $p(\mathbf{z})$. For models such as VAEs, $\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ is standardly interpreted as the description length of $\mathbf{z}$. Hence, we adjust the rate distortion formalism to use $\mathbb{E}_{p_d(\mathbf{x})}\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ in place of $\mathcal{I}(\mathbf{x}, \mathbf{z})$. It can be shown that $\mathbb{E}_{p_d(\mathbf{x})}\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ is a variational upper bound on the standard rate $\mathcal{I}(\mathbf{x}, \mathbf{z})$:

$$\begin{aligned} \mathcal{I}(\mathbf{x}; \mathbf{z}) &\leq \mathcal{I}(\mathbf{x}; \mathbf{z}) + \mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z})) \\ &= \mathbb{E}_{p_d(\mathbf{x})}\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})). \end{aligned} \tag{8}$$

In the context of variational inference, $q(\mathbf{z}|\mathbf{x})$ is the posterior, $q(\mathbf{z}) = \int p_d(\mathbf{x})q(\mathbf{z}|\mathbf{x})d\mathbf{x}$ is the aggregated posterior (Makhzani et al., 2015), and $p(\mathbf{z})$ is the prior. In the context of rate distortion theory, $q(\mathbf{z}|\mathbf{x})$ is the channel conditional, $q(\mathbf{z})$ is the output marginal, and $p(\mathbf{z})$ is the *variational output marginal* distribution. The inequality is tight when $p(\mathbf{z}) = q(\mathbf{z})$, i.e., the variational output marginal (prior) is equal to the output marginal (aggregated posterior). We note that the upper bound of Eq. 8 has been used in other algorithms such as the Blahut-Arimoto algorithm (Arimoto, 1972) or the variational information bottleneck algorithm (Alemi et al., 2016).

3. Variational Rate Distortion Functions

Analogously to the rate distortion function, we define the *variational rate distortion* function $\mathcal{R}_p(D)$ as the minimum value of the variational upper bound $\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ for a given distortion D. More precisely, $\mathcal{R}_p(D)$ is the solution of

$$\min_{q(\mathbf{z}|\mathbf{x})} \mathbb{E}_{p_d(\mathbf{x})}\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})) \quad s.t.\ \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[d(\mathbf{x}, f(\mathbf{z}))] \leq D. \tag{9}$$

We can rewrite the optimization of Eq. 9 using the method of Lagrange multipliers as follows:

$$\min_{q(\mathbf{z}|\mathbf{x})} \mathbb{E}_{p_d(\mathbf{x})}\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})) + \beta \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[d(\mathbf{x}, f(\mathbf{z}))]. \tag{10}$$

Conveniently, the Lagrangian decomposes into independent optimization problems for each $\mathbf{x}$, allowing us to treat this as an optimization problem over $q(\mathbf{z}|\mathbf{x})$ for fixed $\mathbf{x}$. We

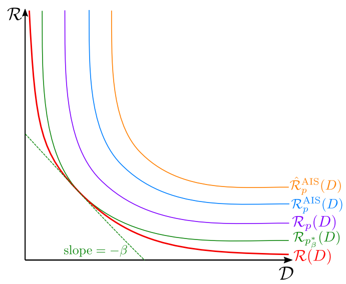


Figure 1. Geometric illustration of the true RD curve and its upper bounds.

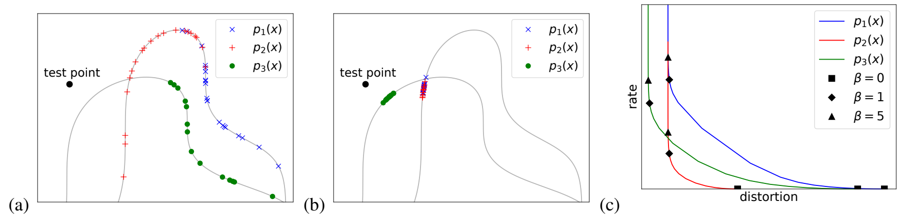


Figure 2. **(a)** Prior samples, or equivalently, reconstructions with $\beta = 0$. **(b)** High-rate reconstructions ($\beta = 5$). **(c)** Variational rate distortion curves for each of these three models.

can compute the rate distortion curve by sweeping over β rather than by sweeping over D.

Now we describe some of the properties of the variational rate distortion function $\mathcal{R}_p(D)$, which are straightforward analogues of well-known properties of the rate distortion function.

Proposition 1. $\mathcal{R}_p(D)$ has the following properties:

(a) $\mathcal{R}_p(D)$ is non-increasing and convex function of D.

(b) We have $\mathcal{R}(D) = \min_{p(\mathbf{z})} \mathcal{R}_p(D)$. As a corollary, for any $p(\mathbf{z})$, we have $\mathcal{R}(D) \leq \mathcal{R}_p(D)$.

(c) The variational rate distortion optimization of Eq. 10 has a unique global optimum which can be expressed as $q_\beta^*(\mathbf{z}|\mathbf{x}) = \frac{1}{Z_\beta(\mathbf{x})} p(\mathbf{z}) \exp(-\beta d(\mathbf{x}, f(\mathbf{z})))$.

Proof. See Appendix A.1.

Fig. 1 shows a geometrical illustration of Prop. 1. As stated by Prop. 1b, for any prior $p(\mathbf{z})$, $\mathcal{R}_p(D)$ is a variational upper bound on $\mathcal{R}(D)$. More specifically, we have $\mathcal{R}(D) = \min_{p(\mathbf{z})} \mathcal{R}(D)$, which implies that for any given β, there exists a prior $p_\beta^*(\mathbf{z}) = \int q_\beta^*(\mathbf{z}|\mathbf{x}) p_d(\mathbf{x}) d\mathbf{x}$, for which the gap between the rate distortion and variational rate distortion functions at β is zero. Furthermore, $\mathcal{R}_{p_\beta^*}(D)$ and $\mathcal{R}_p(D)$ are tangent to each other at the point corresponding to β, and both are tangent to the line with the slope of $-\beta$ passing through this point. In the next section, we will describe how we can use AIS to estimate $\mathcal{R}_p(D)$, and will derive the upper bounds of $\mathcal{R}_p^{\mathrm{AIS}}(D)$ and $\hat{\mathcal{R}}_p^{\mathrm{AIS}}(D)$.

Variational Rate Distortion Functions with NLL Distortion. If the decoder outputs a probability distribution (as in a VAE), we can define the distortion metric to coincide with the negative log-likelihood (NLL): $d(\mathbf{x}, f(\mathbf{z})) = -\log p(\mathbf{x}|\mathbf{z})$. We now describe some of the properties of variational rate distortion functions with NLL distortions.

Proposition 2. The variational rate distortion function $\mathcal{R}_p(D)$ with NLL distortion of $-\log p(\mathbf{x}|\mathbf{z})$ has the following properties:

(a) $\mathcal{R}(D)$ is lower bounded by the linear function of $\mathcal{H}_d - D$, and upper bounded by the variational rate distortion function: $\mathcal{H}_d - D \leq \mathcal{R}(D) \leq \mathcal{R}_p(D)$.

(b) The global optimum of variational rate distortion optimization (Eq. 10) can be expressed as

$q_\beta^*(\mathbf{z}|\mathbf{x}) = \frac{1}{Z_\beta(\mathbf{x})} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z})^\beta$

where $Z_\beta^*(\mathbf{x}) = \int p(\mathbf{z}) p(\mathbf{x}|\mathbf{z})^\beta d\mathbf{z}$.

(c) At $\beta = 1$, the negative summation of rate and distortion is the true log-likelihood:

$\mathcal{L}_p = \mathbb{E}_{p_d(\mathbf{x})}[\log p(\mathbf{x})] = -R_\beta\big|_{\beta=1} - D_\beta\big|_{\beta=1}$.

Proof. See Appendix A.2.

Illustrative Example. We now motivate the rate-distortion tradeoff in generative models using an illustrative example (Fig. 2). Suppose we have three generative models $p_1(\mathbf{x})$, $p_2(\mathbf{x})$ and $p_3(\mathbf{x})$, on a 2-D space with a 1-D Gaussian latent code. The conditional likelihood of these generative models define a 1-D manifold (grey curves in Fig. 2a,b) on the 2-D data space. In Fig. 2a, the colored points represent prior samples from the three models. The $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ models have the same decoder but have different priors. Thus they define different distributions on the same manifold. The $p_3(\mathbf{x})$ model has a different prior and conditional likelihood. Fig. 2c shows the RD curves of each model for the test point shown in Fig. 2a,b. In the low-rate regime, the reconstructions of the models are close to the prior samples (Fig. 2a). Thus, the $p_2(\mathbf{x})$ model achieves a better average reconstruction than the $p_3(\mathbf{x})$ model, and the $p_3(\mathbf{x})$ model is better than the $p_1(\mathbf{x})$. However, in the high-rate regime ($\beta = 5$, Fig. 2b), the $p_3(\mathbf{x})$ model can use a large number of bits to specify reconstruction points on its manifold that are very close to the test-point, and thus outperforms the high-rate reconstructions of both the $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ model. We can also see that at high-rates, the prior is ignored and the compression rates of the $p_1(\mathbf{x})$ and the $p_2(\mathbf{x})$ model match each other.

4. Bounding Rate Distortion Functions with AIS

We've shown that evaluating the variational rate-distortion function $\mathcal{R}_p(D)$ amounts to sampling from and estimating the partition functions of a particular sequence of distributions. Unfortunately, both computations are generally intractable for deep generative models. Furthermore, computing the entire RD curve would seem to require separately performing inference for many values of β.

In this section, we show how to obtain (in practice highly accurate) upper bounds on the RD curve using AIS (see Section 2.1 for background). AIS has several properties that make it remarkably useful for RD curve estimation: (1) the RD target distributions correspond exactly to the set of intermediate distributions commonly used for AIS, (2) it non-asymptotically lower bounds the log partition function in expectation, and (3) the work already done for one distribution helps in doing inference for the next distribution. Due to these properties, we can obtain the entire RD curve in roughly the same time required to estimate a single log-likelihood value.

AIS Chain. We fix a temperature schedule $0 = \beta_0 < \beta_1 < \ldots < \beta_n = \infty$. For the kth intermediate distribution, we use the optimal channel conditional $q_k(\mathbf{z}|\mathbf{x})$ and partition function $Z_k(\mathbf{x})$, corresponding to points along $\mathcal{R}_p(D)$ and derived in Prop. 1c: $q_k(\mathbf{z}|\mathbf{x}) = \frac{1}{Z_k}\tilde{q}_k(\mathbf{z}|\mathbf{x})$, where

$$\tilde{q}_k(\mathbf{z}|\mathbf{x}) = p(\mathbf{z})\exp(-\beta_k d(\mathbf{x}, f(\mathbf{z}))), \tag{11}$$

$$Z_k(\mathbf{x}) = \int \tilde{q}_k(\mathbf{z}|\mathbf{x})d\mathbf{z}. \tag{12}$$

Conveniently, this choice coincides with geometric averages, the typical choice of intermediate distributions for AIS. For the transition operator, we use Hamiltonian Monte Carlo (Neal et al., 2011). At the k-th step, the rate is denoted by $R_k(\mathbf{x}) = \mathrm{KL}(q_k(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ and the distortion is denoted by $D_k(\mathbf{x}) = \mathbb{E}_{q_k(\mathbf{z}|\mathbf{x})}[d(\mathbf{x}, f(\mathbf{z}))]$.

AIS Rate Distortion Curves. For each data point $\mathbf{x}$, we run M independent AIS chains, numbered $i = 1, \ldots, M$, in the forward direction. At the k-th state of the i-th chain, let $\mathbf{z}_k^i$ be the state, w_k^i be the AIS importance weights, and $\tilde{w}_k^i$ be the normalized AIS importance weights. We denote the AIS distribution at the k-th step as the distribution obtained by first sampling from all the M forward distributions $q_f(\mathbf{z}_1^i, \ldots, \mathbf{z}_k^i|\mathbf{x})\Big|_{i=1:M}$, and then re-sampling the samples based on their normalized importance weights $\tilde{w}_k^i$ (see Section 2.1 and Appendix A.4 for more details). More formally $q_k^{\mathrm{AIS}}(\mathbf{z}|\mathbf{x})$ is

$$q_k^{\mathrm{AIS}}(\mathbf{z}|\mathbf{x}) = \mathbb{E}_{\prod_{i=1}^M q_f(\mathbf{z}_1^i,\ldots,\mathbf{z}_k^i|\mathbf{x})}[\sum_{i=1}^M \tilde{w}_k^i \delta(\mathbf{z} - \mathbf{z}_k^i)]. \tag{13}$$

Using the AIS distribution $q_k^{\mathrm{AIS}}(\mathbf{z}|\mathbf{x})$ defined in Eq. 13, we now define the AIS distortion $D_k^{\mathrm{AIS}}(\mathbf{x})$ and the AIS rate $R_k^{\mathrm{AIS}}(\mathbf{x})$ as follows: $D_k^{\mathrm{AIS}}(\mathbf{x}) = \mathbb{E}_{q_k^{\mathrm{AIS}}(\mathbf{z}|\mathbf{x})}[d(\mathbf{x}, f(\mathbf{z}))]$ and $R_k^{\mathrm{AIS}}(\mathbf{x}) = \mathrm{KL}(q_k^{\mathrm{AIS}}(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$. We now define the *AIS rate distortion curve* $\mathcal{R}_p^{\mathrm{AIS}}(D)$ (shown in Fig. 1) as the RD curve obtained by tracing pairs of $\left(R_k^{\mathrm{AIS}}(\mathbf{x}), D_k^{\mathrm{AIS}}(\mathbf{x})\right)$.

Proposition 3. The AIS rate distortion curve upper bounds the variational rate distortion function: $\mathcal{R}_p^{\mathrm{AIS}}(D) \geq \mathcal{R}_p(D)$. *Proof:* See Appendix A.3.

Estimated AIS Rate Distortion Curves. Although the AIS distribution can be easily sampled from, its density is intractable to evaluate. As the result, evaluating $\mathcal{R}_p^{\mathrm{AIS}}(D)$ is also intractable. We now propose to evaluate an upper-bound on $\mathcal{R}_p^{\mathrm{AIS}}(D)$ by finding an upper bound for $R_k^{\mathrm{AIS}}(\mathbf{x})$, and an unbiased estimate for $D_k^{\mathrm{AIS}}(\mathbf{x})$. We use the AIS distribution samples $\mathbf{z}_k^i$ and their corresponding weights $\tilde{w}_k^i$ to obtain the following distortion and partition function estimates:

$$\hat{D}_k^{\mathrm{AIS}}(\mathbf{x}) = \sum_i \tilde{w}_k^i d(\mathbf{x}, f(\mathbf{z}_k^i))), \quad \hat{Z}_k^{\mathrm{AIS}}(\mathbf{x}) = \frac{1}{M}\sum_i w_k^i.$$

Having found the estimates $\hat{D}_k^{\mathrm{AIS}}(\mathbf{x})$ and $\hat{Z}_k^{\mathrm{AIS}}(\mathbf{x})$, we propose to estimate the rate as follows:

$$\hat{R}_k^{\mathrm{AIS}}(\mathbf{x}) = -\log \hat{Z}_k^{\mathrm{AIS}}(\mathbf{x}) - \beta_k \hat{D}_k^{\mathrm{AIS}}(\mathbf{x}). \tag{14}$$

We define the *estimated AIS rate distortion curve* $\hat{\mathcal{R}}_p^{\mathrm{AIS}}(D)$ (shown in Fig. 1) as an RD curve obtained by tracing pairs of rate distortion estimates $\left(\hat{R}_k^{\mathrm{AIS}}(\mathbf{x}), \hat{D}_k^{\mathrm{AIS}}(\mathbf{x})\right)$.

Proposition 4. The estimated AIS rate distortion curve upper bounds the AIS rate distortion curve in expectation: $\mathbb{E}[\hat{\mathcal{R}}_p^{\mathrm{AIS}}(D)] \geq \mathcal{R}_p^{\mathrm{AIS}}(D)$. More specifically, we have

$$\mathbb{E}[\hat{R}_k^{\mathrm{AIS}}(\mathbf{x})] \geq R_k^{\mathrm{AIS}}(\mathbf{x}), \qquad \mathbb{E}[\hat{D}_k^{\mathrm{AIS}}(\mathbf{x})] = D_k^{\mathrm{AIS}}(\mathbf{x}). \tag{15}$$

Proof Sketch. For the complete proof see Appendix A.4. It is straightforward to show that $\mathbb{E}[\hat{D}_k^{\mathrm{AIS}}(\mathbf{x})] = D_k^{\mathrm{AIS}}(\mathbf{x})$, however, the nontrivial part is to show $\mathbb{E}[\hat{R}_k^{\mathrm{AIS}}(\mathbf{x})] \geq R_k^{\mathrm{AIS}}(\mathbf{x})$. Suppose $\mathbf{V}$ is all the AIS states across M parallel chains except the final selected state $\mathbf{z}$. The monotonicity of KL divergence enables us to upper bound the rate $R_k^{\mathrm{AIS}}(\mathbf{x}) = \mathrm{KL}(q_k^{\mathrm{AIS}}(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ in terms of the full KL divergence in the AIS extended state space: $\mathrm{KL}(q_k^{\mathrm{AIS}}(\mathbf{z}, \mathbf{V}|\mathbf{x})\|p(\mathbf{z})p(\mathbf{V}|\mathbf{z}, \mathbf{x}))$, where $q_k^{\mathrm{AIS}}(\mathbf{z}, \mathbf{V}|\mathbf{x})$ is constructed to have the marginal of $q_k^{\mathrm{AIS}}(\mathbf{z}|\mathbf{x})$ (similar to the auxiliary variable construction of Domke & Sheldon (2018)), and $p(\mathbf{z})p(\mathbf{V}|\mathbf{z}, \mathbf{x})$ has the marginal of $p(\mathbf{z})$. We then prove that the AIS estimate of the rate $\hat{R}_k^{\mathrm{AIS}}(\mathbf{x})$ equals the full KL divergence in expectation (see Appendix A.4).

In summary, from Prop. 1, Prop. 3 and Prop. 4, we can conclude that the estimated AIS rate distortion curve upper bounds the true rate distortion curve in expectation (shown in Fig. 1):

$$\mathbb{E}[\hat{\mathcal{R}}_p^{\mathrm{AIS}}(D)] \geq \mathcal{R}_p^{\mathrm{AIS}}(D) \geq \mathcal{R}_p(D) \geq \mathcal{R}(D). \tag{16}$$

In all our experiments, we plot the estimated AIS rate distortion function $\hat{\mathcal{R}}_p^{\text{AIS}}(D)$.

Accuracy of AIS Estimates. While the above discussion focuses on obtaining upper bounds, we note that AIS is one of the most accurate general-purpose methods for estimating partition functions, and therefore we believe our AIS upper bounds to be fairly tight in practice. In theory, for a large number of intermediate distributions, the AIS variance is proportional to $1/MK$ (Neal, 2001; 2005), where M is the number of AIS chains and K is the number of intermediate distributions. For the main experiments of our paper, we evaluate the tightness of the AIS estimate by computing the BDMC gap, and show that in practice our upper bounds are tight (Section 6.1 and Appendix B).

5. Related Work

Evaluation of Implicit Generative Models. Many heuristic measures have been proposed for evaluation of implicit models, such as the Inception score (Salimans et al., 2016) and the Fréchet Inception Distance (FID) (Heusel et al., 2017). One of the main drawbacks of the IS or FID is that they can only provide a single scalar value that cannot distinguish the mode dropping behavior from the mode inventing behavior in generative models. In order to address this, Sajjadi et al. (2018) proposed to study the precision-recall tradeoff for evaluating generative models. The precision-recall tradeoff is analogous to our rate-distortion tradeoff, but has a very different mathematical motivation.

Rate Distortion Theory and Generative Models. Perhaps the closest work to ours is "Fixing a Broken ELBO" (Alemi et al., 2018), which plots variational rate distortion curves for VAEs. Our work is different than Alemi et al. (2018) in two key aspects. First, in Alemi et al. (2018) the variational rate distortion function is evaluated by fixing the architecture of the neural network, and learning the distortion measure $d(\mathbf{x}, f(\mathbf{z}))$ in addition to learning $q(\mathbf{z}|\mathbf{x})$. Whereas, our work follows a conceptually different goal which is to evaluate a particular generative model with a fixed prior and decoder, independent of how the model was trained. The second key difference is that we find the optimal channel conditional $q^*(\mathbf{z}|\mathbf{x})$ by using AIS; while in Alemi et al. (2018), $q(\mathbf{z}|\mathbf{x})$ is a variational distribution restricted to a variational family. In Section 6.5, we will empirically show that AIS obtains significantly tighter RD bounds than variational methods.

Practical Compression Schemes. We have justified our use of compression terminology in terms of Shannon's fundamental result implying that there exists a rate distortion code for any rate distortion pair that is achievable according to the rate distortion function. For *lossless* compression with generative models, there is a practical compression scheme which nearly achieves the theoretical rate (i.e. the negative ELBO): bits-back encoding. The basic scheme was proposed by Wallace (1990); Hinton & Van Camp (1993), and later implemented by Frey & Hinton (1996). Practical versions for modern deep generative models were developed by Townsend et al. (2019); Kingma et al. (2019). Other researchers have developed practical *lossy* coding schemes achieving variational rate distortion bounds for particular latent variable models which exploited the factorial structure of the variational posterior (Ballé et al., 2018; Theis et al., 2017; Yang et al., 2020). These methods are not directly applicable in our setting, since we don't assume an explicit encoder network, and our variational posteriors lack a convenient factorized form. We don't know whether our variational approximation will lead to a practical lossy compression scheme, but the successes for other variational methods give us hope.

6. Experiments

In this section, we use our rate distortion approximations to answer the following questions: How do different generative models such as VAEs, GANs and AAEs perform at different lossy compression rates? What insights can we obtain from the rate distortion curves about different characteristics of generative models? What is the effect of the code size (width), depth of the network, or the learning algorithm on the rate distortion tradeoffs?

The code for reproducing the experiments can be found at https://github.com/BorealisAI/rate_distortion and https://github.com/huangsicong/rate_distortion.

6.1. Validating AIS

Linear VAE and BDMC. We conducted several experiments to validate the correctness of our implementation and the accuracy of the AIS estimates. Firstly, we compared our AIS results with the analytical solution of the variational rate distortion curve on a linear VAE (derived in Appendix B.1) trained on MNIST. As shown in Fig. 3, the RD curve estimated by AIS agrees closely with the analytical solution. Secondly, for the main experiments of the paper, we evaluated the tightness of the AIS estimate by computing the BDMC gap. The largest BDMC gap for VAEs and AAEs was 0.537 nats, and the largest BDMC gap for GANs was 3.724 nats, showing that our AIS upper bounds are tight. More details are provided in Appendix B.

6.2. Rate Distortion Curves of Deep Generative Models

Rate Distortion Curves of GANs. Fig. 4 shows rate distortion curves for GANs trained on MNIST and CIFAR-10. We varied the dimension of the noise vector $\mathbf{z}$, as well as the depth of the decoder. For the GAN experiments on MNIST (Fig. 4a), the label "deep" corresponds to three hidden layers of size 1024, and the label "shallow" corresponds to one hidden layer of size 1024. We trained shallow and deep

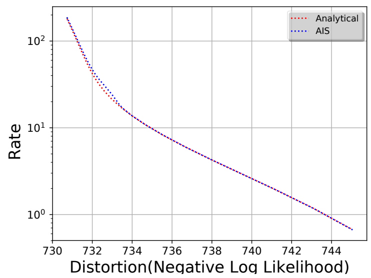
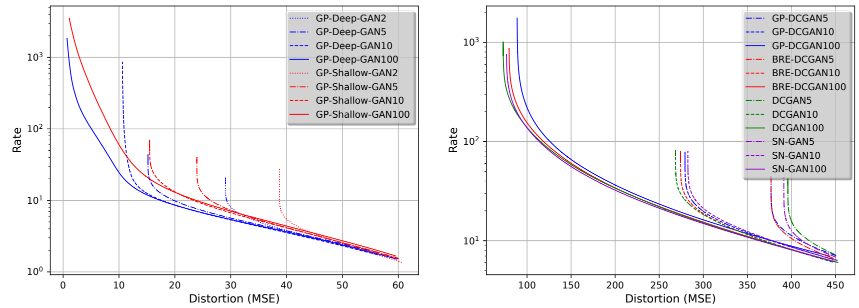


Figure 3. Analytical vs. AIS variational RD curves for a linear VAE.



(a) (b)

Figure 4. Variational rate-distortion curves of GANs. **(a)** MNIST. **(b)** CIFAR-10.

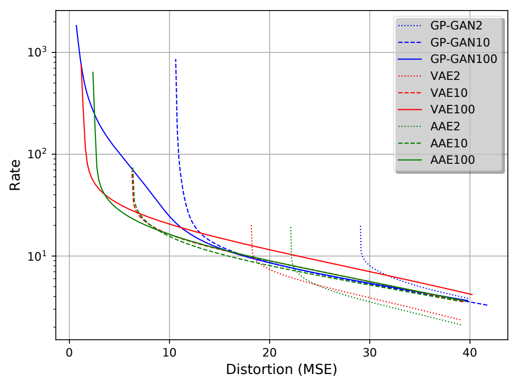


Figure 5. Variational RD curves of VAEs, GANs and AAEs.

GANs with Gradient Penalty (GAN-GP) (Gulrajani et al., 2017) with the code size $d \in \{2, 5, 10, 100\}$ on MNIST. For the GAN experiments on CIFAR-10 (Fig. 4b), we trained the DCGAN (Radford et al., 2015), GAN with Gradient Penalty (GP) (Gulrajani et al., 2017), SN-GAN (Miyato et al., 2018), and BRE-GAN (Cao et al., 2018), with the code size of $d \in \{2, 10, 100\}$. In both the MNIST and CIFAR experiments, we observe that in general increasing the code size has the effect of *extending the curve leftwards*. This is expected, since the high-rate regime is effectively measuring reconstruction ability, and additional dimensions in $\mathbf{z}$ improves the reconstruction.

We also observe from Fig. 4b that different GAN variants with the same code size have nearly identical RD curves, and that the code size dominates the algorithmic differences of GANs.

We can also observe from Fig. 4a that increasing the depth pushes the curves *down and to the left*. In other words, the distortion in both high-rate and mid-rate regimes improves. In these regimes, increasing the depth increases the capacity of the network, which enables the network to make a better use of the information in the code space. In the low-rate regime, however, increasing the depth, similar to increasing the latent size, does not improve the distortion.

Rate Distortion Curves of VAEs. Fig. 5 compares VAEs, AAEs and GP-GANs (Gulrajani et al., 2017) with the code size of $d \in \{2, 10, 100\}$, and the same decoder architecture on the MNIST dataset. In general, we can see that in the mid-rate to high-rate regimes, VAEs achieve better distortions than GANs with the same architecture. This is expected as the VAE is trained with the ELBO objective, which encourages good reconstructions (in the case of factorized Gaussian decoder). We can see from Fig. 5 that in VAEs, increasing the latent capacity pushes the rate distortion curve *up and to the left*. In other words, in contrast with GANs where increasing the latent capacity always improves the rate distortion curve, in VAEs, there is a tradeoff whereby increasing the capacity reduces the distortion at the high-rate regime, at the expense of increasing the distortion in the low-rate regime (or equivalently, increasing the rate required to adequately approximate the data).

We believe the performance drop of VAEs in the low-rate regime is symptomatic of the "holes problem" (Rezende & Viola, 2018; Makhzani et al., 2015) in the code space of VAEs with large code size: because these VAEs allocate a large fraction of their latent spaces to garbage images, it requires many bits to get close to the image manifold. Interestingly, this trade-off could also help explain the well-known problem of blurry samples from VAEs: in order to avoid garbage samples (corresponding to large distortion in the low-rate regime), one needs to reduce the capacity, thereby increasing the distortion at the high-rate regime. By contrast, GANs do not suffer from this tradeoff, and one can train high-capacity GANs without sacrificing performance in the low-rate regime.

Rate Distortion Curves of AAEs. The AAE was introduced by Makhzani et al. (2015) to address the holes problem of VAEs, by directly matching the aggregated posterior to the prior in addition to optimizing the reconstruction cost. Fig. 5 shows the RD curves of AAEs. In comparison

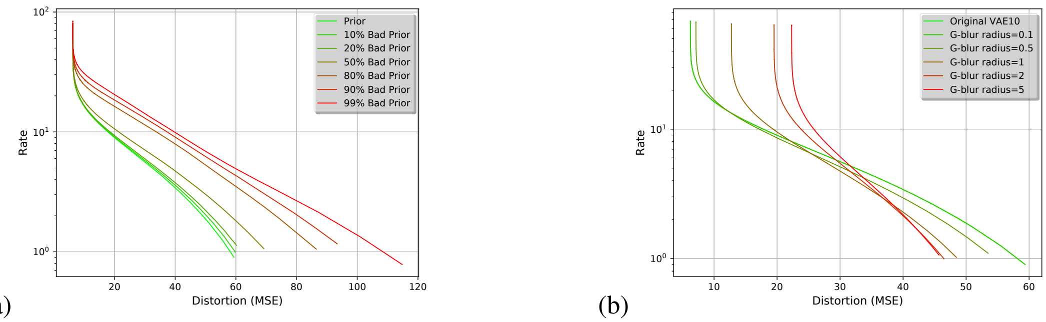


Figure 6. **(a)** Effect of damaging the VAE prior by using a mixture with a bad prior. **(b)** Effect of damaging the conditional likelihood of a VAE by convolving with a Gaussian blur kernel after the last decoder layer.

to GANs, AAEs can match the low-rate performance of GANs, but achieve a better high-rate performance. This is expected as AAEs directly optimize the reconstruction cost as part of their objective. In comparison to VAEs, AAEs perform slightly worse at the high-rate regime, which is expected as the adversarial regularization of AAEs is stronger than the KL regularization of VAEs. But AAEs perform slightly better in the low-rate regime, as they can alleviate the holes problem to some extent.

6.3. Distinguishing Different Failure Modes in Generative Modeling

Since log-likelihoods constitute only a scalar value, they are unable to distinguish different aspects of a generative model which could be good or bad, such as the prior or the observation model. Here, we show that two manipulations which damage a trained VAE in different ways result in very different behavior of the RD curves.

Our first manipulation, originally proposed by Theis et al. (2015), is to use a mixture of the VAE's density and another distribution concentrated away from the data distribution. As pointed out by Theis et al. (2015), this results in a model which achieves high log-likelihood while generating poor samples. Specifically, after training the VAE10 on MNIST, we "damage" its prior $p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$ by altering it to a mixture prior $(1-\alpha)p(\mathbf{z}) + \alpha q(\mathbf{z})$, where $q(\mathbf{z}) = \mathcal{N}(0, 10\mathbf{I})$ is a "poor" prior, which is chosen to be far away from the original prior $p(\mathbf{z})$; and α is close to 1. This process would results in a "poor" generative model that generates garbage samples most of the time (more precisely with the probability of α). Suppose $p(\mathbf{x})$ and $q(\mathbf{x})$ are the likelihood of the good and the poor generative models. It is straightforward to see that $\log q(\mathbf{x})$ is at most 4.6 nats worse that $\log p(\mathbf{x})$, and thus log-likelihood fails to tell these models apart:

$$\log q(\mathbf{x}) = \log \left(0.01 p(\mathbf{x}) + 0.99 \int q(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) d\mathbf{z}\right) \quad (17)$$

$$> \log(0.01 p(\mathbf{x})) \approx \log p(\mathbf{x}) - 4.6 \quad (18)$$

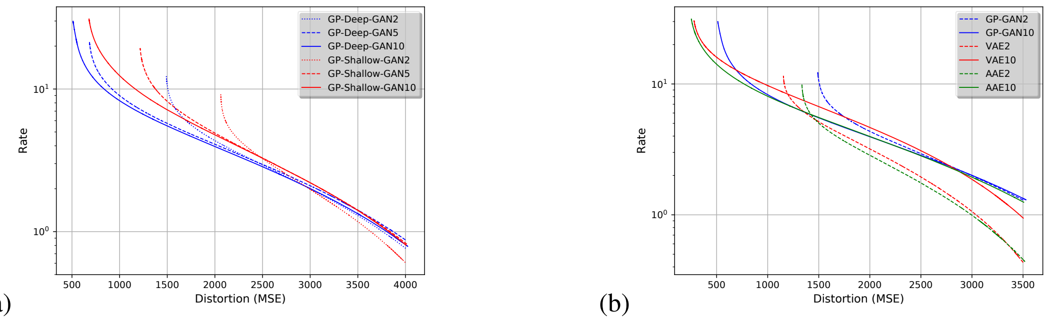
Fig. 6a plots the RD curves of this model for different values of α. We can see that the high-rate and log-likelihood performance of the good and poor generative models are almost identical, whereas in the low-rate regime, the RD curves show a significant drop in the performance and thus successfully detect this failure mode of log-likelihood.

Our second manipulation is to damage the decoder by convolving its output with a Gaussian blur kernel. Fig. 6b shows the rate distortion curves for different radii of the Gaussian kernel. We can see that, in contrast to the mixture prior experiment, the high-rate performance of the VAE drops due to inability of the decoder to output sharp images. However, we can also see an "improvement" in the low-rate performance of the VAE. This is because the data distribution does not necessarily achieve the minimal distortion, and in fact, in the extremely low-rate regime, blurring appears to help by reducing the average Euclidean distance between low-rate reconstructions and the input images. This problem was observed by Theis et al. (2015) in the context of log-likelihood estimation; our current observation indicates that rate distortion analysis does not fix this problem.

Our two manipulations — bad priors and blurring — resulted in very different changes to the RD curves, suggesting that these curves provide a richer picture of the performance of generative models, compared to scalar metrics such as log-likelihoods or FID.

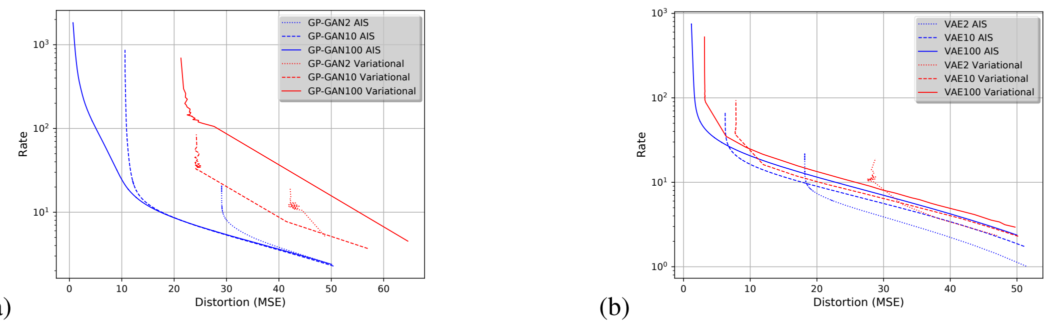
6.4. Beyond Pixelwise Mean Squared Error

The experiments discussed above all used pixelwise MSE as the distortion metric. However, for natural images, one could use more perceptually valid distortion metrics such as SSIM (Wang et al., 2004), MSSIM (Wang et al., 2003), or distances between deep features of a CNN (Johnson et al., 2016). Fig. 7 shows the RD curves of GANs, VAEs, and AAEs on the MNIST, using the MSE on the deep features of a CNN as distortion metric. In all cases, the qualitative behavior of the RD curves with this distortion metric closely matches the qualitative behaviors for pixelwise MSE. We can see from Fig. 7a that similar to the RD curves with MSE distortion, GANs with different depths and code sizes have the same low-rate performance, but as the model



(a) (b)

Figure 7. The RD curves of GANs, VAEs and AAEs with MSE distortion on the deep feature space. The behavior is qualitatively similar to the results for MSE in images (see Fig. 4 and Fig. 5), suggesting that the RD analysis is not particularly sensitive to the particular choice of metric. **(a)** GANs. **(b)** VAEs, GANs, and AAEs.



(a) (b)

Figure 8. Comparison of AIS and variational estimates of variational RD curves. **(a)** GANs. **(b)** VAEs.

gets deeper and wider, the RD curves are pushed down and extended to the left. Similarly, we can see from Fig. 7b that compared to GANs and AAEs, VAEs generally have a better high-rate performance, but worse low-rate performance. The fact that the qualitative behaviors of RD curves with this metric closely match those of pixelwise MSE indicates that the results of our analysis are not overly sensitive to the particular choice of distortion metric.

6.5. Comparison of AIS and Variational RD Curves

In this section, we compare the AIS estimate with the variational estimate of variational RD curves of GANs and VAEs. We conducted an experiment where we trained an amortized variational inference network to estimate the rate and distortion at different values of β (Fig. 8). Since AIS obtains an upper bound on the RD curve (Prop. 4), we know from this figure that the AIS estimate is more accurate than the variational estimate for both VAEs and GANs. In the case of GANs (Fig. 8a), we can see that the variational method completely fails to provide any useful bounds and even predicts a wrong ordering for the comparison of GANs, e.g., it predicts GAN-100 is strictly worse than the low-dimensional GANs. In the case of VAEs (Fig. 8b), the AIS outperforms the variational method, but by a smaller margin. We believe this is because VAEs learn a model whose true posterior could fit to the factorized Gaussian approximation of the posterior, and thus variational approximation could also provide useful bounds for estimating the RD curves. However, in the case of GANs the true posterior is highly multi-modal and thus variational approximations fail to provide useful bounds.

7. Conclusion

In this work, we studied rate distortion approximations for evaluating different generative models such as VAEs, GANs and AAEs. We showed that rate distortion curves provide more insights about the model than the log-likelihood alone while requiring roughly the same computational cost. For instance, we observed that while VAEs with larger code size can generally achieve better lossless compression rates, their performances drop at lossy compression in the low-rate regime. Conversely, expanding the capacity of GANs appears to bring substantial reductions in distortion at the high-rate regime without any corresponding deterioration in quality in the low-rate regime. This may help explain the success of large GAN architectures (Brock et al., 2019; Karras et al., 2018a;b). We also found that increasing the capacity of GANs by increasing the code size (width) has a very different effect than increasing the depth. The former extends the RD curves leftwards, while the latter pushes the curves down. Overall, lossy compression yields a richer and more complete picture of the distribution modeling performance of generative models. The ability to quantitatively measure performance tradeoffs should lead to algorithmic insights which can improve these models.

8. Acknowledgements

Alireza Makhzani and Roger Grosse acknowledge support from the CIFAR Canadian AI Chairs program. Sicong Huang acknowledges the support related to computing from JinSung Kang at Borealis AI.

References

Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. Fixing a broken elbo. In *International Conference on Machine Learning*, pp. 159–168, 2018.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

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

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

Ballé, J., Minnen, D., Singh, S., Hwang, S. J., and Johnston, N. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Cao, Y., Ding, G. W., Lui, K. Y.-C., and Huang, R. Improving GAN training via binarized representation entropy (bre) regularization. *ICLR*, 2018. accepted as poster.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2012.

Danihelka, I., Lakshminarayanan, B., Uria, B., Wierstra, D., and Dayan, P. Comparison of maximum likelihood and gan-based training of real nvps. *arXiv preprint arXiv:1705.05263*, 2017.

Domke, J. and Sheldon, D. R. Importance weighting and variational inference. In *Advances in neural information processing systems*, pp. 4470–4479, 2018.

Frey, B. J. and Hinton, G. E. Free energy coding. In *Proceedings of Data Compression Conference-DCC'96*, pp. 73–81. IEEE, 1996.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Grosse, R. B., Ghahramani, Z., and Adams, R. P. Sandwiching the marginal likelihood using bidirectional monte carlo. *arXiv preprint arXiv:1511.02543*, 2015.

Grover, A., Dhar, M., and Ermon, S. Flow-GAN: Combining maximum likelihood and adversarial learning in generative models. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein GANs. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 6626–6637. Curran Associates, Inc., 2017.

Hinton, G. and Van Camp, D. Keeping neural networks simple by minimizing the description length of the weights. In *in Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory*. Citeseer, 1993.

Huang, G., Yuan, Y., Xu, Q., Guo, C., Sun, Y., Wu, F., and Weinberger, K. An empirical study on evaluation metrics of generative adversarial networks, 2018.

Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018a.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018b.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Neal, R. M. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

Neal, R. M. Estimating ratios of normalizing constants using linked importance sampling. *arXiv preprint math/0511216*, 2005.

Neal, R. M. et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain Monte Carlo*, 2(11):2, 2011.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Rezende, D. J. and Viola, F. Taming vaes. *arXiv preprint arXiv:1810.00597*, 2018.

Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, pp. 5228–5237, 2018.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Salimans, T., Zhang, H., Radford, A., and Metaxas, D. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018.

Theis, L., Oord, A. v. d., and Bethge, M. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

Theis, L., Shi, W., Cunningham, A., and Huszár, F. Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395*, 2017.

Townsend, J., Bird, T., and Barber, D. Practical lossless compression with latent variables using bits back coding. *arXiv preprint arXiv:1901.04866*, 2019.

Wallace, C. S. Classification by minimum-message-length inference. In *International Conference on Computing and Information*, pp. 72–81. Springer, 1990.

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

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Wu, Y., Burda, Y., Salakhutdinov, R., and Grosse, R. On the quantitative analysis of decoder-based generative models. *arXiv preprint arXiv:1611.04273*, 2016.

Yang, Y., Bamler, R., and Mandt, S. Improving inference for neural image compression. *arXiv preprint arXiv:2006.04240*, 2020.