Appendix: Evaluating Lossy Compression Rates of Deep Generative Models

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

A. Proofs

A.1. Proof of Prop. 1.

Proof of Prop. 1a. As D increases, $\mathcal{R}_p(D)$ is minimized over a larger set, so $\mathcal{R}_p(D)$ is non-increasing function of D.

The distortion $\mathbb{E}_{q(\mathbf{x},\mathbf{z})}[d(\mathbf{x}, f(\mathbf{z}))]$ is a linear function of the channel conditional distribution $q(\mathbf{z}|\mathbf{x})$. The mutual information is a convex function of $q(\mathbf{z}|\mathbf{x})$. The $\mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z}))$ is also convex function of $q(\mathbf{z})$, which itself is a linear function of $q(\mathbf{z}|\mathbf{x})$. Thus $\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ is a convex function of $q(\mathbf{z}|\mathbf{x})$. Suppose for the distortions D_1 and D_2, $q_1(\mathbf{z}|\mathbf{x})$ and $q_2(\mathbf{z}|\mathbf{x})$ achieve the optimal rates in Eq. 6 respectively. Suppose the conditional $q_\lambda(\mathbf{z}|\mathbf{x})$ is defined as $q_\lambda(\mathbf{z}|\mathbf{x}) = \lambda q_1(\mathbf{z}|\mathbf{x}) + (1-\lambda)q_2(\mathbf{z}|\mathbf{x})$. The $\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ objective that the conditional $q_\lambda(\mathbf{z}|\mathbf{x})$ achieves is $\mathcal{I}_\lambda(\mathbf{z};\mathbf{x}) + \mathrm{KL}(q_\lambda(\mathbf{z})\|p(\mathbf{z}))$, and the distortion D_λ that this conditional achieves is $D_\lambda = \lambda D_1 + (1-\lambda)D_2$. Now we have

$$\begin{aligned}
\mathcal{R}_p(D_\lambda) &\le \mathcal{I}_\lambda(\mathbf{z};\mathbf{x}) + \mathrm{KL}(q_\lambda(\mathbf{z})\|p(\mathbf{z})) && (19)\\
&\le \lambda\mathcal{I}_1(\mathbf{z};\mathbf{x}) + \lambda\mathrm{KL}(q_1(\mathbf{z})\|p(\mathbf{z})) + (1-\lambda)\mathcal{I}_2(\mathbf{z};\mathbf{x}) + (1-\lambda)\mathrm{KL}(q_2(\mathbf{z})\|p(\mathbf{z})) && (20)\\
&= \lambda\mathcal{R}_p(D_1) + (1-\lambda)\mathcal{R}_p(D_2) && (21)
\end{aligned}$$

which proves the convexity of $\mathcal{R}_p(D)$.

Alternative Proof of Prop. 1a. We know that $\mathbb{E}_{p_d(\mathbf{x})}\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ is a convex function of $q(\mathbf{z}|\mathbf{x})$, and $\mathbb{E}_{q(\mathbf{x},\mathbf{z})}[d(\mathbf{x}, f(\mathbf{z}))]$ is a linear and thus convex function of $q(\mathbf{z}|\mathbf{x})$. As the result, the following optimization problem is a convex optimization problem.

$$\min_{q(\mathbf{z}|\mathbf{x})} \mathbb{E}_{p_d(\mathbf{x})}\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})) \quad s.t.\ \mathbb{E}_{q(\mathbf{x},\mathbf{z})}[d(\mathbf{x}, f(\mathbf{z}))] \le 0. \tag{22}$$

The rate distortion function $\mathcal{R}_p(D)$ is the perturbation function of the convex optimization problem of Eq. 22. The convexity of $\mathcal{R}_p(D)$ follows from the fact that the perturbation function of any convex optimization problem is a convex function (Boyd & Vandenberghe, 2004).

Proof of Prop. 1b. We have

$$\begin{aligned}
\min_{p(\mathbf{z})} \mathcal{R}_p(D) &= \min_{p(\mathbf{z})} \min_{q(\mathbf{z}|\mathbf{x}):\mathbb{E}[d(\mathbf{x},f(\mathbf{z}))]\le D} \mathcal{I}(\mathbf{x};\mathbf{z}) + \mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z})) && (23)\\
&= \min_{q(\mathbf{z}|\mathbf{x}):\mathbb{E}[d(\mathbf{x},f(\mathbf{z}))]\le D} \min_{p(\mathbf{z})} \mathcal{I}(\mathbf{x};\mathbf{z}) + \mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z})) && (24)\\
&= \min_{q(\mathbf{z}|\mathbf{x}):\mathbb{E}[d(\mathbf{x},f(\mathbf{z}))]\le D} \mathcal{I}(\mathbf{x};\mathbf{z}) && (25)\\
&= \mathcal{R}(D). && (26)
\end{aligned}$$

where in Eq. 24, we have used the fact that for any function $f(x, y)$, we have

$$\min_x \min_y f(x,y) = \min_y \min_x f(x,y) = \min_{x,y} f(x,y), \tag{27}$$

[*]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.

and in Eq. 25, we have used the fact that $\mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z}))$ is minimized when $p(\mathbf{z}) = q(\mathbf{z})$.

Proof of Prop. 1c. In Prop. 1a, we showed that $\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ is a convex function of $q(\mathbf{z}|\mathbf{x})$, and that the distortion is a linear function of $q(\mathbf{z}|\mathbf{x})$. So the summation of them in Eq. 10 will be a convex function of $q(\mathbf{z}|\mathbf{x})$. The unique global optimum of this convex optimization can be found by rewriting Eq. 10 as

$$\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})) + \beta\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[d(\mathbf{x}, f(\mathbf{z}))] = \mathrm{KL}\big(q(\mathbf{z}|\mathbf{x})\|\frac{1}{Z(\mathbf{x})}p(\mathbf{z})\exp(-\beta d(\mathbf{x}, f(\mathbf{z})))\big) - \log Z_\beta(\mathbf{x}) \tag{28}$$

where $Z_\beta(\mathbf{x}) = \int p(\mathbf{z})\exp(-\beta d(\mathbf{x}, f(\mathbf{z})))d\mathbf{z}$. The minimum of Eq. 28 is obtained when the KL divergence is zero. Thus the optimal channel conditional is

$$q^*_\beta(\mathbf{z}|\mathbf{x}) = \frac{1}{Z_\beta(\mathbf{x})}p(\mathbf{z})\exp(-\beta d(\mathbf{x}, f(\mathbf{z}))). \tag{29}$$

A.2. Proof of Prop. 2.

Proof of Prop. 2a. $\mathcal{R}(D) \le \mathcal{R}_p(D)$ was proved in Prop. 1b. To prove the first inequality, note that the summation of rate and distortion is

$$\mathcal{R}_p(D) + D = \mathcal{I}(\mathbf{z};\mathbf{x}) + \mathbb{E}_{q^*(\mathbf{x},\mathbf{z})}[-\log p(\mathbf{x}|\mathbf{z})] \tag{30}$$

$$= \mathcal{H}_d + \mathbb{E}_{q^*(\mathbf{z})}\mathrm{KL}(q^*(\mathbf{x}|\mathbf{z})\|p(\mathbf{x}|\mathbf{z})) \ge \mathcal{H}_d. \tag{31}$$

where $q^*(\mathbf{x},\mathbf{z})$ is the optimal joint channel conditional, and $q^*(\mathbf{z})$ and $q^*(\mathbf{x}|\mathbf{z})$ are its marginal and conditional. The equality happens if there is a joint distribution $q(\mathbf{x},\mathbf{z})$, whose conditional $q(\mathbf{x}|\mathbf{z}) = p(\mathbf{x}|\mathbf{z})$, and whose marginal over $\mathbf{x}$ is $p_d(\mathbf{x})$. But note that such a joint distribution might not exist for an arbitrary $p(\mathbf{x}|\mathbf{z})$.

Proof of Prop. 2b. The proof can be easily obtained by using $d(\mathbf{x}, f(\mathbf{z})) = -\log p(\mathbf{x}|\mathbf{z})$ in Prop. 1c.

Proof of Prop. 2c. Based on Prop. 2b, at $\beta = 1$, we have

$$Z^*_\beta(\mathbf{x}) = \int p(\mathbf{z})p(\mathbf{x}|\mathbf{z})d\mathbf{z} = p(\mathbf{x}). \tag{32}$$

A.3. Proof of Prop. 3.

The set of pairs of $\big(R^{\mathrm{AIS}}_k(\mathbf{x}), D^{\mathrm{AIS}}_k(\mathbf{x})\big)$ are achievable variational rate distortion pairs (achieved by $q^{\mathrm{AIS}}_k(\mathbf{z}|\mathbf{x})$). Thus, by the definition of $\mathcal{R}_p(D)$, $\mathcal{R}^{\mathrm{AIS}}_p(D)$ falls in the achievable region of $\mathcal{R}_p(D)$ and, thus maintains an upper bound on it: $\mathcal{R}^{\mathrm{AIS}}_p(D) \ge \mathcal{R}_p(D)$.

A.4. Proof of Prop. 4.

AIS has the property that for any step k of the algorithm, the set of chains up to step k, and the partial computation of their weights, can be viewed as the result of a complete run of AIS with target distribution $q^*_k(\mathbf{z}|\mathbf{x})$. Hence, we assume without loss of generality that we are looking at a complete run of AIS (but our analysis applies to the intermediate distributions as well).

Let $q^{\mathrm{AIS}}_k(\mathbf{z}|\mathbf{x})$ denote the distribution of final samples produced by AIS. More precisely, it is a distribution encoded by the following procedure:

1. For each data point $\mathbf{x}$, we run M independent AIS chains, numbered $i = 1, \dots, M$. Let $\mathbf{z'}^i_k$ denotes the k-th state of the i-th chain. The joint distribution of the forward pass up to the k-th state is denoted by $q_f(\mathbf{z'}^i_1, \dots, \mathbf{z'}^i_k|\mathbf{x})$. The un-normalized joint distribution of the backward pass is denoted by

$$\tilde{q}_b(\mathbf{z'}^i_1, \dots, \mathbf{z'}^i_k|\mathbf{x}) = p(\mathbf{z'}^i_k)\exp(-\beta_k d(\mathbf{x}, f(\mathbf{z'}^i_k)))q_b(\mathbf{z'}^i_1, \dots, \mathbf{z'}^i_{k-1}|\mathbf{z'}^i_k, \mathbf{x}).$$

2. Compute the importance weights and normalized importance weights of each chain using

$$w^i_k = \frac{\tilde{q}_b(\mathbf{z'}^i_1, \dots, \mathbf{z'}^i_k|\mathbf{x})}{q_f(\mathbf{z'}^i_1, \dots, \mathbf{z'}^i_k|\mathbf{x})} \quad \text{and} \quad \tilde{w}^i_k = \frac{w^i_k}{\sum_{i=1}^M w^i_k}. \tag{33}$$

3. Select a chain index S with probability of $\tilde{w}_k^i$.
4. Assign the selected chain values to $(\mathbf{z}_1^1, \dots, \mathbf{z}_k^1)$:
$$(\mathbf{z}_1^1, \dots, \mathbf{z}_k^1) = (\mathbf{z'}_1^S, \dots, \mathbf{z'}_k^S). \tag{34}$$
5. Keep the unselected chain values and re-label them as $(\mathbf{z}_1^{2:M}, \dots, \mathbf{z}_k^{2:M})$:
$$(\mathbf{z}_1^{2:M}, \dots, \mathbf{z}_k^{2:M}) = (\mathbf{z'}_1^{-S}, \dots, \mathbf{z'}_k^{-S}). \tag{35}$$
where $-S$ denotes the set of all indices except the selected index S.
6. Return $\mathbf{z} = \mathbf{z}_k^1$.

More formally, the AIS distribution is

$$q_k^{\text{AIS}}(\mathbf{z}|\mathbf{x}) = \mathbb{E}_{\prod_{i=1}^M q_f(\mathbf{z'}_1^i, \dots, \mathbf{z'}_k^i|\mathbf{x})}[\sum_{i=1}^M \tilde{w}_k^i \delta(\mathbf{z} - \mathbf{z'}_k^i)]. \tag{36}$$

Using the AIS distribution $q_k^{\text{AIS}}(\mathbf{z}|\mathbf{x})$ defined as above, we define the AIS distortion $D_k^{\text{AIS}}(\mathbf{x})$ and the AIS rate $R_k^{\text{AIS}}(\mathbf{x}) = \text{KL}(q_k^{\text{AIS}}(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ as follows:

$$D_k^{\text{AIS}}(\mathbf{x}) = \mathbb{E}_{q_k^{\text{AIS}}(\mathbf{z}|\mathbf{x})}[d(\mathbf{x}, f(\mathbf{z}))] \tag{37}$$
$$R_k^{\text{AIS}}(\mathbf{x}) = \text{KL}(q_k^{\text{AIS}}(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})). \tag{38}$$

In order to estimate $R_k^{\text{AIS}}(\mathbf{x})$ and $D_k^{\text{AIS}}(\mathbf{x})$, we define

$$\hat{D}_k^{\text{AIS}}(\mathbf{x}) = \sum_{i=1}^M \tilde{w}_k^i d(\mathbf{x}, f(\mathbf{z'}_k^i)), \tag{39}$$
$$\hat{Z}_k^{\text{AIS}}(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^M w_k^i, \tag{40}$$
$$\hat{R}_k^{\text{AIS}}(\mathbf{x}) = -\log \hat{Z}_k^{\text{AIS}}(\mathbf{x}) - \beta_k \hat{D}_k^{\text{AIS}}(\mathbf{x}). \tag{41}$$

We would like to prove that

$$\mathbb{E}_{\prod_{i=1}^M q_f(\mathbf{z'}_1^i, \dots, \mathbf{z'}_k^i|\mathbf{x})}[\hat{D}_k^{\text{AIS}}(\mathbf{x})] = D_k^{\text{AIS}}(\mathbf{x}), \tag{42}$$
$$\mathbb{E}_{\prod_{i=1}^M q_f(\mathbf{z'}_1^i, \dots, \mathbf{z'}_k^i|\mathbf{x})}[\hat{R}_k^{\text{AIS}}(\mathbf{x})] \geq R_k^{\text{AIS}}(\mathbf{x}). \tag{43}$$

The proof of Eq. 42 is straightforward:

$$\begin{aligned}
D_k^{\text{AIS}}(\mathbf{x}) &= \mathbb{E}_{q_k^{\text{AIS}}(\mathbf{z}|\mathbf{x})}[d(\mathbf{x}, f(\mathbf{z}))], && (44)\\
&= \int q_k^{\text{AIS}}(\mathbf{z}|\mathbf{x}) d(\mathbf{x}, f(\mathbf{z})) d\mathbf{z}, && (45)\\
&= \int \mathbb{E}_{\prod_{i=1}^M q_f(\mathbf{z'}_1^i, \dots, \mathbf{z'}_k^i|\mathbf{x})}[\sum_{i=1}^M \tilde{w}_k^i \delta(\mathbf{z} - \mathbf{z'}_k^i)] d(\mathbf{x}, f(\mathbf{z})) d\mathbf{z}, && (46)\\
&= \mathbb{E}_{\prod_{i=1}^M q_f(\mathbf{z'}_1^i, \dots, \mathbf{z'}_k^i|\mathbf{x})} \sum_{i=1}^M \tilde{w}_k^i \big[\int \delta(\mathbf{z} - \mathbf{z'}_k^i) d(\mathbf{x}, f(\mathbf{z})) d\mathbf{z}\big], && (47)\\
&= \mathbb{E}_{\prod_{i=1}^M q_f(\mathbf{z'}_1^i, \dots, \mathbf{z'}_k^i|\mathbf{x})} \sum_{i=1}^M \tilde{w}_k^i d(\mathbf{x}, f(\mathbf{z'}_k^i)), && (48)\\
&= \mathbb{E}_{\prod_{i=1}^M q_f(\mathbf{z'}_1^i, \dots, \mathbf{z'}_k^i|\mathbf{x})}[\hat{D}_k^{\text{AIS}}(\mathbf{x})]. && (49)
\end{aligned}$$

Eq. 44 shows that $\hat{D}_k^{\text{AIS}}(\mathbf{x})$ is an unbiased estimate of $D_k^{\text{AIS}}(\mathbf{x})$. We also know $\log \hat{Z}_k^{\text{AIS}}(\mathbf{x})$ obtained by Eq. 40 is the estimate of the log partition function, and by the Jenson's inequality lower bounds in expectation the true log partition function: $\mathbb{E}[\log \hat{Z}_k^{\text{AIS}}(\mathbf{x})] \leq \log Z_k(\mathbf{x})$. After obtaining $\hat{D}_k^{\text{AIS}}(\mathbf{x})$ and $\log \hat{Z}_k^{\text{AIS}}(\mathbf{x})$, we use Eq. 41 to obtain $\hat{R}_k^{\text{AIS}}(\mathbf{x})$. Now, it remains to prove Eq. 43, which states that $\hat{R}_k^{\text{AIS}}(\mathbf{x})$ upper bounds the AIS rate term $R_k^{\text{AIS}}(\mathbf{x})$ in expectation.

Let $q_k^{\text{AIS}}(\mathbf{z}_1^{1:M}, \ldots, \mathbf{z}_k^{1:M}|\mathbf{x})$ denote the joint AIS distribution over all states of $\{\mathbf{z}_1^{1:M}, \ldots, \mathbf{z}_k^{1:M}\}$, defined in Eq. 34 and Eq. 35. It can be shown that (see Domke & Sheldon (2018))

$$q_k^{\text{AIS}}(\mathbf{z}_1^{1:M}, \ldots, \mathbf{z}_k^{1:M}|\mathbf{x}) = \frac{\tilde{q}_b(\mathbf{z}_1^1, \ldots, \mathbf{z}_k^1|\mathbf{x}) \prod_{i=2}^M q_f(\mathbf{z}_1^i, \ldots, \mathbf{z}_k^i|\mathbf{x})}{\hat{Z}_k^{\text{AIS}}(\mathbf{x})} \tag{50}$$

$$= \frac{p(\mathbf{z}_k^1) \exp(-\beta_k d(\mathbf{x}, f(\mathbf{z}_k^1))) \, q_b(\mathbf{z}_1^1, \ldots, \mathbf{z}_{k-1}^1|\mathbf{z}_k^1, \mathbf{x}) \prod_{i=2}^M q_f(\mathbf{z}_1^i, \ldots, \mathbf{z}_k^i|\mathbf{x})}{\hat{Z}_k^{\text{AIS}}(\mathbf{x})} \tag{51}$$

In order to simplify notation, suppose $\mathbf{z}_k^1$ is denoted by $\mathbf{z}$, and all the other variables $\{\mathbf{z}_1^{1:M}, \ldots, \mathbf{z}_{k-1}^{1:M}, \mathbf{z}_k^{2:M}\}$ are denoted by $\mathbf{V}$. Using this notation, we define $p(\mathbf{V}|\mathbf{z}, \mathbf{x})$ and $q_k^{\text{AIS}}(\mathbf{z}, \mathbf{V}|\mathbf{x})$ as follows:

$$p(\mathbf{V}|\mathbf{z}, \mathbf{x}) := q_b(\mathbf{z}_1^1, \ldots, \mathbf{z}_{k-1}^1|\mathbf{z}_k^1, \mathbf{x}) \prod_{i=2}^M q_f(\mathbf{z}_1^i, \ldots, \mathbf{z}_k^i|\mathbf{x}), \tag{52}$$

$$q_k^{\text{AIS}}(\mathbf{z}, \mathbf{V}|\mathbf{x}) := q_k^{\text{AIS}}(\mathbf{z}_1^{1:M}, \ldots, \mathbf{z}_k^{1:M}|\mathbf{x}) \tag{53}$$

Using the above notation, Eq. 51 can be re-written as

$$\hat{Z}_k^{\text{AIS}}(\mathbf{x}) = \frac{p(\mathbf{z}) \, \exp(-\beta_k d(\mathbf{x}, f(\mathbf{z}))) \, p(\mathbf{V}|\mathbf{z}, \mathbf{x})}{q_k^{\text{AIS}}(\mathbf{z}, \mathbf{V}|\mathbf{x})}. \tag{54}$$

Hence,

$$\begin{aligned} \mathbb{E}[\log \hat{Z}_k^{\text{AIS}}(\mathbf{x})] &= \mathbb{E}[\log p(\mathbf{z}) - \log q_k^{\text{AIS}}(\mathbf{z}, \mathbf{V}|\mathbf{x}) + \log p(\mathbf{V}|\mathbf{x}, \mathbf{z})] - \beta_k \mathbb{E}[d(\mathbf{x}, f(\mathbf{z}))] \\ &= -\text{KL}(q_k^{\text{AIS}}(\mathbf{z}, \mathbf{V}|\mathbf{x}) \| p(\mathbf{z}) p(\mathbf{V}|\mathbf{z}, \mathbf{x})) - \beta_k \mathbb{E}[d(\mathbf{x}, f(\mathbf{z}))] \\ &\leq -\text{KL}(q_k^{\text{AIS}}(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) - \beta_k \mathbb{E}[d(\mathbf{x}, f(\mathbf{z}))], \end{aligned} \tag{55}$$

where the inequality follows from the monotonicity of KL divergence. Rearranging terms, we bound the rate:

$$R_k^{\text{AIS}}(\mathbf{x}) = \text{KL}(q_k^{\text{AIS}}(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) \leq -\mathbb{E}[\log \hat{Z}_k^{\text{AIS}}(\mathbf{x})] - \beta_k \mathbb{E}[d(\mathbf{x}, f(\mathbf{z}))] = \mathbb{E}[\hat{R}_k^{\text{AIS}}(\mathbf{x})]. \tag{56}$$

Eq. 56 shows that $\hat{R}_k^{\text{AIS}}(\mathbf{x})$ upper bounds the AIS rate $R_k^{\text{AIS}}(\mathbf{x})$ in expectation. We also showed $\hat{D}_k^{\text{AIS}}(\mathbf{x})$ is an unbiased estimate of the AIS distortion $D_k^{\text{AIS}}(\mathbf{x})$. Hence, the estimated AIS rate distortion curve upper bounds the AIS rate distortion curve in expectation: $\mathbb{E}[\hat{\mathcal{R}}_p^{\text{AIS}}(D)] \geq \mathcal{R}_p^{\text{AIS}}(D)$.

B. Validation of AIS experiments

B.1. Analytical Solution of the Variational Rate Distortion Optimization on the Linear VAE

We compared our AIS results with the analytical solution of the variational rate distortion optimization on a linear VAE trained on MNIST as shown in Fig. 3.

In order to derive the analytical solution, we first find the optimal distribution $q_\beta^*(\mathbf{z}|\mathbf{x})$ from Prop. 2b. For simplicity, we assume a fixed identity covariance matrix I at the output of the conditional likelihood of the linear VAE decoder. In other words, the decoder of the VAE is simply: $\mathbf{x} = \mathbf{W}\mathbf{z} + \mathbf{b} + \epsilon$, where $\mathbf{x}$ is the observation, $\mathbf{z}$ is the latent code vector, $\mathbf{W}$ is the decoder weight matrix and $\mathbf{b}$ is the bias. The observation noise of the decoder is $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. It's easy to show that the conditional likelihood raised to a power β is: $p(\mathbf{x}|\mathbf{z})^\beta = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \mathbf{b}, \frac{1}{\beta}\mathbf{I})$. Then, $q_\beta^*(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mu_\beta, \mathbf{\Sigma}_\beta)$, where

$$\boldsymbol{\mu}_\beta = \mathbb{E}_{q_\beta^*(\mathbf{z}|\mathbf{x})}[\mathbf{z}] = \mathbf{W}^\intercal(\mathbf{W}\mathbf{W}^\intercal + \beta^{-1}\mathbf{I})^{-1}(\mathbf{x} - \mathbf{b}), \tag{57}$$

$$\mathbf{\Sigma}_\beta = \text{Cov}_{q_\beta^*(\mathbf{z}|\mathbf{x})}[\mathbf{z}] = \mathbf{I} - \mathbf{W}^\intercal(\mathbf{W}\mathbf{W}^\intercal + \beta^{-1}\mathbf{I})^{-1}\mathbf{W}. \tag{58}$$

For numerical stability, we can further simplify the above by taking the SVD of $\mathbf{W}$: Suppose we have $\mathbf{W} = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal$. We can use the Woodbury Matrix Identity to the matrix inversion operation to obtain

$$\boldsymbol{\mu}_\beta = \mathbf{V}\mathbf{R}_\beta\mathbf{U}^\intercal(\mathbf{x} - \mathbf{b}), \tag{59}$$
$$\boldsymbol{\Sigma}_\beta = \mathbf{V}\mathbf{S}_\beta\mathbf{V}^\intercal, \tag{60}$$

where $\mathbf{R}_\beta$ is a diagonal matrix with the i-th diagonal entry being $\frac{d_i}{d_i^2 + \frac{1}{\beta}}$ and $\mathbf{S}_\beta$ is a diagonal matrix with the i-th diagonal entry being $\frac{1}{\beta d_i^2 + 1}$, where d_i is the i-th diagonal entry of $\mathbf{D}$. The analytical solution for optimal rate is:

$$D_{KL}(q_\beta^*(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) = D_{KL}(\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)||\mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})) \tag{61}$$
$$= \frac{1}{2}\left(\mathrm{tr}\left(\boldsymbol{\Sigma}_\beta\right) + (-\boldsymbol{\mu}_\beta)^\intercal(-\boldsymbol{\mu}_\beta) - k + \ln\left((\det \boldsymbol{\Sigma}_\beta)^{-1}\right)\right) \tag{62}$$
$$= \frac{1}{2}\left(\mathrm{tr}\left(\boldsymbol{\Sigma}_\beta\right) + (\boldsymbol{\mu}_\beta)^\intercal(\boldsymbol{\mu}_\beta) - k - \ln\left(\det \boldsymbol{\Sigma}_\beta\right)\right) \tag{63}$$

Where k is the dimension of the latent code $\mathbf{z}$. With negative log-likelihood as the distortion metric, the analytically form of distortion term is:

$$\mathbb{E}_{q_\beta^*(\mathbf{z}|\mathbf{x})}\left[-\log p(\mathbf{x}|\mathbf{z})\right] = \int_{-\infty}^{\infty} -\log((2\pi)^{-k/2}\exp\left\{-\frac{1}{2}(\mathbf{x} - (\mathbf{W}\mathbf{z} + \mathbf{b}))^\intercal(\mathbf{x} - (\mathbf{W}\mathbf{z} + \mathbf{b}))\right\})q_\beta^*(\mathbf{z}|\mathbf{x})d\mathbf{z} \tag{64}$$
$$= -(\log((2\pi)^{-k/2}) + \frac{1}{2}\int_{-\infty}^{\infty}\left\{(\mathbf{x} - (\mathbf{W}\mathbf{z} + \mathbf{b}))^\intercal(\mathbf{x} - (\mathbf{W}\mathbf{z} + \mathbf{b}))\right\}q_\beta^*(\mathbf{z}|\mathbf{x})d\mathbf{z} \tag{65}$$
$$= \frac{k}{2}\log(2\pi) + \frac{1}{2}(\mathbf{x} - \mathbf{b})^\intercal(\mathbf{x} - \mathbf{b}) - (\mathbf{W}\boldsymbol{\mu}_\beta)^\intercal(\mathbf{x} - \mathbf{b}) + \frac{1}{2}\mathbb{E}_{q_\beta^*(\mathbf{z}|\mathbf{x})}\left[(\mathbf{W}\mathbf{z})^\intercal(\mathbf{W}\mathbf{z})\right] \tag{66}$$

where $\mathbb{E}_{q_\beta^*(\mathbf{z}|\mathbf{x})}\left[(\mathbf{W}\mathbf{z})^\intercal(\mathbf{W}\mathbf{z})\right]$ can be obtained by change of variable: Let $\mathbf{y} = \mathbf{W}\mathbf{z}$, then:

$$\mathbb{E}_{q^*(y)}\left[\mathbf{y}\right] = \mathbf{W}\boldsymbol{\mu}_\beta = \mathbf{U}(\mathbf{I} - \mathbf{S}_\beta)\mathbf{U}^\intercal(\mathbf{x} - \mathbf{b}) \tag{67}$$
$$\mathrm{Cov}_{q^*(y)}\left[\mathbf{y}\right] = \mathbf{W}\boldsymbol{\Sigma}_\beta\mathbf{W}^\intercal = \mathbf{U}\mathbf{D}\mathbf{S}_\beta\mathbf{D}\mathbf{U}^\intercal \tag{68}$$
$$\mathbb{E}_{q_\beta^*(\mathbf{z}|\mathbf{x})}\left[(\mathbf{W}\mathbf{z})^\intercal(\mathbf{W}\mathbf{z})\right] = \mathbb{E}_{q^*(y)}\left[\mathbf{y}^\intercal\mathbf{y}\right] = \mathbb{E}_{q^*(y)}\left[\mathbf{y}\right]^\intercal\mathbb{E}_{q^*(y)}\left[\mathbf{y}\right] + \mathrm{tr}(\mathrm{Cov}_{q^*(y)}\left[\mathbf{y}\right]) \tag{69}$$

B.2. The BDMC Gap

We evaluated the tightness of the AIS estimate by computing the BDMC gaps using the same AIS settings. Fig. 9, shows the BDMC gaps at different compression rates for the VAE, GAN and AAE experiments on the MNIST dataset. 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.

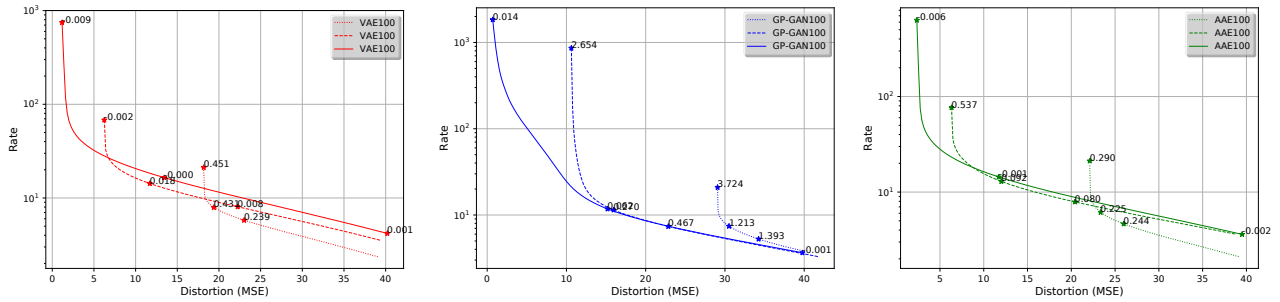
C. Experimental Details

The code for reproducing all the experiments of this paper can be found at: https://github.com/huangsicong/rate_distortion.

C.1. Datasets and Models

We used MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky & Hinton, 2009) datasets in our experiments.

Real-Valued MNIST. For the VAE experiments on the real-valued MNIST dataset (Fig. 5a), we used the "VAE-50" architecture described in (Wu et al., 2016), and only changed the code size in our experiments. The decoder variance is a global parameter learned during the training. The network was trained for 1000 epochs with the learning rate of 0.0001 using the Adam optimizer (Kingma & Ba, 2014), and the checkpoint with the best validation loss was used for the rate distortion evaluation.



(a) VAEs (b) GANs (c) AAEs

Figure 9. The BDMC gaps annotated on estimated AIS Variational Rate Distortion curves of (a) VAEs, (b) GANs, and (c) AAEs.

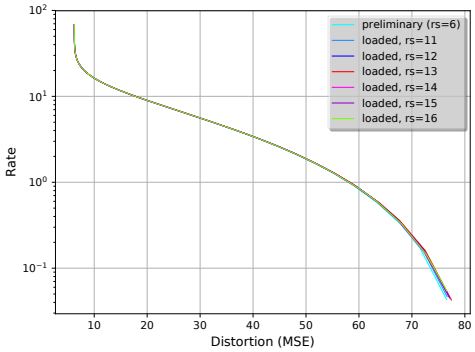


Figure 10. The variational rate distortion curves obtained by adaptively tuning the HMC parameters in the preliminary run, and preloading the HMC parameters in the second formal run. "rs" in the legend indicates the random seed used in the second run.

For the GAN experiments on MNIST (Fig. 4a), we used the "GAN-50" architecture described in (Wu et al., 2016). In order to stabilize the training dynamic, we used the gradient penalty (GP) (Salimans et al., 2016). In our deep architectures, we used code sizes of $d \in \{2, 5, 10, 100\}$ and three hidden layers each having 1024 hidden units to obtain the following GAN models: Deep-GAN2, Deep-GAN5, Deep-GAN10 and Deep-GAN100. The shallow GANs architectures are similar to the deep architectures but with one layer of hidden units.

CIFAR-10. For the CIFAR-10 experiments (Fig. 4b), we experimented with different GAN models such as DCGAN (Radford et al., 2015), DCGAN with Gradient Penalty (GP-GAN) (Gulrajani et al., 2017), Spectral Normalization (SN-GAN) (Miyato et al., 2018), and DCGAN with Binarized Representation Entropy Regularization (BRE-GAN) (Cao et al., 2018). The numbers at the end of each GAN name in Fig. 4b indicate the code size.

C.2. AIS Settings for RD Curves

We evaluated each RD curve at 2000 points corresponding to different values of β, with $N \gg 2000$ intermediate distributions. We used a sigmoid temperature schedule as used in Wu et al. (2016). We used $\beta_{\text{max}} \approx 3000$ for 100 dimensional models (GAN100, VAE100, and AAE100), and used $\beta_{\text{max}} \approx 36000$ for the rest of the models (2, 5 and 10 dimensional). For the 2, 5 and 10 dimensional models, we used $N = 60000$ intermediate distributions. For 100 dimensional models, we used $N = 1600000$ intermediate distributions in order to obtain small BDMC gaps. We used 20 leap frog steps for HMC, 40 independent chains, on a single batch of 50 images. On the MNIST dataset, we also tested with a larger batch size of 500 MNIST images, but did not observe a significant difference in average rates and distortions. On a P100 GPU, for MNIST, it takes 4-7 hours to compute an RD curve with $N = 60000$ intermediate distributions and takes around 7 days for $N = 160000$ intermediate distributions. For all of the CIFAR experiments, we used the temperature schedule with $N = 60000$ intermediate distributions, and each experiment takes about 7 days to complete.

C.3. Adaptive Tuning of HMC Parameters.

While running the AIS chain, the parameters of the HMC kernel cannot be adaptively tuned, since it would violate the Markovian property of the chain. So in order to be able to adaptively tune HMC parameters such as the number of leapfrog steps and the step size, in all our experiments, we first do a preliminary run where the HMC parameters are adaptively tuned to yield an average acceptance probability of 65% as suggested in Neal (2001). Then in the second "formal" run, we pre-load and fix the HMC parameters found in the preliminary run, and start the chain with a new random seed to obtain our final results. Interestingly, we observed that the difference in the RD curves obtained from the preliminary run and the formal runs with different random seeds is very small, as shown in Fig. 10. This figure shows that the AIS with the HMC kernel is robust against different choices of random seeds for approximating the RD curve of VAE10.

D. High-Rate vs. Low-Rate Reconstructions

In this section, we visualize the high-rate ($\beta \approx 3500$) and low-rate ($\beta = 0$) reconstructions of the MNIST images for VAEs, GANs and AAEs with different hidden code sizes. The qualitative results are shown in Fig. 11 and Fig. 12, which is consistent with the quantitative results presented in the experiment section of the paper.

(a) Original MNIST test images

(b) Low Rate VAE2

(c) Low Rate AAE2

(d) Low Rate GAN2

(e) Low Rate VAE10

(f) Low Rate AAE10

(g) Low Rate GAN10

(h) Low Rate VAE100

(i) Low Rate AAE100

(j) Low Rate GAN100

Figure 11. Low-rate reconstructions ($\beta = 0$) of VAEs, GANs and AAEs on MNIST.

(a) Original MNIST test images.

(b) High Rate VAE2

(c) High Rate AAE2

(d) High Rate GAN2

(e) High Rate VAE10

(f) High Rate AAE10

(g) High Rate GAN10

(h) High Rate VAE100

(i) High Rate AAE100

(j) High Rate GAN100

Figure 12. High-rate reconstructions (β_{max}) of VAEs, GANs and AAEs on MNIST. $\beta_{\text{max}} = 3333$ for 100 dimensional models, and $\beta_{\text{max}} = 36000$ for the 2 and 10 dimensional models.