IMPROVING MUTUAL INFORMATION ESTIMATION WITH ANNEALED AND ENERGY-BASED BOUNDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mutual information (MI) is a fundamental quantity in information theory and machine learning. However, direct estimation of mutual information is intractable, even if the true joint probability density for the variables of interest is known, as it involves estimating a potentially high dimensional log partition function. In this work, we view mutual information estimation from the perspective of importance sampling. Since naive importance sampling with the marginal density as a proposal requires exponential sample complexity in the true mutual information, we propose Multi-Sample Annealed Importance Sampling (AIS) bounds on mutual information to bridge between the marginal and joint distributions. In settings where the full joint distribution is available, we provide lower and upper bounds that can tightly estimate large values of MI in our experiments. In settings where only a single marginal distribution is known, we improve upon existing variational methods by directly optimizing a tighter lower bound on MI, using energy-based training to estimate gradients and multi-sample AIS for evaluation. Our methods are particularly suitable for evaluating MI in deep generative models, since explicit forms for the marginal or joint densities are often available. We evaluate our bounds on estimating the MI of VAEs and GANs trained on the MNIST and CIFAR datasets, and showcase significant gains over existing bounds in these challenging settings with high ground truth MI.

1 INTRODUCTION

Mutual information (MI) is among the most general measures of dependence between two random variables. Among many other applications in machine learning, mutual information has been used for both training (Alemi et al., 2016; 2018; Chen et al., 2016; Zhao et al., 2018) and evaluating (Alemi & Fischer, 2018; Huang et al., 2020) generative models. Furthermore, a recent success in neural network function approximation has encouraged a wave of variational or contrastive methods for representation learning (van den Oord et al., 2018) or MI estimation from samples alone (Belghazi et al., 2018; Poole et al., 2019). However, McAllester & Stratos (2020) have shown strong theoretical limitations on any estimator which does not have access to an analytical form of at least one marginal distribution. In light of these limitations, we consider MI estimation in settings where a single marginal or the full joint distribution are available.

In this work, we propose novel Multi-Sample Annealed Importance Sampling (AIS) bounds which can tightly estimate large values of MI when the full joint distribution is known. We motivate these bounds by viewing MI estimation from the perspective of importance sampling, which also suggests novel variational approaches for the case where only a single marginal is known. First, our *Generalized* IWAE (GIWAE) bounds highlight how multi-sample contrastive estimation can complement variational learning to improve MI lower bounds, and includes INFONCE (van den Oord et al., 2018; Poole et al., 2019) and importance-weighted autoencoder (IWAE) (Burda et al., 2016; Sobolev & Vetrov, 2019) bounds as special cases. We also propose MINE-AIS, which uses an improved energy-based lower bound compared with Mutual Information Neural Estimation (MINE) (Belghazi et al., 2018) and corresponds to the infinite-sample limit of the GIWAE framework. We propose an efficient training procedure using Markov Chain Monte Carlo (MCMC) and utilize our Multi-Sample AIS evaluation bounds to show notable improvement over finite-sample methods for lower bounding MI. In Fig. 1, we summarize the MI bounds discussed in this paper and the relationships between them.

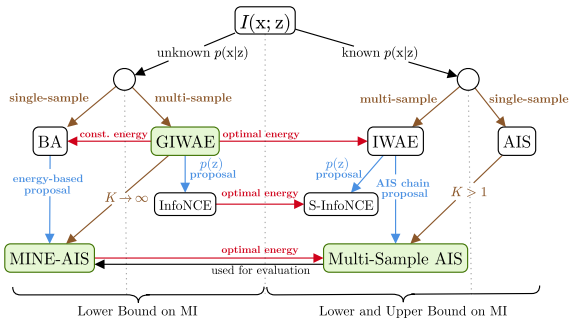


Figure 1: Schematic of MI bounds discussed in this paper. Green shading indicates our contributions, while columns and gold labels indicate single- or multi-sample bounds. Blue arrows indicate special cases using the indicated proposal distribution. Relationships based on learned energy functions are indicated by red arrows. We obtain only lower bounds on MI with unknown $p(\mathbf{x}|\mathbf{z})$, but both upper and lower bounds with known $p(\mathbf{x}|\mathbf{z})$. All bounds require a known marginal $p(\mathbf{z})$ for evaluation, apart from (Structured) INFO-NCE.

1.1 PROBLEM SETTING

The mutual information between two random variables $\mathbf{x}$ and $\mathbf{z}$ with joint distribution $p(\mathbf{x}, \mathbf{z})$ is

$$I(\mathbf{x};\mathbf{z}) = \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})p(\mathbf{z})}\right] = \mathcal{H}(\mathbf{x}) - \mathcal{H}(\mathbf{x}|\mathbf{z}) = \mathbb{E}_{p(\mathbf{x},\mathbf{z})}[\log p(\mathbf{x}|\mathbf{z})] - \mathbb{E}_{p(\mathbf{x})}[\log p(\mathbf{x})], \quad (1)$$

where $\mathcal{H}(\mathbf{x}|\mathbf{z})$ denotes the conditional entropy $-\mathbb{E}_{p(\mathbf{x},\mathbf{z})} \log p(\mathbf{x}|\mathbf{z})$. We primarily focus on bounds that assume either a single marginal distribution or the full joint distribution are available. A natural setting where the full joint distribution is available is estimating the mutual information in deep generative models between the latent variables, with a known prior $\mathbf{z} \sim p(\mathbf{z})$, and data $\mathbf{x} \sim p(\mathbf{x})$ simulated from the model (Alemi & Fischer, 2018). [1] Settings where a single marginal is available appear in simulation-based inference (Cranmer et al., 2020), where information about input parameters θ is known and a simulator can generate $\mathbf{x}$ for a given θ, but the likelihood $p(\mathbf{x}|\theta)$ is intractable.

While sampling from the posterior $p(\mathbf{z}|\mathbf{x})$ for an arbitrary $\mathbf{x}$ is often intractable, we can obtain a single posterior sample for $\mathbf{x} \sim p(\mathbf{x}), \mathbf{z} \sim p(\mathbf{z}|\mathbf{x})$ in cases where samples from the joint distribution of interest $p(\mathbf{x})p(\mathbf{z}|\mathbf{x})$ are available. Throughout this paper, we will refer to bounds which involve only a single posterior sample as practical, and those involving multiple posterior samples as impractical.

When the conditional $p(\mathbf{x}|\mathbf{z})$ is tractable to sample and evaluate, simple Monte Carlo sampling provides an unbiased, low variance estimate of the conditional entropy term in Eq. 1. In this case, the difficulty of MI estimation reduces to estimating the log partition function, for which importance sampling (IS) based methods are among the most well studied and successful solutions.

2 MULTI-SAMPLE BOUNDS ON MUTUAL INFORMATION

In this section, we review single- and multi-sample MI bounds from the perspective of importance sampling. This provides insight into existing bounds and will provide foundation for our primary contributions in Sec. 3 and 4.

2.1 SIMPLE IMPORTANCE SAMPLING BOUNDS

For an arbitrary conditional distribution $q_\theta(\mathbf{z}|\mathbf{x})$, we can obtain lower and upper bounds on $\log p(\mathbf{x})$:

$$\mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{q_\theta(\mathbf{z}|\mathbf{x})}\right] \leq \log p(\mathbf{x}) \leq \mathbb{E}_{p(\mathbf{z}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{q_\theta(\mathbf{z}|\mathbf{x})}\right] \quad (2)$$

where the left-hand side is the familiar Evidence Lower Bound (ELBO) with a gap of $D_{\mathrm{KL}}[q_\theta(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})] = \log p(\mathbf{x}) - \mathrm{ELBO}(q_\theta)$. We refer to the right-hand side as the Evidence Upper Bound (EUBO), with gap $D_{\mathrm{KL}}[p(\mathbf{z}|\mathbf{x})||q_\theta(\mathbf{z}|\mathbf{x})] = \mathrm{EUBO}(q_\theta) - \log p(\mathbf{x})$. Throughout this paper, we will derive bounds which correspond to extended-state space versions of Eq. 2 (see App. A).

By plugging the importance sampling lower and upper bounds on $\log p(\mathbf{x})$ from Eq. 2 into Eq. 1, we obtain upper and lower bounds on mutual information as follows

$$I_{\mathrm{BA}_L}(q_\theta) = \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\theta(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right] \leq I(\mathbf{x};\mathbf{z}) \leq \mathbb{E}_{p(\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x})}\left[\log \frac{q_\theta(\mathbf{z}|\mathbf{x})}{p(\mathbf{x},\mathbf{z})}\right] - \mathcal{H}(\mathbf{x}|\mathbf{z}) = I_{\mathrm{BA}_U}(q_\theta) \quad (3)$$

The left hand side of Eq. 3 is the well-known Barber-Agakov (BA) bound (Barber & Agakov, 2003; Agakov, 2006), which we denote by $I_{\mathrm{BA}_L}(q_\theta) = I(\mathbf{x};\mathbf{z}) - D_{\mathrm{KL}}[p(\mathbf{z}|\mathbf{x})||q_\theta(\mathbf{z}|\mathbf{x})]$. We refer to the right hand side as the BA upper bound $I_{\mathrm{BA}_U}(q_\theta) = I(\mathbf{x};\mathbf{z}) + D_{\mathrm{KL}}[q_\theta(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})]$.

[1] An alternative, 'encoding' MI between the real data and the latent code is often of interest in representation learning (see App. F), but cannot be directly estimated using our methods due to the unavailability of densities for the data distribution $p_d(\mathbf{x})$ or the 'aggregated posterior' $q(\mathbf{z}) = \int p_d(\mathbf{x})q(\mathbf{z}|\mathbf{x})d\mathbf{x}$.

Relationship with CLUB When the prior is used as the proposal, $I_{\text{BA}_U}(p(\mathbf{z}))$ corresponds to the Contrastive Log-Ratio Upper Bound (CLUB) (Cheng et al., 2020) with known $p(\mathbf{x}|\mathbf{z})$.

2.2 Importance Weighted Autoencoder

For a given variational distribution $q_\theta(\mathbf{z}|\mathbf{x})$, IWAE bounds (Burda et al., 2016; Sobolev & Vetrov, 2019) improve upon simple importance sampling by considering multi-sample proposal and target distributions in an extended state space (Domke & Sheldon, 2018; Cremer et al., 2017). Consider a proposal $q_{\text{PROP}}^{\text{IWAE}}$ with K independent samples from the variational distribution $q_\theta(\mathbf{z}|\mathbf{x})$. The extended state space target $p_{\text{TGT}}^{\text{IWAE}}$ is a mixture distribution involving a single sample from the posterior $p(\mathbf{z}|\mathbf{x})$ or joint $p(\mathbf{x},\mathbf{z})$ distribution and $K-1$ samples from $q_\theta(\mathbf{z}|\mathbf{x})$

$$q_{\text{PROP}}^{\text{IWAE}}(\mathbf{z}^{(1:K)}|\mathbf{x}) = \prod_{k=1}^{K} q_\theta(\mathbf{z}^{(k)}|\mathbf{x}) \qquad p_{\text{TGT}}^{\text{IWAE}}(\mathbf{x},\mathbf{z}^{(1:K)}) = \frac{1}{K}\sum_{s=1}^{K} p(\mathbf{x},\mathbf{z}^{(s)}) \prod_{k=1,k\neq s}^{K} q_\theta(\mathbf{z}^{(k)}|\mathbf{x})\,. \tag{4}$$

The log ratio $\log \frac{p_{\text{TGT}}^{\text{IWAE}}(\mathbf{x},\mathbf{z}^{(1:K)})}{q_{\text{PROP}}^{\text{IWAE}}(\mathbf{z}^{(1:K)}|\mathbf{x})}$ reduces to familiar ratio appearing in the IWAE objective, while the ratio of normalization constants corresponds to $\log p(\mathbf{x})$. As in App. A, taking expectations under the proposal yields a lower bound, whereas taking expectations under the target yields an upper bound

$$\underbrace{\mathbb{E}_{q_\theta(\mathbf{z}^{(1:K)}|\mathbf{x})}\left[\log \frac{1}{K}\sum_{i=1}^{K}\frac{p(\mathbf{x},\mathbf{z}^{(k)})}{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}\right]}_{\text{ELBO}_{\text{IWAE}}(q_\theta,K)} \leq \log p(\mathbf{x}) \leq \underbrace{\mathbb{E}_{p(\mathbf{z}^{(1)}|\mathbf{x})q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \frac{1}{K}\sum_{i=1}^{K}\frac{p(\mathbf{x},\mathbf{z}^{(k)})}{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}\right]}_{\text{EUBO}_{\text{IWAE}}(q_\theta,K)}\,. \tag{5}$$

Note that the indices $\mathbf{z}^{(1:K)}$ denote a set of independent draws from a given distribution. As for the ELBO and EUBO, the gap in the lower and upper bounds are $D_{KL}[q_{\text{PROP}}^{\text{IWAE}}||p_{\text{TGT}}^{\text{IWAE}}]$ and $D_{KL}[p_{\text{TGT}}^{\text{IWAE}}||q_{\text{PROP}}^{\text{IWAE}}]$, respectively. See App. B for detailed derivations.

As discussed in Sec. 1.1, lower and upper bounds on the log partition function translate to upper and lower bounds, respectively, on MI when $p(\mathbf{x}|\mathbf{z})$ is known. We refer to these bounds arising from IWAE estimation as $I_{\text{IWAE}_L}(q_\theta,K)$ and $I_{\text{IWAE}_U}(q_\theta,K)$. For example, $I_{\text{IWAE}_L}(q_\theta,K)$ can be written as

$$I(\mathbf{x};\mathbf{z}) \geq I_{\text{IWAE}_L}(q_\theta,K) := \mathbb{E}_{p(\mathbf{x})p(\mathbf{z}^{(1)}|\mathbf{x})q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \frac{p(\mathbf{x}|\mathbf{z}^{(1)})}{\frac{1}{K}\sum_{i=1}^{K}\frac{p(\mathbf{x},\mathbf{z}^{(k)})}{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}}\right] \tag{6}$$

While it is well-known that increasing K leads to tighter IWAE bounds (Burda et al., 2016; Sobolev & Vetrov, 2019), we characterize the logarithmic rate of bias reduction for $I_{\text{IWAE}_L}(q_\theta,K)$ in the following proposition, with proof in App. B.2.

Proposition 2.1. IWAE *bounds on* MI *improve upon the* BA *bounds with the following relationships:*

$$I_{BA_L}(q_\theta) \leq I_{IWAE_L}(q_\theta,K) \leq I_{BA_L}(q_\theta) + \log K, \qquad I_{IWAE_U}(q_\theta,K) \leq I_{BA_U}(q_\theta). \tag{7}$$

2.3 Generalized IWAE

In this section, we consider a family of *Generalized* IWAE (GIWAE) lower bounds, which interpolate between $I_{\text{BA}_L}(q_\theta)$ and $I_{\text{IWAE}_L}(q_\theta,K)$ using a contrastive energy function $T_\phi(\mathbf{x},\mathbf{z})$. While similar bounds appear in (Lawson et al., 2019; Sobolev, 2019b), we provide a thorough discussion of special cases, and empirical analysis in the context of MI estimation in Sec. 5.2. We also show that our MINE-AIS method in Sec. 4 corresponds to the limiting behavior of GIWAE as $K \to \infty$.

To derive the GIWAE lower bound, we begin by further extending the state space of the IWAE target distribution in Eq. 4, using a uniform index variable $p(s) = \frac{1}{K}\,\forall s$ that specifies which sample $\mathbf{z}^{(s)} \sim p(\mathbf{z}|\mathbf{x})$ is drawn from the posterior

$$p_{\text{TGT}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)},s|\mathbf{x}) = \frac{1}{K}\,p(\mathbf{z}^{(s)}|\mathbf{x}) \prod_{k=1,k\neq s}^{K} q(\mathbf{z}^{(k)}|\mathbf{x})\,. \tag{8}$$

Note that marginalization over s leads to the IWAE target distribution $p_{\text{TGT}}^{\text{IWAE}}(\mathbf{z}^{(1:K)}|\mathbf{x})$. For the GIWAE extended state space proposal, consider a categorical index variable $q_{\text{PROP}}^{\text{GIWAE}}(s|\mathbf{z}^{(1:K)},\mathbf{x})$ drawn using self-normalized importance sampling (SNIS) with weights calculated by a learned energy function T_ϕ

$$q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)},s|\mathbf{x}) = \left(\prod_{k=1}^{K} q(\mathbf{z}^{(k)}|\mathbf{x})\right) q_{\text{PROP}}^{\text{GIWAE}}(s|\mathbf{z}^{(1:K)},\mathbf{x}) \text{ where } q_{\text{PROP}}^{\text{GIWAE}}(s|\mathbf{z}^{(1:K)},\mathbf{x}) = \frac{e^{T_\phi(\mathbf{x},\mathbf{z}^{(s)})}}{\sum_{k=1}^{K} e^{T_\phi(\mathbf{x},\mathbf{z}^{(k)})}}.$$

We provide detailed discussion of the GIWAE probabilistic interpretation in App. C.1. Taking the expectation under $p_{\text{TGT}}^{\text{GIWAE}}\mathbf{z}^{(1:K)}, s|\mathbf{x})$ of the log importance ratio $\log \frac{p_{\text{TGT}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)},s|\mathbf{x})}{q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)},s|\mathbf{x})}$ as in App. A, we obtain an upper bound on $\log p(\mathbf{x})$ and lower bound on MI as

$$I_{\text{GIWAE}_L}(q_\theta, T_\phi, K) = \underbrace{\mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\theta(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right]}_{I_{\text{BA}_L}(q_\theta)} + \underbrace{\mathbb{E}_{p(\mathbf{x})p(\mathbf{z}^{(1)}|\mathbf{x})q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \frac{e^{T_\phi(\mathbf{x},\mathbf{z}^{(1)})}}{\frac{1}{K}\sum_{i=1}^{K} e^{T_\phi(\mathbf{x},\mathbf{z}^{(k)})}}\right]}_{\text{contrastive term}\leq \log K} . \tag{9}$$

We see that the GIWAE lower bound decomposes into the sum of two terms, where the first is the BA variational lower bound for $q_\theta(\mathbf{z}|\mathbf{x})$ and the second is a contrastive term which distinguishes negative samples from $q_\theta(\mathbf{z}|\mathbf{x})$ and positive samples from $p(\mathbf{z}|\mathbf{x})$. Similarly to BA, GIWAE requires access to the analytical form of $p(\mathbf{z})$ to *evaluate* the bound on MI. However, if the goal is to *optimize* mutual information in applications such as representation learning, both the BA and GIWAE lower bounds can be used even if no marginal distribution is available. See App. F for more detailed discussion.

Relationship with BA Choosing constant $T_\phi(\mathbf{x},\mathbf{z}) = c$ means that the second term in GIWAE vanishes, so that we have $I_{\text{GIWAE}_L}(q, T_\phi = c, K) = I_{\text{BA}_L}(q_\theta)$ for all K. Similarly, the single sample $I_{\text{GIWAE}_L}(q, T_\phi, K = 1)$ also equals the BA lower bound for all T_ϕ.

Relationship with InfoNCE When the prior $p(\mathbf{z})$ is used in place of $q_\theta(\mathbf{z}|\mathbf{x})$, we can recognize the second term in Eq. 9 as the INFONCE contrastive lower bound (van den Oord et al., 2018; Poole et al., 2019), with $I_{\text{INFONCE}_L}(T_\phi, K) = I_{\text{GIWAE}_L}(p(\mathbf{z}), T_\phi, K)$. From this perspective, the GIWAE lower bound highlights how variational learning can complement contrastive bounds to improve MI estimation beyond the known $\log K$ limitations of INFONCE (van den Oord et al., 2018). However, using the prior as the proposal in INFONCE does allow the energy function to admit a bi-linear implementation $T(\mathbf{x},\mathbf{z}) = T_{\mathbf{x}}(\mathbf{x})T_{\mathbf{z}}(\mathbf{z})$, which requires only $N + K$ forward passes, rather than NK for GIWAE, where N is the batch size and K is the total number of positive and negative samples.

Relationship with IWAE The following proposition characterizes the relationship between the GIWAE lower bound in Eq. 9 and the IWAE lower bound on MI from Sec. 2.2. See App. C.4 for proofs.

Proposition 2.2. *For a given $q_\theta(\mathbf{z}|\mathbf{x})$, the optimal energy function equals the true log importance weights up to an arbitrary constant: $T^*(\mathbf{x},\mathbf{z}) = \log \frac{p(\mathbf{x},\mathbf{z})}{q_\theta(\mathbf{z}|\mathbf{x})} + c(\mathbf{x})$. With this choice of $T^*(\mathbf{x},\mathbf{z})$,*

$$I_{\text{GIWAE}_L}(q_\theta, T^*) = I_{\text{IWAE}_L}(q_\theta) . \tag{10}$$

Corollary 2.3. *For a given $q_\theta(\mathbf{z}|\mathbf{x})$, let $T_{\phi^*}(\mathbf{x},\mathbf{z})$ denote the energy function that maximizes the* GIWAE *lower bound when optimization is performed with a neural network parameterized by ϕ. Then*

$$I_{\text{BA}_L}(q_\theta) \leq I_{\text{GIWAE}_L}(q, T_{\phi^*}, K) \leq I_{\text{IWAE}_L}(q, K) \leq I_{\text{BA}_L}(q_\theta) + \log K . \tag{11}$$

While the GIWAE lower bound does not assume access to the full joint distribution $p(\mathbf{x},\mathbf{z})$, Prop. 2.2 suggests that the role of the energy function is to learn the true log importance weights for $q_\theta(\mathbf{z}|\mathbf{x})$. Thus, when $p(\mathbf{x}|\mathbf{z})$ is known, IWAE is always preferable to GIWAE. Cor. 2.3 shows that while GIWAE can improve upon BA, this improvement is at most logarithmic with respect to K.

Relationship with Structured INFONCE Since INFONCE and Structured INFONCE (S-INFONCE, see App. B.5, Poole et al. (2019)) are special cases of GIWAE and IWAE that use $p(\mathbf{z})$ as the variational distribution, Cor. 2.3 suggests the following relationship

$$0 \leq I_{\text{INFONCE}_L}(T_\phi, K) \leq I_{\text{S-INFONCE}_L}(K) \leq \log K . \tag{12}$$

3 MULTI-SAMPLE AIS BOUNDS FOR ESTIMATING MUTUAL INFORMATION

In the previous sections, we derived probabilistic interpretations of extended state space importance sampling using multiple samples K, as in IWAE. In this section, we first review AIS, which extends the state space using T intermediate distributions. We then show that these approaches are complementary, and derive two practical Multi-Sample AIS methods, which provide tighter bounds by combining insights from IWAE and AIS.

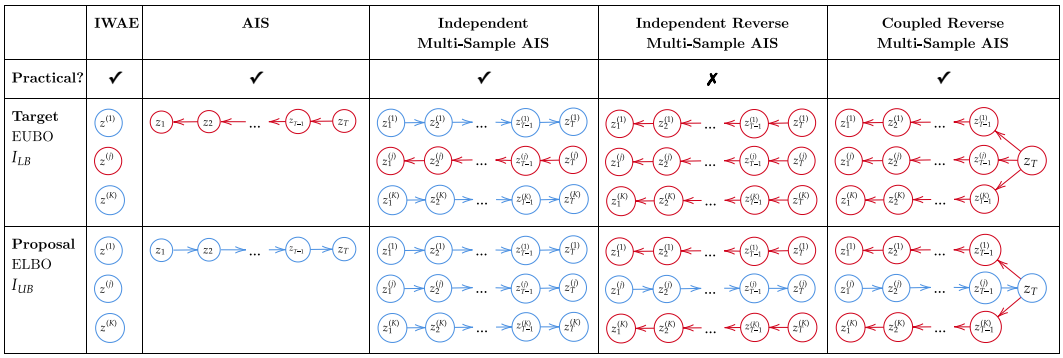


Figure 2: Extended state-space probabilistic interpretations of multi-sample AIS bounds. Forward chains are colored in blue, and backward chains are colored in red. Note that ELBOs and EUBOs are obtained by taking the expectation of the log unnormalized importance weights $\log p_{\text{TGT}}(\cdot)/q_{\text{PROP}}(\cdot)$ under either the proposal or target distribution, and can then be translated to MI bounds.

3.1 Annealed Importance Sampling Background

AIS (Neal, 2001) constructs a sequence of intermediate distributions $\{\pi_t(\mathbf{z})\}_{t=0}^T$, which bridge between a normalized initial distribution $\pi_0(\mathbf{z}|\mathbf{x})$ and target distribution $\pi_T(\mathbf{z}|\mathbf{x}) = p(\mathbf{z}|\mathbf{x})$ with unnormalized density $\pi_T(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}, \mathbf{z})$ and normalizing constant $\mathcal{Z}_T(\mathbf{x}) = p(\mathbf{x})$. A common choice for intermediate distributions is the geometric mixture path parameterized by $\{\beta_t\}_{t=0}^T$:

$$\pi_t(\mathbf{z}|\mathbf{x}) = \frac{\pi_0(\mathbf{z}|\mathbf{x})^{1-\beta_t}\, \pi_T(\mathbf{x}, \mathbf{z})^{\beta_t}}{\mathcal{Z}_t(\mathbf{x})} \quad \text{where} \quad \mathcal{Z}_t(\mathbf{x}) = \int \pi_0(\mathbf{z}|\mathbf{x})^{1-\beta_t}\, \pi_T(\mathbf{x}, \mathbf{z})^{\beta_t} d\mathbf{z}. \tag{13}$$

In the probabilistic interpretation of AIS, we consider an extended state space proposal $q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}|\mathbf{x})$, obtained by sampling from the initial $\pi_0(\mathbf{z}|\mathbf{x})$ and constructing transitions $\mathcal{T}_f(\mathbf{z}_{t+1}|\mathbf{z}_t, \mathbf{x})$ which leave $\pi_t(\mathbf{z}|\mathbf{x})$ invariant. The target distribution $p_{\text{TGT}}^{\text{AIS}}(\mathbf{z}_{0:T}|\mathbf{x})$ is given by running the reverse transitions $\mathcal{T}_r(\mathbf{z}_t|\mathbf{z}_{t+1}, \mathbf{x})$ starting from a target or posterior sample $\pi_T(\mathbf{z}|\mathbf{x})$, as shown in Fig. 2.

$$q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}|\mathbf{x}) = \pi_0(\mathbf{z}_0|\mathbf{x}) \prod_{t=0}^{T-1} \mathcal{T}_f(\mathbf{z}_{t+1}|\mathbf{z}_t) \qquad p_{\text{TGT}}^{\text{AIS}}(\mathbf{x}, \mathbf{z}_{0:T}) = \pi_T(\mathbf{x}, \mathbf{z}_T) \prod_{t=0}^{T-1} \mathcal{T}_r(\mathbf{z}_t|\mathbf{z}_{t+1}, \mathbf{x}) \tag{14}$$

As described in App. H, taking expectations of the log importance weights under the proposal and target yields a lower and upper bound on the log partition function $\log p(\mathbf{x})$. These single-chain lower and upper bounds translate to upper and lower bounds on MI, $I_{\text{AIS}_L}(\pi_0, T)$ and $I_{\text{AIS}_U}(\pi_0, T)$, which were suggested for MI estimation in the blog post of Sobolev (2019a).

3.2 Independent Multi-Sample AIS Bounds

To derive *Independent* Multi-Sample AIS (IM-AIS), we construct an extended state space proposal by running K independent AIS forward chains $\mathbf{z}_{0:T}^{(k)} \sim q_{\text{PROP}}^{\text{AIS}}$ in parallel. Similarly to the IWAE upper bound (Eq. 5), the extended state space target involves selecting a index s uniformly at random, and running a backward AIS chain $\mathbf{z}_{0:T}^{(s)} \sim p_{\text{TGT}}^{\text{AIS}}$ starting from a true posterior sample $\mathbf{z}_T \sim p(\mathbf{z}|\mathbf{x})$. The remaining $K-1$ samples are obtained by running forward AIS chains, as visualized in Fig. 2

$$q_{\text{PROP}}^{\text{IM-AIS}}(\mathbf{z}_{0:T}^{(1:K)}|\mathbf{x}) = \prod_{k=1}^{K} q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(k)}|\mathbf{x}), \qquad p_{\text{TGT}}^{\text{IM-AIS}}(\mathbf{z}_{0:T}^{(1:K)}, \mathbf{x}) = \frac{1}{K}\sum_{s=1}^{K} p_{\text{TGT}}^{\text{AIS}}(\mathbf{x}, \mathbf{z}_{0:T}^{(s)}) \prod_{\substack{k=1 \\ k\neq s}}^{K} q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(k)}|\mathbf{x}). \tag{15}$$

where $q_{\text{PROP}}^{\text{AIS}}$ and $p_{\text{TGT}}^{\text{AIS}}$ were defined in Eq. 14. Note that sampling from the extended state space target distribution is practical, as it only requires one sample from the true posterior distribution.

Taking the expectation of the log unnormalized density ratio under the proposal and target yields lower and upper bounds on $\log p(\mathbf{x})$, respectively (see App. A)

$$\underbrace{\mathbb{E}_{\mathbf{z}_{0:T}^{(1:K)} \sim q_{\text{PROP}}^{\text{AIS}}} \left[\log \frac{1}{K} \sum_{k=1}^{K} \frac{p_{\text{TGT}}^{\text{AIS}}(\mathbf{x}, \mathbf{z}_{0:T}^{(k)})}{q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(k)}|\mathbf{x})} \right]}_{\text{ELBO}_{\text{IM-AIS}}(\pi_0, K, T)} \leq \log p(\mathbf{x}) \leq \underbrace{\mathbb{E}_{\substack{\mathbf{z}_{0:T}^{(1)} \sim p_{\text{TGT}}^{\text{AIS}} \\ \mathbf{z}_{0:T}^{(2:K)} \sim q_{\text{PROP}}^{\text{AIS}}}} \left[\log \frac{1}{K} \sum_{k=1}^{K} \frac{p_{\text{TGT}}^{\text{AIS}}(\mathbf{x}, \mathbf{z}_{0:T}^{(k)})}{q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(k)}|\mathbf{x})} \right]}_{\text{EUBO}_{\text{IM-AIS}}(\pi_0, K, T)}. \tag{16}$$

which again have KL divergences as the gap in their bounds. Independent Multi-Sample AIS reduces to IWAE for $T = 1$, and reduces to single-sample AIS for $K = 1$. Both the upper and lower bounds become exact as $K \to \infty$ or $T \to \infty$, and translate to lower and upper bounds on MI as in Sec. 1.1.

Proposition 3.1 (**Complexity in** K)**.** *For a given T, the bias of the Independent Multi-Sample* AIS *lower bound on* MI*, $I_{\text{IM-AIS}_L}(\pi_0, T)$, decreases (at most) logarithmically in K.*

$$I_{\text{AIS}_L}(\pi_0, T) \leq I_{\text{IM-AIS}_L}(\pi_0, K, T) \leq I_{\text{AIS}_L}(\pi_0, T) + \log K \tag{17}$$

To characterize the bias reduction for AIS with increasing T, we prove the following proposition.

Proposition 3.2 (**Complexity in** T)**.** *Assuming perfect transitions and a geometric annealing path with linearly-spaced $\{\beta_t\}_{t=1}^T$, the sum of the gaps in the multi-sample* AIS *sandwich bounds on* MI*, $\delta^{T,K} = I_{\text{IM-AIS}_U}(\pi_0, K, T) - I_{\text{IM-AIS}_L}(\pi_0, K, T)$, reduces linearly with increasing T.*

This generalizes Thm. 1 of Grosse et al. (2013), which holds as $T \to \infty$. See App. H.3 for proof.

3.3 Coupled Reverse Multi-Sample AIS Bounds

We can change the role of the forward and backward annealing chains in Independent Multi-Sample AIS to obtain alternative bounds on the log partition function. We define *Independent Reverse Multi-Sample* AIS (IR-AIS) using the following proposal and target distribution, as shown in Fig. 2.

$$q_{\text{PROP}}^{\text{IR-AIS}}(\mathbf{x}, \mathbf{z}_{0:T}^{(1:K)}) = \frac{1}{K}\sum_{s=1}^{K} q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(s)}|\mathbf{x}) \prod_{k=1,k\neq s}^{K} p_{\text{TGT}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(k)}, \mathbf{x}) \qquad p_{\text{TGT}}^{\text{IR-AIS}}(\mathbf{z}_{0:T}^{(1:K)}, \mathbf{x}) = \prod_{k=1}^{K} p_{\text{TGT}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(k)}, \mathbf{x})$$

Note that partition function ratio is $\mathcal{Z}_{\text{TGT}}/\mathcal{Z}_{\text{PROP}} = p(\mathbf{x})^K/p(\mathbf{x})^{K-1} = p(\mathbf{x})$. Using these distributions, we derive $\log p(\mathbf{x})$ and MI bounds in App. J. However, these will be impractical in most settings since they require multiple true posterior samples (see Sec. 1.1).

To address this, we propose *Coupled* Reverse Multi-Sample AIS (CR-AIS). The extended state space target distribution in Fig. 2 initializes K backward chains from a single target sample $\mathbf{z}_T \sim \pi_T(\mathbf{z}|\mathbf{x})$, with remaining transitions denoted $p_{\text{TGT}}^{\text{AIS}}(\mathbf{z}_{0:T-1}|\mathbf{z}_T)$ since they are identical to standard AIS in Eq. 14.

$$p_{\text{TGT}}^{\text{CR-AIS}}(\mathbf{z}_{0:T-1}^{(1:K)}, \mathbf{z}_T, \mathbf{x}) = \pi_T(\mathbf{z}_T, \mathbf{x}) \prod_{k=1}^{K} p_{\text{TGT}}^{\text{AIS}}(\mathbf{z}_{0:T-1}^{(k)}|\mathbf{z}_T, \mathbf{x}). \tag{18}$$

The extended state space proposal is obtained by selecting an index s uniformly at random and running a single forward AIS chain. We then run $K-1$ backward chains, all starting from the last state of the selected forward chain, as visualized in Fig. 2

$$q_{\text{PROP}}^{\text{CR-AIS}}(\mathbf{z}_{0:T-1}^{(1:K)}, \mathbf{z}_T|\mathbf{x}) = \frac{1}{K}\sum_{s=1}^{K} q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T-1}^{(s)}, \mathbf{z}_T|\mathbf{x}) \prod_{k=1,k\neq s}^{K} p_{\text{TGT}}^{\text{AIS}}(\mathbf{z}_{0:T-1}^{(k)}|\mathbf{z}_T, \mathbf{x}) \tag{19}$$

Taking the expected log ratio under the proposal and target yields lower and upper bounds on $\log p(\mathbf{x})$,

$$\underbrace{-\mathbb{E}_{\substack{\mathbf{z}_{0:T-1}^{(1)}, \mathbf{z}_T \sim q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}|\mathbf{x}) \\ \mathbf{z}_{0:T-1}^{(2:K)} \sim p_{\text{TGT}}^{\text{AIS}}(\mathbf{z}_{0:T-1}|\mathbf{z}_T, \mathbf{x})}} \left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{q_{\text{PROP}}^{\text{AIS}}(\cdot)}{p_{\text{TGT}}^{\text{AIS}}(\cdot)}\right]}_{\text{ELBO}_{\text{CR-AIS}}(\pi_0, K, T)} \leq \log p(\mathbf{x}) \leq \underbrace{-\mathbb{E}_{\substack{\mathbf{z}_T \sim \pi_T(\mathbf{z}_T|\mathbf{x}) \\ \mathbf{z}_{0:T-1}^{(1:K)} \sim p_{\text{TGT}}^{\text{AIS}}(\mathbf{z}_{0:T-1}|\mathbf{z}_T, \mathbf{x})}} \left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{q_{\text{PROP}}^{\text{AIS}}(\cdot)}{p_{\text{TGT}}^{\text{AIS}}(\cdot)}\right]}_{\text{EUBO}_{\text{CR-AIS}}(\pi_0, K, T)}$$

Proposition 3.3. *For a given T, the bias of the Coupled Reverse Multi-Sample* AIS *upper bound on* MI*, $I_{\text{CR-AIS}_U}(\pi_0, K, T)$, decreases (at most) logarithmically in K.*

$$I_{\text{AIS}_U}(\pi_0, T) - \log K \leq I_{\text{CR-AIS}_U}(\pi_0, K, T) \leq I_{\text{AIS}_U}(\pi_0, T). \tag{20}$$

3.4 Discussion

Relationship with BDMC While Bidirectional Monte Carlo (BDMC) (Grosse et al., 2015; 2016) was the first method proposing multi-sample log partition function bounds using AIS chains, our probabilistic interpretations provide novel perspective on BDMC. Perhaps surprisingly, we find that BDMC lower and upper bounds do not correspond to the same extended state space proposal and target distributions. In particular, the BDMC lower bound on $\log p(\mathbf{x})$ corresponds to the lower bound of Independent Multi-Sample AIS (Fig. 2 Col. 4, Row 4), while the upper bound of BDMC matches the upper bound of Coupled Reverse Multi-Sample AIS (Fig. 2 Col. 6, Row 3).

Effect of K and T We have shown in Prop. 3.2 that our multi-sample AIS bounds can achieve *linear* bias reduction with increasing T, although this computation must be done in serial fashion. While increasing K involves parallel computation, its bias reduction is often only *logarithmic* for both IWAE and Multi-Sample AIS, as we have shown in Prop. 2.1, 3.1, 3.3 and further argue in App. B.2-B.4. Based on these arguments, we recommend increasing K until computation can no longer be parallelized on given hardware and allocating all remaining resources to increasing T.

Comparing Multi-Sample AIS Bounds In App. M.1 Fig. 9 , we compare the performance of our various Multi-Sample AIS bounds in order to make recommendations for which bounds to use in practice. For the upper bound on MI, we recommend the Independent Multi-Sample AIS ELBO, which uses independent samples and is not limited to $\log K$ improvement compared with the Coupled Reverse Multi-Sample AIS ELBO (Prop. 3.3). The results are less conclusive for the lower bounds on MI. While the MI lower bound obtained from $\text{EUBO}_{\text{IM-AIS}}(\pi_0, K, T)$ (RHS of Eq. 16) can only improve upon single-sample AIS by $\log K$ (Prop. 3.1), this improvement is easily obtained for low T and may be useful for quickly estimating MI at a similar order of magnitude as $\log K$. The Coupled Reverse AIS lower bound on MI in requires at least moderate values of T to match or marginally improve upon Independent Multi-Sample AIS, suggesting that the preferred MI lower bounds may vary based on the scale of the true MI and the amount of computation available.

4 MINE-AIS Estimation of Mutual Information

In this section, we present MINE-AIS, which uses our multi-sample AIS bounds from the previous section to bound the mutual information in settings where the true conditional density $p(\mathbf{x}|\mathbf{z})$ is unknown. We consider a flexible energy-based distribution $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$ as an approximation to the posterior $p(\mathbf{z}|\mathbf{x})$ (Poole et al., 2019; Arbel et al., 2020)

$$\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x}) = \frac{1}{\mathcal{Z}_{\theta,\phi}(\mathbf{x})} q_\theta(\mathbf{z}|\mathbf{x}) e^{T_\phi(\mathbf{x},\mathbf{z})} \quad \text{where} \quad \mathcal{Z}_{\theta,\phi}(\mathbf{x}) = \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right]. \tag{21}$$

Plugging $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$ into the BA lower bound, we denote the resulting bound as the Implicit Barber-Agakov Lower Bound (IBAL), since it is often difficult to evaluate explicitly due to the intractable log partition function term. After simplifying in App. D.1, we obtain

$$I(\mathbf{x};\mathbf{z}) \geq \text{IBAL}(q_\theta, T_\phi) := I_{\text{BA}_L}(\pi_{\theta,\phi}) = \underbrace{\mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\theta(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right]}_{I_{\text{BA}_L}(q_\theta)} + \underbrace{\mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{e^{T_\phi(\mathbf{x},\mathbf{z})}}{\mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right]}\right]}_{\leq \mathbb{E}_{p(\mathbf{x})}[D_{KL}[p(\mathbf{z}|\mathbf{x})||q_\theta(\mathbf{z}|\mathbf{x})]]}, \tag{22}$$

with the gap of the IBAL equal to $\mathbb{E}_{p(\mathbf{x})}[D_{\text{KL}}[p(\mathbf{z}|\mathbf{x}))||\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})]]$. Note that the IBAL generalizes the Unnormalized Barber-Agakov bound from Poole et al. (2019), with $I_{\text{UBA}}(T_\phi) = \text{IBAL}(p(\mathbf{z}), T_\phi)$.

Relationship with MINE As discussed in Poole et al. (2019), MINE-DV and MINE-F (Belghazi et al., 2018) utilize *lower bounds* on $\text{IBAL}(p(\mathbf{z}), T_\phi)$. Instead, MINE-AIS will optimize Eq. 22 directly using an improved energy-based training scheme and evaluate the IBAL using AIS.

Relationship with GIWAE We can immediately notice similarities between GIWAE and MINE-AIS, including that their optimal energy functions match and that both bounds include a contrastive term which improves upon the BA lower bound. In fact, we prove the following proposition in App. E.2.

Proposition 4.1. *For given $q_\theta(\mathbf{z}|\mathbf{x})$ and $T_\phi(\mathbf{x},\mathbf{z})$,* $\lim_{K\to\infty} I_{\text{GIWAE}_L}(q_\theta, T_\phi, K) = \text{IBAL}(q_\theta, T_\phi)$.

Thus, we may view the IBAL as the limiting behavior of the finite-sample GIWAE bounds as $K \to \infty$. While Cor. 2.3 shows that $I_{\text{GIWAE}_L}(q_\theta, T)$ can improve $I_{\text{BA}_L}(q_\theta)$ by at most $\log K$ nats, we show in App. E.1 that the IBAL contrastive term is flexible enough to close the entire gap in the BA bound.

$$I_{\text{BA}_L}(q_\theta) \leq \text{IBAL}(q_\theta, T_{\phi^*}) \leq I(\mathbf{x};\mathbf{z}) = I_{\text{BA}_L}(q_\theta) + \mathbb{E}_{p(\mathbf{x})}\left[D_{KL}[p(\mathbf{z}|\mathbf{x})||q_\theta(\mathbf{z}|\mathbf{x})]\right] \tag{23}$$

However, this flexible contrastive term comes at the cost of tractability, as MINE-AIS involves an intractable partition function $\log \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}[e^{T_\phi(\mathbf{x},\mathbf{z})}]$ compared to the finite-sum term in GIWAE.

Training of MINE-AIS Although the log partition function $\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})$ in the IBAL is intractable to evaluate, we only require an unbiased estimator of its gradient for training. Differentiating Eq. 22 with respect to the parameters θ and ϕ of the variational and energy function, respectively, we obtain

$$\frac{\partial}{\partial\theta}\text{IBAL}(q_\theta, T_\phi) = \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\frac{\partial}{\partial\theta}\log q_\theta(\mathbf{z}|\mathbf{x})\right] - \mathbb{E}_{p(\mathbf{x})\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})}\left[\frac{\partial}{\partial\theta}\log q_\theta(\mathbf{z}|\mathbf{x})\right], \tag{24}$$

$$\frac{\partial}{\partial\phi}\text{IBAL}(q_\theta, T_\phi) = \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\frac{\partial}{\partial\phi}T_\phi(\mathbf{x},\mathbf{z})\right] - \mathbb{E}_{p(\mathbf{x})\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})}\left[\frac{\partial}{\partial\phi}T_\phi(\mathbf{x},\mathbf{z})\right]. \tag{25}$$

Table 1: MI Estimation with IWAE (with varying K), and Multi-Sample AIS (with varying T) on MNIST (left) and CIFAR (right). Tight sandwich bounds, with a gap of less than 2 nats, are in bold.

Method	Proposal	VAE2	VAE10	VAE100	GAN2	GAN10	GAN100
AIS T=1	$p(\mathbf{z})$	$(0, 223.93)$	$(0, 1900.27)$	$(0, 5936.13)$	$(0, 767.39)$	$(0, 851.20)$	$(0, 845.96)$
	$q(\mathbf{z}\|\mathbf{x})$	$(7.67, 9.15)$	$(21.09, 64.14)$	$(34.15, 366.82)$	$(7.22, 13.87)$	$(3.83, 340.29)$	$(2.57, 512.63)$
AIS T=500	$p(\mathbf{z})$	$(8.23, 17.11)$	$(33.52, 107.36)$	$(78.71, 493.92)$	$(8.23, 98.09)$	$(21.52, 167.76)$	$(24.45, 228.71)$
	$q(\mathbf{z}\|\mathbf{x})$	**$(8.97, 8.98)$**	$(34.18, 53.31)$	$(81.93, 254.35)$	**$(10.45, 11.13)$**	$(22.10, 94.37)$	$(25.06, 165.05)$
AIS T=30k	$p(\mathbf{z})$	**$(8.86, 8.97)$**	**$(34.28, 34.32)$**	**$(80.93, 81.03)$**	**$(10.35, 10.54)$**	**$(22.52, 22.59)$**	**$(26.43, 26.60)$**
	$q(\mathbf{z}\|\mathbf{x})$	**$(8.98, 8.98)$**	**$(34.29, 34.29)$**	**$(80.93, 80.97)$**	**$(10.56, 10.56)$**	**$(22.54, 22.59)$**	**$(26.49, 26.56)$**
IWAE K=1	$p(\mathbf{z})$	$(0, 800.25)$	$(0, 3968.43)$	$(0, 10919.71)$	$(0, 1678.36)$	$(0, 1688.52)$	$(0, 1602.63)$
	$q(\mathbf{z}\|\mathbf{x})$	$(7.64, 14.50)$	$(21.14, 147.77)$	$(34.56, 702.71)$	$(7.18, 110.65)$	$(3.82, 1098.28)$	$(2.54, 1346.11)$
IWAE K=1k	$p(\mathbf{z})$	$(6.77, 29.28)$	$(6.91, 1201.26)$	$(6.91, 4244.00)$	$(6.87, 106.53)$	$(6.91, 444.64)$	$(6.91, 483.34)$
	$q(\mathbf{z}\|\mathbf{x})$	**$(9.01, 9.01)$**	$(28.00, 46.02)$	$(41.47, 266.91)$	**$(10.49, 10.77)$**	$(10.73, 150.65)$	$(9.44, 291.38)$
IWAE K=1M	$p(\mathbf{z})$	**$(8.99, 9.00)$**	$(13.82, 379.34)$	$(13.82, 2259.10)$	**$(10.46, 10.61)$**	$(13.79, 84.88)$	$(13.82, 121.35)$
	$q(\mathbf{z}\|\mathbf{x})$	**$(9.00, 9.00)$**	**$(33.16, 34.88)$**	$(48.37, 170.37)$	**$(10.58, 10.58)$**	$(17.37, 31.99)$	$(16.29, 69.39)$

Method	Proposal	GAN100
AIS T=1	$p(\mathbf{z})$	$(0, 11827.18)$
	$q(\mathbf{z}\|\mathbf{x})$	$(7.55, 8751.58)$
AIS T=500	$p(\mathbf{z})$	$(80.28, 6104.59)$
	$q(\mathbf{z}\|\mathbf{x})$	$(83.346, 5316.53)$
AIS T=100k	$p(\mathbf{z})$	$(195.20, 485.57)$
	$q(\mathbf{z}\|\mathbf{x})$	$(174.66, 202.83)$
IWAE K=1	$p(\mathbf{z})$	$(0, 26510.82)$
	$q(\mathbf{z}\|\mathbf{x})$	$(11.50, 20962.96)$
IWAE K=1k	$p(\mathbf{z})$	$(6.91, 9309.36)$
	$q(\mathbf{z}\|\mathbf{x})$	$(18.40, 5042.09)$
IWAE K=20k	$p(\mathbf{z})$	$(9.90, 7698.18)$
	$q(\mathbf{z}\|\mathbf{x})$	$(21.40, 3951.44)$

We use MCMC with Hamiltonian Monte Carlo (HMC) transition kernels (Neal, 2011) to draw approximate samples from $\pi_{\theta,\phi}$, with additional discussion in App. E.4. Remarkably, using Prop. 4.1 and Eq. 24-25, we observe that direct gradient-based optimization of the contrastive term in finite-sample GIWAE bounds translates to energy-based training in the limit of $K \to \infty$. The goal in both cases is to raise the value of T_ϕ on positive examples and lower it on negative examples, although, in MINE-AIS, we use only a single, approximate sample from $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$.

Evaluation of IBAL in MINE-AIS Given an energy function learned using our MINE-AIS training scheme, we still need to evaluate the IBAL lower bound on MI. We have an unbiased, low variance estimate for $\mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[T_\phi(\mathbf{x},\mathbf{z})\right]$, and will use Multi-Sample AIS to estimate $\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})$ by annealing between $\pi_0(\mathbf{z}|\mathbf{x}) = q_\theta(\mathbf{z}|\mathbf{x})$ and $\pi_T(\mathbf{x},\mathbf{z}) = q_\theta(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})}$. To preserve a lower bound on $I(\mathbf{x};\mathbf{z})$, we would like an upper bound on $\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})$. However, this requires true samples from $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$ to initialize backward AIS chains. Since these samples are unavailable, we argue in App. E.5 that we can preserve an approximate upper bound on $\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})$ by initializing backward chains from the true posterior $p(\mathbf{z}|\mathbf{x})$ instead of $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$. Using a Multi-Sample AIS lower bound on $\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})$, we can upper bound the IBAL but cannot guarantee that this is a stochastic lower bound on $I(\mathbf{x};\mathbf{z})$.

5 Experiments

In this section, we evaluate our proposed Multi-Sample AIS, MINE-AIS, and GIWAE bounds for estimating MI in deep generative models such as VAEs and GANs trained on MNIST and CIFAR.

5.1 Multi-Sample AIS Estimation of Mutual Information

We compare Multi-Sample AIS MI estimation against IWAE, as both methods assume the full joint distribution is available. For the initial distribution of AIS or variational distribution of IWAE, we can use any distribution that is tractable to sample and evaluate. We experiment using both the prior $p(\mathbf{z})$ and a Gaussian $q(\mathbf{z}|\mathbf{x})$ trained to optimize the BA lower bound. Tables 1 summarize our results.

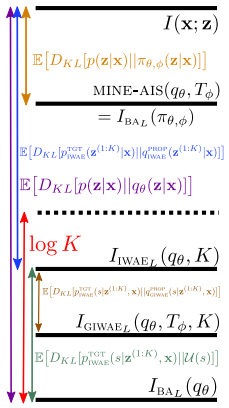
IWAE As described in Sec. 2.2, IWAE bounds encompass a wide range of MI estimators. The $K = 1$ bounds with learned $q_\theta(\mathbf{z}|\mathbf{x})$ correspond to BA bounds, while the upper bound with a prior $p(\mathbf{z})$ proposal matches the CLUB. For $K > 1$ and $p(\mathbf{z})$ as the proposal, we obtain Structured InfoNCE.

We see that the improvement of the IWAE lower bound is limited by $\log K$, and thus requires an exponentially large sample size to close the gap from the BA lower bound ($K = 1$) to the true MI. IWAE can successfully sandwich the MI of VAE2, VAE10 and GAN2 on MNIST in cases where the gap is less than $\log K$. However, using Gaussian $q_\theta(\mathbf{z}|\mathbf{x})$, tight estimation with IWAE is clearly infeasible on VAE100 and GAN100. On VAE100, for example, at least e^{46} total samples are required to improve $I_{\text{BA}_L}(q_\theta) = I_{\text{IWAE}_L}(q_\theta, 1) = 34.56$ to match the ground truth $I(\mathbf{x};\mathbf{z}) \approx 80.95$ estimated by AIS.

Multi-Sample AIS We evaluate our Multi-Sample AIS bounds with $K = 48$ chains on MNIST and $K = 12$ on CIFAR, and a varying number of intermediate T. We show results for the Independent Multi-Sample AIS MI lower bound and Coupled Reverse Multi-Sample AIS upper bound in Table 1. Using large enough values of T, Multi-Sample AIS can tightly sandwich large values of ground truth MI in our MNIST-VAE models. Our bounds are more loose in the challenging CIFAR-GAN setting, but these results highlight that increasing T in Multi-Sample AIS is a practical way to reduce bias using additional computation. This is in stark contrast to the exponential sample complexity required for IWAE, as we saw in the previous discussion. We provide runtime details in App. L.3.

Input Used	Bound \ Model	Linear VAE10	MNIST VAE20	MNIST GAN20
$p(\mathbf{x}, \mathbf{z})$	Analytical	23.22	N/A	N/A
$p(\mathbf{x}, \mathbf{z})$ Joint Samples	Multi-Sample AIS Sandwich Bound	(23.21, 23.25)	(36.51, 36.66)	(53.69, 53.73)
	IWAE LB ($K = 1000$)	$20.5 + 2.6 = 23.2$	$22.6 + 6.6 = 29.2$	$20.4 + 6.9 = 27.3$
	IWAE LB ($K = 100$)	$21.4 + 1.7 = 23.1$	$24.2 + 4.6 = 28.8$	$20.9 + 4.6 = 25.5$
	Structured InfoNCE LB ($K = 1000$)	6.90	6.90	6.90
	Structured InfoNCE LB ($K = 100$)	4.60	4.60	4.60
$p(\mathbf{z})$ Joint Samples	MINE-AIS Sandwich Bound on IBAL	(23.04, 23.05)	(30.05, 30.05)	(38.61, 38.61)
	Generalized IWAE LB ($K = 100$)	$22.7 + 0.1 = 22.8$	$23.1 + 3.9 = 27.0$	$20.3 + 4.6 = 24.9$
	Barber-Agakov LB	22.76	24.48	20.98
Joint Samples	InfoNCE LB ($K = 1000$)	6.90	6.90	6.90
	InfoNCE LB ($K = 100$)	4.60	4.60	4.60

(a)



(b)

Figure 3: Comparing energy-based bounds (GIWAE and MINE-AIS) with other MI bounds.

5.2 Energy-Based Estimation of Mutual Information

In this section, we evaluate the family of GIWAE and MINE-AIS bounds, which assume access to a known marginal $p(\mathbf{z})$ but not the conditional $p(\mathbf{x}|\mathbf{z})$. We summarize these bounds in Fig. 3b.

BA, IWAE and GIWAE Bounds Recall that $I_{\mathrm{GIWAE}_L}(q, T_\phi, K)$ in Eq. 9 consists of the BA lower bound and a contrastive term. In Fig. 3a, we report the contribution of each term for different model complexity and special cases of GIWAE. We find that the results respect the inequalities shown in Cor. C.2, with $I_{\mathrm{BA}_L}(q_\theta) \leq I_{\mathrm{GIWAE}_L}(q_\theta, T_{\phi^*}, K) \leq I_{\mathrm{IWAE}_L}(q_\theta, K) \leq I_{\mathrm{BA}_L}(q_\theta) + \log K$. [2]

For the Linear VAE, the true posterior is in the Gaussian variational family and $I_{\mathrm{BA}_L}(q_\theta)$ almost exactly matches the analytical MI. In this case, the contrastive term provides much less than $\log K$ improvement for GIWAE and IWAE, since even the optimal energy function cannot distinguish $q_\theta(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z}|\mathbf{x})$. The contrastive term increases for more complex MNIST-VAE posteriors, while the fact that BA term in IWAE is lower for $K = 1000$ than for $K = 100$ indicates a tradeoff between the two terms. A worse q_θ allows the contrastive term to achieve closer to the full $\log K$ improvement, and yields a higher overall bound. This observation is similar to Rainforth et al. (2018), where increasing K in $\mathrm{ELBO}_{\mathrm{IWAE}}(q_\theta, K)$ reduces the quality of the posterior approximation in variational inference. Finally, there is essentially no tradeoff between the terms for MNIST-GAN, since the variational family is far enough from the true posterior that both learned GIWAE and optimal IWAE energy functions can achieve $\log K$ improvement over BA.

MINE-AIS Bounds We can see in Fig. 3a that MINE-AIS improves over BA due to its more flexible variational family, and improves upon GIWAE, IWAE, and (Structured) INFONCE since it is their limiting behavior. [3] In Fig. 3a, we can compare AIS estimates of MI and the IBAL for MINE-AIS to indicate the quality of the learned energy function. We observe that MINE-AIS lower bounds underestimate the value obtained from Multi-Sample AIS by $1\%, 18\%$, and 28% on linear VAE, MNIST-VAE, and MNIST-GAN, respectively.

6 Conclusion

In this work, we have provided a unifying view of mutual information estimation from the perspective of importance sampling. We derived probabilistic interpretations of each bound, which shed light on the limitations of existing estimators, and motivate our novel GIWAE, Multi-Sample AIS, and MINE-AIS bounds. When the conditional is not known, our GIWAE bounds highlights how variational bounds can complement contrastive learning to improve lower bounds on MI beyond the known $\log K$ limitations. When the full joint distribution is known, our Multi-Sample AIS bounds are the only practical bounds which have been demonstrated to estimate large values of MI without exponential sample complexity, and thus should be considered the gold standard for MI estimation in these settings. Finally, MINE-AIS extends Multi-Sample AIS evaluation to unknown conditional densities, and can be viewed as the infinite-sample behavior of GIWAE and existing contrastive bounds. Our MINE-AIS and Multi-Sample AIS methods highlight how MCMC techniques can be used to improve mutual information estimation when a single analytic marginal or conditional density is available.

[2] Note that the BA term in GIWAE and IWAE may not match exactly, since the overall training objectives are different. For example, GIWAE jointly learns q_θ and T_ϕ rather than only optimizing q_θ to maximize $I_{\mathrm{BA}_L}(q_\theta)$.

[3] Note that our MINE-AIS results use $p(\mathbf{z})$ as the base distribution, which is the limiting behavior of INFONCE. We might expect further improvement by jointly learning a variational base distribution $q_\theta(\mathbf{z}|\mathbf{x})$ and energy T_ϕ.

REFERENCES

Felix Agakov. *Variational information maximization in stochastic environments*. PhD thesis, University of Edinburgh, 2006.

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, 2018.

Alexander A Alemi and Ian Fischer. Gilbo: one metric to measure them all. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 7037–7046, 2018.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Shun-Ichi Amari. Alpha-divergence is unique, belonging to both f-divergence and bregman divergence classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931, 2009.

Michael Arbel, Liang Zhou, and Arthur Gretton. Generalized energy based models. *arXiv e-prints*, pp. arXiv–2003, 2020.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in Neural Information Processing Systems*, 32:15535–15545, 2019.

David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pp. 201–208, 2003.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pp. 531–540. PMLR, 2018.

Yuri Burda, Roger B Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2016.

Sourav Chatterjee, Persi Diaconis, et al. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 2180–2188, 2016.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning*, pp. 1779–1788. PMLR, 2020.

Andrzej Cichocki and Shun-ichi Amari. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Chris Cremer, Quaid Morris, and David Duvenaud. Reinterpreting importance-weighted autoencoders. *arXiv preprint arXiv:1704.02916*, 2017.

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

Tomas Geffner and Justin Domke. Mcmc variational inference via uncorrected hamiltonian annealing. *Advances in Neural Information Processing Systems*, 2021.

Roger B Grosse, Chris J Maddison, and Ruslan Salakhutdinov. Annealing between distributions by averaging moments. In *NIPS*, pp. 2769–2777. Citeseer, 2013.

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

Roger B Grosse, Siddharth Ancha, and Daniel M Roy. Measuring the reliability of mcmc inference with bidirectional monte carlo. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 2459–2467, 2016.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein GANs. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

Umang Gupta, Aaron Ferber, Bistra Dilkina, and Greg Ver Steeg. Controllable guarantees for fair outcomes via contrastive information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7610–7619, 2021.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

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

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

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

Dieterich Lawson, George Tucker, Bo Dai, and Rajesh Ranganath. Energy-inspired models: learning with sampler-induced distributions. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 8501–8513, 2019.

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

Chris J Maddison, Dieterich Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Whye Teh. Filtering variational objectives. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6576–6586, 2017.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Vaden Masrani, Tuan Anh Le, and Frank Wood. The thermodynamic variational objective. *Advances in Neural Information Processing Systems*, 32:11525–11534, 2019.

Bogdan Mazoure, Remi Tachet des Combes, Thang Long DOAN, Philip Bachman, and R Devon Hjelm. Deep reinforcement and infomax learning. *Advances in Neural Information Processing Systems*, 33, 2020.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pp. 875–884. PMLR, 2020.

Tom Minka et al. Divergence measures and message passing. Technical report, Citeseer, 2005.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. *Advances in Neural Information Processing Systems*, 28:2125–2133, 2015.

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

Radford M Neal. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2011.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, 2019.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Tom Rainforth, Adam Kosiorek, Tuan Anh Le, Chris Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. In *International Conference on Machine Learning*, pp. 4277–4285. PMLR, 2018.

Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pp. 1218–1226. PMLR, 2015.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Artem Sobolev. Thoughts on mutual information estimation: More estimators. *Blog post*, 2019a. URL http://artem.sobolev.name/posts/2019-08-10-thoughts-on-mutual-information-more-estimators.html.

Artem Sobolev. Thoughts on mutual information estimation: Formal limitations. *Blog post*, 2019b. URL http://artem.sobolev.name/posts/2019-08-14-thoughts-on-mutual-information-formal-limitations.html.

Artem Sobolev and Dmitry P Vetrov. Importance weighted hierarchical variational inference. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2019.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Yuhuai Wu, Yuri Burda, Ruslan Salakhutdinov, and Roger Grosse. On the quantitative analysis of decoder-based generative models. *arXiv preprint arXiv:1611.04273*, 2016.

Guodong Zhang, Kyle Hsu, Jianing Li, Chelsea Finn, and Roger Grosse. Differentiable annealed importance sampling and the perils of gradient noise. *Advances in Neural Information Processing Systems*, 2021.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. The information autoencoding family: A lagrangian perspective on latent variable generative models. In *Proc. 34th Conference on Uncertainty in Artificial Intelligence*, 2018.

Supplementary Material

Table of Contents

A A General Approach for Deriving Extended State Space Bounds on Log Partition Functions **14**

B IWAE **15**
- B.1 Probabilistic Interpretation and Bounds . . . 15
- B.2 Proof of Proposition 2.1 . . . 16
- B.3 Logarithmic Improvement Results for $I_{\text{IWAE}_L}(q_\theta, K)$ . . . 17
- B.4 Bias Reduction of IWAE Lower Bound on $\log p(\mathbf{x})$ / Upper Bound on MI . . . 18
- B.5 Relationship with Structured InfoNCE . . . 18

C Generalized IWAE **18**
- C.1 Probabilistic Interpretation and Bounds . . . 18
- C.2 GIWAE Upper Bound on MI Does Not Provide Benefit Over IWAE . . . 20
- C.3 Relationship between IWAE and GIWAE Probabilistic Interpretations . . . 20
- C.4 Proof of Proposition 2.2 and Corollary 2.3 . . . 21

D Generalized Mutual Information Neural Estimation (GMINE) **22**
- D.1 Probabilistic Interpretation of IBAL . . . 22
- D.2 Probabilistic Interpretation of Generalized MINE-DV . . . 22
- D.3 Probabilistic Interpretation of Generalized MINE-F . . . 24

E MINE-AIS **25**
- E.1 Properties of MINE-AIS . . . 25
- E.2 Proof of Proposition 4.1 . . . 26
- E.3 Convergence of GIWAE SNIS Distribution to MINE-AIS . . . 28
- E.4 Details for Training MINE-AIS . . . 29
- E.5 Approximate Reverse Annealing for Evaluating MINE-AIS . . . 29
- E.6 Multi-Sample AIS Convergence Results for MINE-AIS . . . 30

F Applications to Mutual Information Estimation without Known Marginals **31**
- F.1 BA Lower Bound . . . 31
- F.2 GIWAE Lower Bound . . . 32
- F.3 MINE-AIS / IBAL Lower Bound . . . 33

G Background on AIS **33**

H Independent Multi-Sample AIS **34**
- H.1 Probabilistic Interpretation and Bounds . . . 34
- H.2 Proof of Proposition 3.1 . . . 35
- H.3 Proof of Proposition 3.2 . . . 35

I Reverse IWAE **37**
- I.1 Probabilistic Interpretation and Bounds . . . 37
- I.2 Proof of Logarithmic Improvement of RIWAE . . . 37

J Independent Reverse Multi-Sample AIS **38**

K Coupled Reverse Multi-Sample AIS **39**
- K.1 Probabilistic Interpretation and Bounds . . . 39
- K.2 Proof of Proposition 3.3 . . . 39

L Experimental Details **40**
- L.1 Datasets and Models . . . 40
- L.2 Multi-Sample AIS Settings . . . 40

L.3 Runtime Comparison 40
L.4 Initial Distribution of AIS 40
L.5 Analytical Solution of the Mutual Information on the Linear VAE 41

M Additional Experiments **42**
M.1 Comparison of Multi-Sample AIS Bounds 42
M.2 Confidence Intervals for Multi-Sample AIS Experiments 42

A A General Approach for Deriving Extended State Space Bounds on Log Partition Functions

We begin by describing a general methodology for deriving lower and upper bounds on the log partition function $\log p(\mathbf{x})$ using importance sampling in an extended state space. We employ this approach throughout this Appendix, where the state space may be extended using multiple samples ($K > 1$) as in IWAE or approximate samples from intermediate densities ($T > 1$) as in AIS.

Suppose we are given a tractable initial distribution $\pi_0(\mathbf{z})$, which may correspond to a variational distribution $q_\theta(\mathbf{z}|\mathbf{x})$ or prior distribution $p(\mathbf{z})$, and are interested in estimating the normalization constant $p(\mathbf{x})$ for an intractable posterior distribution $p(\mathbf{z}|\mathbf{x}) \propto p(\mathbf{x}, \mathbf{z})$. In Fig. 4, we summarize various extended state space proposal (second column of Fig. 4) and target distributions (third column of Fig. 4) constructed from these initial and posterior distributions, with q_{PROP} and p_{TGT} chosen such that $\frac{\mathcal{Z}_{\text{TGT}}}{\mathcal{Z}_{\text{PROP}}} = \log p(\mathbf{x})$. Each pair of extended state-space proposal and target distributions then provides both upper and lower bounds on the log partition function using importance sampling, by taking the expected log ratio of unnormalized densities under either the proposal or target distribution

$$\mathbb{E}_{q_{\text{PROP}}}\left[\log \frac{\tilde{p}_{\text{TGT}}}{\tilde{q}_{\text{PROP}}}\right] \leq \log \frac{\mathcal{Z}_{\text{TGT}}}{\mathcal{Z}_{\text{PROP}}} \leq \mathbb{E}_{p_{\text{TGT}}}\left[\log \frac{\tilde{p}_{\text{TGT}}}{\tilde{q}_{\text{PROP}}}\right] \tag{26}$$

Note that we use $\tilde{q}, \tilde{p}$ to indicate unnormalized densities over $\mathbf{x}$ and a possibly extended (latent) state space. For example, $p_{\text{TGT}}(\mathbf{z}_{0:T}^{(1:K)}|\mathbf{x}) = \frac{1}{\mathcal{Z}_{\text{TGT}}}\tilde{p}_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{0:T}^{(1:K)})$ represents a state space extended by multiple proposal samples $K \geq 1$ as in IWAE and multiple intermediate samples $T \geq 1$ as in AIS.

For all q_{PROP} and p_{TGT}, the gap in the lower bound in Eq. 26 is the forward KL divergence, $D_{KL}(q_{\text{PROP}}||p_{\text{TGT}})$, while the gap in the upper bound is the reverse KL divergence, $D_{KL}(p_{\text{TGT}})||q_{\text{PROP}})$.

$$\mathbb{E}_{q_{\text{PROP}}}\left[\frac{\tilde{p}_{\text{TGT}}}{\tilde{q}_{\text{PROP}}}\right] = \mathbb{E}_{q_{\text{PROP}}}\left[\frac{p_{\text{TGT}}}{q_{\text{PROP}}}\right] + \log \frac{\mathcal{Z}_{\text{TGT}}}{\mathcal{Z}_{\text{PROP}}} = \underbrace{\log \frac{\mathcal{Z}_{\text{TGT}}}{\mathcal{Z}_{\text{PROP}}} - D_{KL}[q_{\text{PROP}}||p_{\text{TGT}}]}_{\text{ELBO}(q_{\text{PROP}}, p_{\text{TGT}})} \leq \log \frac{\mathcal{Z}_{\text{TGT}}}{\mathcal{Z}_{\text{PROP}}} \tag{27}$$

$$\mathbb{E}_{p_{\text{TGT}}}\left[\frac{\tilde{p}_{\text{TGT}}}{\tilde{q}_{\text{PROP}}}\right] = \mathbb{E}_{p_{\text{TGT}}}\left[\frac{p_{\text{TGT}}}{q_{\text{PROP}}}\right] + \log \frac{\mathcal{Z}_{\text{TGT}}}{\mathcal{Z}_{\text{PROP}}} = \underbrace{\log \frac{\mathcal{Z}_{\text{TGT}}}{\mathcal{Z}_{\text{PROP}}} + D_{KL}[p_{\text{TGT}}||q_{\text{PROP}}]}_{\text{EUBO}(q_{\text{PROP}}, p_{\text{TGT}})} \geq \log \frac{\mathcal{Z}_{\text{TGT}}}{\mathcal{Z}_{\text{PROP}}} \tag{28}$$

Thus, the bounds in Eq. 26 directly generalize the standard ELBO ($\log p(\mathbf{x}) - D_{KL}[q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})]$) and EUBO ($\log p(\mathbf{x}) + D_{KL}[p(\mathbf{z}|\mathbf{x})||q(\mathbf{z}|\mathbf{x})]$), which appear as special cases with $K = 1$, $T = 1$, $q(\mathbf{z}|\mathbf{x})$ as the proposal distribution, and $p(\mathbf{z}|\mathbf{x})$ as the target distribution. We can alternatively derive the bounds in Eq. 26 by expanding the expression for the $D_{KL}[q_{\text{PROP}}||p_{\text{TGT}}]$ in the extended state space to obtain $\log \mathcal{Z}_{\text{TGT}}/\mathcal{Z}_{\text{PROP}} - \text{ELBO}(q_{\text{PROP}}, p_{\text{TGT}})$ as in Eq. 27.

In Fig. 4, we highlight that the initial or variational distribution (blue circles) and posterior distribution (red circles) distributions may be used multiple times, in either or both of the extended state space proposal and target distributions. Similarly, forward AIS chains (blue circles) starting from the initial distribution and the backward AIS chains (shown in red circles) starting from the target distribution may be used repeatedly in both the proposal or target.

In the next sections, we proceed to derive each of the bounds in Fig. 4 along with probabilistic interpretations of the underlying generative processes.

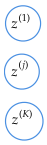
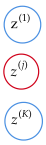
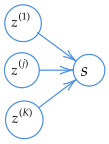
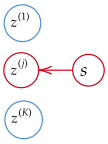
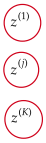
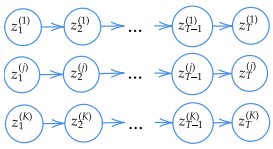
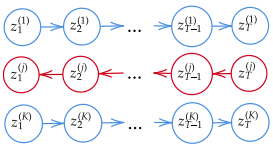
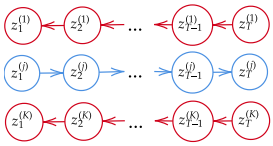
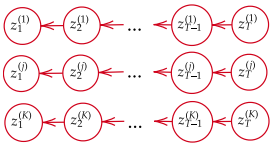
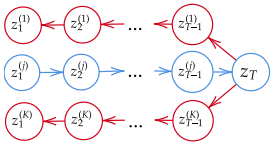
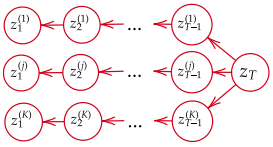
	Practicality	**Extended State-Space Proposal** Lower Bound on $\log p(x)$ (ELBO) Upper Bound on $I(x;z)$	**Extended State-Space Target** Upper Bound on $\log p(x)$ (EUBO) Lower Bound on $I(x;z)$
Simple IS	✓		
IWAE	✓		
Generalized IWAE	✓		
Reverse IWAE	✗		
AIS	✓		
Independent Multi-Sample AIS	✓		
Independent Reverse Multi-Sample AIS	✗		
Coupled Reverse Multi-Sample AIS	✓		

Figure 4: Comparison of the probabilistic extended state-space interpretations of different multi-sample bounds. Forward chains of AIS and variational distributions in IS/IWAE are colored in blue. Backward chains in AIS or posterior distributions in IS/IWAE are colored in red. Note that a lower bound on $\log p(\mathbf{x})$, translates to an upper bound on MI, and vice versa.

B IWAE

B.1 Probabilistic Interpretation and Bounds

Consider a K-sample proposal distribution $q^{\text{IWAE}}_{\text{PROP}}(\mathbf{z}^{(1:K)})$ using independent draws from an initial distribution $q_\theta(\mathbf{z}|\mathbf{x})$. The target distribution is defined as a uniform mixture of K components, where each component replaces the initial sample $q_\theta(\mathbf{z}^{(k)}|\mathbf{x})$ in index k with a sample from the target

distribution $p(\mathbf{z}|\mathbf{x})$.

$$q_{\text{PROP}}^{\text{IWAE}}(\mathbf{z}^{(1:K)}|\mathbf{x}) = \prod_{k=1}^{K} q_\theta(\mathbf{z}^{(k)}|\mathbf{x}), \qquad p_{\text{TGT}}^{\text{IWAE}}(\mathbf{x}, \mathbf{z}^{(1:K)}) = \frac{1}{K}\sum_{k=1}^{K} p(\mathbf{x}, \mathbf{z}^{(s)}) \prod_{k=1, k\neq s}^{K} q_\theta(\mathbf{z}^{(s)}|\mathbf{x}) \,. \tag{29}$$

Note that the normalizing constant of $p_{\text{TGT}}^{\text{IWAE}}(\mathbf{x}, \mathbf{z}^{(1:K)})$, or $\int p_{\text{TGT}}^{\text{IWAE}}(\mathbf{x}, \mathbf{z}^{(1:K)}) d\mathbf{z}^{(1:K)}$, equals $p(\mathbf{x})$ since $q_\theta(\mathbf{z}|\mathbf{x})$ is normalized and $\int p(\mathbf{x}, \mathbf{z}) d\mathbf{z} = p(\mathbf{x})$. To connect this with the IWAE bound, we show that the log importance ratio reduces to

$$\begin{aligned}
\log \frac{p_{\text{TGT}}^{\text{IWAE}}(\mathbf{x}, \mathbf{z}^{(1:K)})}{q_{\text{PROP}}^{\text{IWAE}}(\mathbf{x}, \mathbf{z}^{(1:K)})} &= \log \frac{\frac{1}{K}\sum_{k=1}^{K} p(\mathbf{x}, \mathbf{z}^{(s)}) \prod_{\substack{j=1 \\ j\neq k}}^{K} q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}{\prod_{k=1}^{K} q_\theta(\mathbf{z}^{(k)}|\mathbf{x})} \\
&= \log \left(\frac{1}{K}\sum_{k=1}^{K} p(\mathbf{x}, \mathbf{z}^{(k)}) \prod_{\substack{j=1 \\ j\neq k}}^{K} q_\theta(\mathbf{z}^{(k)}|\mathbf{x}) \cdot \prod_{k=1}^{k} q_\theta(\mathbf{z}^{(k)}|\mathbf{x})^{-1} \right) \\
&= \log \frac{1}{K}\sum_{k=1}^{K} \frac{p(\mathbf{x}, \mathbf{z}^{(k)})}{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}
\end{aligned} \tag{30}$$

As in Eq. (26), we can obtain lower and upper bounds on $\log p(\mathbf{x})$ by taking expectations under the proposal and target distributions, respectively.

Lower Bound on $\log p(\mathbf{x})$ and Upper Bound on MI First, taking expectations under $q_{\text{PROP}}^{\text{IWAE}}$, we obtain a lower bound on $\log p(\mathbf{x})$ as show in App. A

$$\text{ELBO}_{\text{IWAE}}(q_\theta, K) := \mathbb{E}_{q_\theta(\mathbf{z}^{(1:K)})}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{p(\mathbf{x}, \mathbf{z}^{(k)})}{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}\right] \tag{31}$$

This corresponds to the following upper bound on MI for known $p(\mathbf{x}|\mathbf{z})$

$$I(\mathbf{x}; \mathbf{z}) \leq I_{\text{IWAE}_U}(q_\theta, K) = -\mathcal{H}(\mathbf{x}|\mathbf{z}) - \text{ELBO}_{\text{IWAE}}(q_\theta, K) \tag{32}$$

$$= \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log p(\mathbf{x}|\mathbf{z})\right] - \mathbb{E}_{q_\theta(\mathbf{z}^{(1:K)}|\mathbf{x})}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{p(\mathbf{x}, \mathbf{z}^{(k)})}{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}\right] \tag{33}$$

Upper Bound on $\log p(\mathbf{x})$ and Lower Bound on MI Similarly, with expectations under $p_{\text{TGT}}^{\text{IWAE}}(\mathbf{z}^{(1:K)}|\mathbf{x})$, we obtain an upper bound on $\log p(\mathbf{x})$. Since, for independent draws, the uniform mixture $p_{\text{TGT}}^{\text{IWAE}}$ is invariant to index permutations, we may choose the target sample to be $\mathbf{z}^{(1)}$ and obtain the upper bound of Sobolev & Vetrov (2019):

$$\text{EUBO}_{\text{IWAE}}(q_\theta, K) := \mathbb{E}_{p(\mathbf{z}^{(1)}|\mathbf{x})q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \frac{1}{K}\sum_{i=1}^{K} \frac{p(\mathbf{x}, \mathbf{z}^{(k)})}{q(\mathbf{z}^{(k)}|\mathbf{x})}\right] \leq \log p(\mathbf{x}) \tag{34}$$

Translating this to a lower bound on MI with known $p(\mathbf{x}|\mathbf{z})$,

$$I(\mathbf{x}; \mathbf{z}) \geq I_{\text{IWAE}_L}(q_\theta, K) = -\mathcal{H}(\mathbf{x}|\mathbf{z}) - \text{EUBO}_{\text{IWAE}}(q_\theta, K) \tag{35}$$

$$= \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log p(\mathbf{x}|\mathbf{z})\right] - \mathbb{E}_{p(\mathbf{z}^{(1)}|\mathbf{x})q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{p(\mathbf{x}, \mathbf{z}^{(k)})}{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}\right] \tag{36}$$

B.2 Proof of Proposition 2.1

We first recall results that IWAE bounds tighten with increasing K.

Proposition B.1. *For given* $q_\theta(\mathbf{z}|\mathbf{x})$, $\text{ELBO}_{\text{IWAE}}(q_\theta, K+1) \geq \text{ELBO}_{\text{IWAE}}(q_\theta, K)$ *(Burda et al., 2016) and* $\text{EUBO}_{\text{IWAE}}(q_\theta, K+1) \leq \text{EUBO}_{\text{IWAE}}(q_\theta, K)$ *(Sobolev & Vetrov, 2019).*

Proposition B.2. *The bias of the* IWAE *upper bound on* $\log p(\mathbf{x})$ *decreases logarithmically in* K*. In particular, the* K*-sample upper bound* $\text{EUBO}_{\text{IWAE}}(q_\theta, K)$ *can improve upon the single-sample* EUBO $:= \log p(\mathbf{x}) + D_{KL}[p(\mathbf{z}|\mathbf{x})||q_\theta(\mathbf{z}|\mathbf{x})]$ *by at most* $\log K$ *nats.*

$$\text{EUBO}(q_\theta) - \log K \leq \text{EUBO}_{\text{IWAE}}(q_\theta, K) \leq \text{EUBO}(q_\theta) \tag{37}$$

Proof. We include a simple proof of Prop. B.2 here, which provides intuition for the role of the 'negative' samples $q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})$ and forms the basis for similar results later in the paper.

$$\begin{aligned}\log p(\mathbf{x}) \leq \mathbb{E}_{p(\mathbf{z}^{(1)}|\mathbf{x})q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \frac{1}{K}\sum_{k=1}^{K}\frac{p(\mathbf{x},\mathbf{z}^{(k)})}{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}\right] &= \mathbb{E}_{p(\mathbf{z}^{(1)}|\mathbf{x})q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \sum_{k=1}^{K}\frac{p(\mathbf{x},\mathbf{z}^{(k)})}{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}\right] - \log K \\ &\geq \mathbb{E}_{p(\mathbf{z}^{(1)}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{z}^{(1)})}{q_\theta(\mathbf{z}^{(1)}|\mathbf{x})}\right] - \log K \end{aligned} \tag{38}$$

where, in the second line, we have ignored the contribution of the 'negative' samples in the summation since the likelihood ratio is non-negative. Using the 'positive' sample from the posterior $p(\mathbf{z}|\mathbf{x})$, the first term is the EUBO and is still an upper bound on $\log p(\mathbf{x})$. However, note that EUBO $- \log K$ is no longer guaranteed to be an upper bound on $\log p(\mathbf{x})$. Prop. B.1 implies that $\text{EUBO}_{\text{IWAE}}(q_\theta, K) \leq \text{EUBO}(q_\theta)$. □

Prop. 2.1 follows directly by translating these $\log p(\mathbf{x})$ results to bounds on MI.

Proposition 2.1. IWAE *bounds on* MI *improve upon the* BA *bounds with the following relationships:*

$$I_{BA_L}(q_\theta) \leq I_{IWAE_L}(q_\theta, K) \leq I_{BA_L}(q_\theta) + \log K, \qquad I_{IWAE_U}(q_\theta, K) \leq I_{BA_U}(q_\theta). \tag{7}$$

Proof. Recall from Sec. 1.1 and Sec. 2.1 that $I_{\text{BA}_L}(q_\theta)) = -\text{EUBO}(q_\theta) - \mathcal{H}(\mathbf{x}|\mathbf{z})$, and $I_{\text{IWAE}_L}(q_\theta, K) = -\text{EUBO}_{\text{IWAE}}(q_\theta, K) - \mathcal{H}(\mathbf{x}|\mathbf{z})$. Using Prop. B.2, $\text{EUBO}(q_\theta, K) - \log K \leq \text{EUBO}_{\text{IWAE}}(q_\theta, K)$ implies that

$$I_{\text{IWAE}_L}(q_\theta, K) - I_{\text{BA}_L}(q_\theta) = \text{EUBO}(q_\theta) - \text{EUBO}_{\text{IWAE}}(q_\theta, K) \leq \log K \tag{39}$$

so that $I_{\text{IWAE}_L}(q_\theta, K) \leq I_{\text{BA}_L}(q_\theta)) + \log K$ as desired.

$I_{\text{BA}_L}(q_\theta) \leq I_{\text{IWAE}_L}(q_\theta, K)$ and $I_{\text{IWAE}_U}(q_\theta, K) \leq I_{\text{BA}_U}(q_\theta)$ follow from the fact that IWAE bounds tighten with increasing K in Prop. B.1. □

B.3 Logarithmic Improvement Results for $I_{\text{IWAE}_L}(q_\theta, K)$

In Sec. 2.3, we showed that IWAE is a special case of GIWAE, which decomposes into the sum of a variational BA lower bound and a K-sample contrastive term. This suggests that the IWAE lower bound on MI, which arises from an upper bound on $\log p(\mathbf{x})$, may be written as

$$I_{\text{IWAE}_L}(q_\theta, K) = \underbrace{\mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\theta(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right]}_{I_{\text{BA}}(q)} + \underbrace{\mathbb{E}_{p(\mathbf{x})p(\mathbf{z}^{(1)}|\mathbf{x})q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \frac{\frac{p(\mathbf{x},\mathbf{z}^{(1)})}{q_\theta(\mathbf{z}^{(1)}|\mathbf{x})}}{\frac{1}{K}\sum_{i=1}^{K}\frac{p(\mathbf{x},\mathbf{z}^{(k)})}{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}}\right]}_{0 \leq \text{ contrastive term } \leq \log K} \tag{40}$$

This way of writing $I_{\text{IWAE}_L}(q_\theta, K)$ provides intuition for the result in Prop. 2.1, since the improvement over $I_{\text{BA}_L}(q_\theta)$ is simply the contrastive term, which is limited to $\log K$.

Table 2 and Table 3 show the IWAE objective decomposition to the BA term and the contrastive term, on VAEs and GANs trained MNIST and CIFAR10 dataset. We can see that in all the experiments, the contribution of the contrastive term is always less than $\log K$. For low dimensional VAEs and GANs, the contrastive term contributes less than $\log K$, since the posterior is not complicated and the BA term can get close to the ground truth MI. However, for higher dimensional VAE and GAN models, the contribution of the contrastive term is almost exactly $\log K$. This sheds light on the inherent exponential sample complexity of IWAE, and motives our Multi-Sample AIS methods for MI estimation.

Method	$\log K$	Proposal	VAE2	VAE10	VAE100	GAN2	GAN10	GAN100
IWAE (K=1)	0	$p(\mathbf{z})$	$0+0=0$	$0+0=0$	$0+0=0$	$0+0=0$	$0+0=0$	$0+0=0$
		$q(\mathbf{z}\|\mathbf{x})$	$7.64+0=7.64$	$21.14+0=21.14$	$34.56+0=34.56$	$7.18+0=7.18$	$3.82+0=3.82$	$2.54+0=2.54$
IWAE (K=1K)	6.91	$p(\mathbf{z})$	$0+6.77=6.77$	$0+6.91=6.91$	$0+6.91=6.91$	$0+6.87=6.87$	$0+6.91=6.91$	$0+6.91=6.91$
		$q(\mathbf{z}\|\mathbf{x})$	$7.64+1.37=9.01$	$21.14+6.86=28.00$	$34.56+6.91=41.47$	$7.18+3.31=10.49$	$3.82+6.91=10.73$	$2.54+6.90=9.44$
IWAE (K=1M)	13.81	$p(\mathbf{z})$	$0+8.99=8.99$	$0+13.82=13.82$	$0+13.82=13.82$	$0+10.46=10.46$	$0+13.79=13.79$	$0+13.82=13.82$
		$q(\mathbf{z}\|\mathbf{x})$	$7.64+1.36=9.00$	$21.14+12.02=33.16$	$34.56+13.81=48.37$	$7.18+3.4=10.58$	$3.82+13.55=17.37$	$2.54+13.75=16.29$

Table 2: MNIST-VAE/GAN: Decomposition of IWAE LBs on MI to BA term and contrastive term ($< \log K$).

Method	$\log K$	Proposal	GAN100
IWAE K=1	0	$p(\mathbf{z})$	$0+0=0$
		$q(\mathbf{z}\|\mathbf{x})$	$11.50+0=11.50$
IWAE K=1k	6.91	$p(\mathbf{z})$	$0+6.91=6.91$
		$q(\mathbf{z}\|\mathbf{x})$	$11.50+6.90=18.40$
IWAE K=20k	9.90	$p(\mathbf{z})$	$0+9.90=9.90$
		$q(\mathbf{z}\|\mathbf{x})$	$11.50+9.90=21.40$

Table 3: CIFAR-GAN: Decomposition of IWAE LBs on MI to BA term and contrastive term ($< \log K$).

B.4 Bias Reduction of IWAE Lower Bound on $\log p(\mathbf{x})$ / Upper Bound on MI

We were unable to show a similar result for the IWAE lower bound on $\log p(\mathbf{x})$ and upper bound on $I(\mathbf{x};\mathbf{z})$, although Chatterjee et al. (2018) suggests that exponential sample complexity $K \propto \exp\{D_{KL}[p(\mathbf{z}|\mathbf{x})||q_\theta(\mathbf{z}|\mathbf{x})]\}$ is required for accurate importance sampling estimation. Maddison et al. (2017); Domke & Sheldon (2018) find that the bias of the IWAE lower bound in the limit of $K \to \infty$ reduces at the rate of $\mathcal{O}(\frac{1}{2K}\text{Var}[\frac{p(\mathbf{z}|\mathbf{x})}{q_\theta(\mathbf{z}|\mathbf{x})}])$, although the $\text{Var}[\frac{p(\mathbf{x},\mathbf{z})}{q_\theta(\mathbf{z}|\mathbf{x})}]$ term is at least exponential in $D_{KL}[p(\mathbf{z}|\mathbf{x})||q_\theta(\mathbf{z}|\mathbf{x})]$ (Song & Ermon, 2019).

B.5 Relationship with Structured InfoNCE

We can recognize the Structured InfoNCE upper and lower bounds for known $p(\mathbf{x}|\mathbf{z})$ (Poole et al. (2019) Sec. 2.5) as simply applying the standard IWAE bounds, using the marginal $p(\mathbf{z})$ in place of the variational $q_\theta(\mathbf{z}|\mathbf{x})$

$$\mathbb{E}_{p(\mathbf{x})p(\mathbf{z}^{(1)}|\mathbf{x})p(\mathbf{z}^{(2:K)})}\left[\log \frac{p(\mathbf{x}|\mathbf{z}^{(1)})}{\frac{1}{K}\sum_{k=1}^{K} p(\mathbf{x}|\mathbf{z}^{(k)})}\right] \leq I(\mathbf{x};\mathbf{z}) \leq \mathbb{E}_{p(\mathbf{x})p(\mathbf{z}^{(1:K)})}\left[\log \frac{1}{\frac{1}{K}\sum_{k=1}^{K} p(\mathbf{x}|\mathbf{z}^{(k)})}\right] - \mathcal{H}(\mathbf{x}|\mathbf{z}).$$

We refer to the lower bound as $I_{\text{S-InfoNCE}_L}(K)$ and the upper bound as $I_{\text{S-InfoNCE}_U}(K)$. From Prop. 2.1, we obtain an alternative proof that the structured InfoNCE lower bound is upper bounded by $\log K$. Since $I_{BA}(p(\mathbf{z})) \leq I_{\text{IWAE}_L}(p(\mathbf{z}), K) \leq I_{BA}(p(\mathbf{z})) + \log K$ and the BA bound with a prior proposal equals 0 from Eq. 3, we have that $0 \leq I_{\text{S-InfoNCE}} \leq \log K$.

C Generalized IWAE

C.1 Probabilistic Interpretation and Bounds

To derive a probabilistic interpretation for GIWAE, our starting point is to further extend the state space of the IWAE target distribution in Eq. 4, using a uniform index variable $p(s) = \frac{1}{K}\,\forall s$ that specifies which sample $\mathbf{z}^{(k)}$ is drawn from the posterior $p(\mathbf{z}|\mathbf{x})$. This is shown in Fig. 4, and leads to a joint distribution over $(\mathbf{z}^{(1:K)}, s)$

$$p_{\text{TGT}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)}, s|\mathbf{x}) = \frac{1}{K}\, p(\mathbf{z}^{(s)}|\mathbf{x}) \prod_{\substack{k=1\\k\neq s}}^{K} q(\mathbf{z}^{(k)}|\mathbf{x})\,, \tag{41}$$

with marginalization over s leading to $p_{\text{TGT}}^{\text{IWAE}}(\mathbf{z}^{(1:K)}|\mathbf{x})$ in Eq. 4. The posterior over the index variable s, which infers the 'positive' sample drawn from $p(\mathbf{z}|\mathbf{x})$ given a set of samples $\mathbf{z}^{(1:K)}$, corresponds

to the normalized importance weights

$$p_{\text{TGT}}^{\text{GIWAE}}(s|\mathbf{x}, \mathbf{z}^{(1:K)}) = \frac{\frac{p(\mathbf{x},\mathbf{z}^{(s)})}{q_\theta(\mathbf{z}^{(s)}|\mathbf{x})}}{\sum_{k=1}^{K} \frac{p(\mathbf{x},\mathbf{z}^{(k)})}{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}} \tag{42}$$

For the GIWAE extended state space proposal distribution, we consider a categorical index variable $q_{\text{PROP}}^{\text{GIWAE}}(s|\mathbf{z}^{(1:K)}, \mathbf{x})$ drawn according SNIS, with weights calculated using the energy function T_ϕ.

$$q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)}, s|\mathbf{x}) = \left(\prod_{k=1}^{K} q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}^{(k)}|\mathbf{x})\right) q(s|\mathbf{z}^{(1:K)}, \mathbf{x}) \quad \text{where} q_{\text{PROP}}^{\text{GIWAE}}(s|\mathbf{z}^{(1:K)}, \mathbf{x}) = \frac{e^{T_\phi(\mathbf{x},\mathbf{z}^{(s)})}}{\sum\limits_{k=1}^{K} e^{T_\phi(\mathbf{x},\mathbf{z}^{(k)})}}. \tag{43}$$

Similarly to Lawson et al. (2019) and as shown in App. C.3, the induced SNIS distribution performs variational inference of the posterior $p_{\text{TGT}}^{\text{GIWAE}}(s|\mathbf{x}, \mathbf{z}^{(1:K)})$. As we show in Cor. C.3, the optimal energy function $T_\phi(\mathbf{x}, \mathbf{z}) = \log \frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z}|\mathbf{x})} + c(\mathbf{x})$ recovers the IWAE probabilistic interpretation of Domke & Sheldon (2018).

Upper Bound on $\log p(\mathbf{x})$ and Lower Bound on MI To derive bounds on the log partition function, we first calculate the unnormalized density ratio

$$\log \frac{p_{\text{TGT}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)}, s, \mathbf{x})}{q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)}, s|\mathbf{x})} = \log \frac{\frac{1}{K} p(\mathbf{x}, \mathbf{z}^{(s)}) \prod\limits_{\substack{k=1 \\ k\neq s}}^{K} q(\mathbf{z}^{(k)}|\mathbf{x})}{q(s|\mathbf{z}^{(1:K)}, \mathbf{x}) \prod\limits_{k=1}^{K} q(\mathbf{z}^{(k)}|\mathbf{x})} \tag{44}$$

$$= \log \frac{1}{K} \frac{\sum_k e^{T(\mathbf{x},\mathbf{z}^{(k)})}}{e^{T(\mathbf{x},\mathbf{z}^{(s)})}} \frac{p(\mathbf{x}, \mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}|\mathbf{x})} \tag{45}$$

$$= \log \frac{p(\mathbf{x}, \mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}|\mathbf{x})} - T(\mathbf{x}, \mathbf{z}^{(s)}) + \log \frac{1}{K} \sum_{k=1}^{K} e^{T(\mathbf{x},\mathbf{z}^{(k)})} \tag{46}$$

Taking the expectation of the unnormalized log ratio under the proposal or target distribution yields a lower or upper bound, respectively, on $\log p(\mathbf{x})$

$$\underbrace{\mathbb{E}_{q_{\text{PROP}}^{\text{GIWAE}}}\left[\log \frac{p_{\text{TGT}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)}, s, \mathbf{x})}{q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)}, s|\mathbf{x})}\right]}_{\text{ELBO}_{\text{GIWAE}}(q_\theta, T_\phi, K)} \leq \log p(\mathbf{x}) \leq \underbrace{\mathbb{E}_{p_{\text{TGT}}^{\text{GIWAE}}}\left[\log \frac{p_{\text{TGT}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)}, s, \mathbf{x})}{q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)}, s|\mathbf{x})}\right]}_{\text{EUBO}_{\text{GIWAE}}(q_\theta, T_\phi, K)}. \tag{47}$$

To derive an explicit form for the upper bound on $\log p(\mathbf{x})$,

$$\text{EUBO}_{\text{GIWAE}}(q_\theta, T_\phi, K) = \mathbb{E}_{p_{\text{TGT}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)}, s|\mathbf{x})}\left[\log \frac{p(\mathbf{x}, \mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}|\mathbf{x})} - T(\mathbf{x}, \mathbf{z}^{(s)}) + \log \frac{1}{K} \sum_{k=1}^{K} e^{T(\mathbf{x},\mathbf{z}^{(k)})}\right]$$

$$= \mathbb{E}_{p(\mathbf{z}^{(s)}|\mathbf{x})}\left[\log \frac{p(\mathbf{x}, \mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}|\mathbf{x})}\right] - \mathbb{E}_{p(\mathbf{z}^{(s)}|\mathbf{x}) \prod\limits_{k\neq s} q(\mathbf{z}^{(k)}|\mathbf{x})}\left[\log \frac{e^{T(\mathbf{x},\mathbf{z}^{(s)})}}{\frac{1}{K} \sum\limits_{k=1}^{K} e^{T(\mathbf{x},\mathbf{z}^{(k)})}}\right] \tag{48}$$

Translating this to the lower bound on MI from Eq. 9 corresponds to taking $I(\mathbf{x}; \mathbf{z}) = -\mathbb{E}_{p(\mathbf{x})}[\log p(\mathbf{x})] - \mathcal{H}(\mathbf{x}|\mathbf{z}) \geq -\text{EUBO}_{\text{GIWAE}}(q_\theta, T_\phi, K) - \mathcal{H}(\mathbf{x}|\mathbf{z})$. Writing the conditional entropy

term over the index s,

$$I(\mathbf{x};\mathbf{z}) \geq \mathbb{E}_{p(\mathbf{z}^{(s)}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z}^{(s)})] \tag{49}$$

$$- \left(\mathbb{E}_{p(\mathbf{z}^{(s)}|\mathbf{x})} \left[\log \frac{p(\mathbf{x},\mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}|\mathbf{x})} \right] - \mathbb{E}_{p(\mathbf{z}^{(s)}|\mathbf{x})q(\mathbf{z}^{(2:K)}|\mathbf{x})} \left[\log \frac{e^{T(\mathbf{x},\mathbf{z}^{(s)})}}{\frac{1}{K}\sum_{k=1}^{K} e^{T(\mathbf{x},\mathbf{z}^{(k)})}} \right] \right)$$

$$= \mathbb{E}_{p(\mathbf{x},\mathbf{z}^{(s)})} \left[\log \frac{q(\mathbf{z}^{(s)}|\mathbf{x})}{p(\mathbf{z}^{(s)})} \right] + \mathbb{E}_{p(\mathbf{x})p(\mathbf{z}^{(s)}|\mathbf{x})q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})} \left[\log \frac{e^{T_\theta(\mathbf{x},\mathbf{z}^{(s)})}}{\frac{1}{K}\sum_{i=1}^{K} e^{T_\theta(\mathbf{x},\mathbf{z}^{(k)})}} \right] . \tag{50}$$

C.2 GIWAE Upper Bound on MI Does Not Provide Benefit Over IWAE

Taking expectations under $q_{\text{PROP}}^{\text{GIWAE}}$ to obtain a lower bound on $\log p(\mathbf{x})$ or upper bound on MI, notice that we will obtain expectations under the marginal of $\mathbf{z}^{(s)}$, where $\mathbf{z}^{(s)}$ is selected via SNIS as in Eq. 43. In translating this to a bound on mutual information, the conditional entropy term $\mathcal{H}(\mathbf{x}|\mathbf{z})$ takes expectation with respect to $p(\mathbf{z}|\mathbf{x})$. In Eq. (50), this term canceled with the expectation over $\mathbf{z}^{(s)} \sim p_{\text{TGT}}^{\text{GIWAE}}$. However, this will not occur in the case of the GIWAE upper bound on MI.

This implies that knowledge of the joint density $p(\mathbf{x},\mathbf{z})$ is required to evaluate both the conditional entropy and $\log \frac{p(\mathbf{x},\mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}|\mathbf{x})}$ terms. However, if both $p(\mathbf{z})$ and $p(\mathbf{x}|\mathbf{z})$ are known, then we will show in Cor. C.2 below that the optimal energy function in GIWAE yields the true importance weights and the resulting bound matches the IWAE upper bound (see Prop. 2.2). We thus conclude that this upper bound is not useful in practice.

C.3 Relationship between IWAE and GIWAE Probabilistic Interpretations

Proposition C.1. *For a given $q_\theta(\mathbf{z}|\mathbf{x})$, the* IWAE *lower bound on* MI *is tighter than the* GIWAE *lower bound for any $T_\phi(\mathbf{x},\mathbf{z})$, and their difference is the average* KL *divergence between the normalized importance weights $p_{\text{TGT}}^{\text{GIWAE}}(s|\mathbf{z}^{(1:K)},\mathbf{x})$ and the variational distribution $q(s|\mathbf{z}^{(1:K)},\mathbf{x})$ (Eq. 43)*

$$I_{\text{IWAE}_L}(q,K) = I_{\text{GIWAE}_L}(q,T,K) + \mathbb{E}_{p(\mathbf{x})p_{\text{TGT}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)}|\mathbf{x})} \left[D_{KL}[p_{\text{TGT}}^{\text{GIWAE}}(s|\mathbf{z}^{(1:K)},\mathbf{x})||q_{\text{PROP}}^{\text{GIWAE}}(s|\mathbf{z}^{(1:K)},\mathbf{x})] \right] . \tag{51}$$

Proof. To show this result, we consider the posterior over the index variable $p(s|\mathbf{z}^{(1:K)},\mathbf{x})$

$$p_{\text{TGT}}^{\text{GIWAE}}(s|\mathbf{z}^{(1:K)},\mathbf{x}) = \frac{p_{\text{TGT}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)},s,\mathbf{x})}{\sum_{s=1}^{K} p_{\text{TGT}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)},s,\mathbf{x})} = \frac{\frac{1}{K}p(\mathbf{z}^{(s)}|\mathbf{x})\prod_{\substack{k=1\\k\neq s}}^{K} q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}{\frac{1}{K}\sum_{s=1}^{K} p(\mathbf{z}^{(s)}|\mathbf{x})\prod_{\substack{k=1\\k\neq s}}^{K} q_\theta(\mathbf{z}^{(k)}|\mathbf{x})} = \frac{\frac{p(\mathbf{z}^{(s)}|\mathbf{x})}{q(\mathbf{z}^{(s)}|\mathbf{x})}}{\sum_{s=1}^{K}\frac{p(\mathbf{z}^{(s)}|\mathbf{x})}{q(\mathbf{z}^{(s)}|\mathbf{x})}}$$

Now, note that $p(\mathbf{z}^{(1:K)},s,\mathbf{x}) = \big(\sum_{s=1}^{K} p(\mathbf{z}^{(1:K)},s,\mathbf{x})\big)\cdot p(s|\mathbf{z}^{(1:K)},\mathbf{x})$, so that we can decompose the KL divergence as

$$D_{KL}[p_{\text{TGT}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)},s,\mathbf{x})||q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)}s,\mathbf{x})] = \mathbb{E}_{p_{\text{TGT}}^{\text{GIWAE}}} \left[\log \frac{\sum_{s=1}^{K} p(\mathbf{z}^{(1:K)},s,\mathbf{x})}{\prod_{k=1}^{K} q_\theta(\mathbf{z}^{(k)}|\mathbf{x})} \frac{p(s|\mathbf{z}^{(1:K)},\mathbf{x})}{q(s|\mathbf{z}^{(1:K)},\mathbf{x})} \right]$$

$$= D_{KL}[\frac{1}{K}\sum_{s=1}^{K} p(\mathbf{z}^{(s)}|\mathbf{x}) \prod_{\substack{k=1\\k\neq s}}^{K} q_\theta(\mathbf{z}^{(k)}|\mathbf{x})|| \prod_{k=1}^{K} q_\theta(\mathbf{z}^{(k)}|\mathbf{x})] + D_{KL}[p(s|\mathbf{z}^{(1:K)},\mathbf{x})||q(s|\mathbf{z}^{(1:K)},\mathbf{x})]$$

$$= D_{KL}[p_{\text{TGT}}^{\text{IWAE}}(\mathbf{z}^{(1:K)}|\mathbf{x})||q_{\text{PROP}}^{\text{IWAE}}(\mathbf{z}^{(1:K)}|\mathbf{x})] + D_{KL}[p_{\text{TGT}}^{\text{GIWAE}}(s|\mathbf{z}^{(1:K)},\mathbf{x})||q_{\text{PROP}}^{\text{GIWAE}}(s|\mathbf{z}^{(1:K)},\mathbf{x})]$$

Now, the left hand side equals the gap in the GIWAE upper bound on $\log p(\mathbf{x})$ in Eq. (47), while the first term on the RHS is the gap in the IWAE upper bound on $\log p(\mathbf{x})$ in Eq. 36. □

C.4 PROOF OF PROPOSITION 2.2 AND COROLLARY 2.3

We show the following results, which expand upon Prop. 2.2 and Cor. 2.3 in the main text.

Corollary C.2. *The* GIWAE *lower bound has the following properties:*

(a) For a given $q_\theta(\mathbf{z}|\mathbf{x})$, the optimal energy function equals the true log importance weights up to an arbitrary constant: $T^*(\mathbf{x},\mathbf{z}) = \log \frac{p(\mathbf{x},\mathbf{z})}{q_\theta(\mathbf{z}|\mathbf{x})} + c(\mathbf{x})$. With this choice of $T^*(\mathbf{x},\mathbf{z})$,

$$I_{\text{GIWAE}_L}(q_\theta, T^*) = I_{\text{IWAE}_L}(q_\theta) \tag{52}$$

In other words, when the joint distribution $p(\mathbf{x},\mathbf{z})$ is known, the GIWAE *bound does not provide any advantage over the* IWAE *bound.*

(b) For a given $q_\theta(\mathbf{z}|\mathbf{x})$, let $T_{\phi^*}(\mathbf{x},\mathbf{z})$ denote the energy function that maximizes the GIWAE *lower bound when optimization is performed with a neural network parameterized by ϕ. Then, Prop. 2.1, and Prop. C.1 imply*

$$I_{\text{BA}_L}(q_\theta) \leq I_{\text{GIWAE}_L}(q_\theta, T_{\phi^*}, K) \leq I_{\text{IWAE}_L}(q_\theta, K) \leq I_{\text{BA}_L}(q_\theta) + \log K\,. \tag{53}$$

Corollary C.3. *Using the prior $q_\theta(\mathbf{z}|\mathbf{x}) = p(\mathbf{z})$ as in* INFONCE*,*

(a) For the optimal energy function, Cor. C.2(a) implies $T^*(\mathbf{x},\mathbf{z}) = \log p(\mathbf{x}|\mathbf{z}) + c(\mathbf{x})$, and

$$I_{\text{GIWAE}_L}(p(\mathbf{z}), T^*, K) = I_{\text{INFONCE}_L}(T^*, K) = I_{\text{S-INFONCE}_L}(K) \tag{54}$$

(b) For an arbitrary energy function $T_\phi(\mathbf{x},\mathbf{z})$, Cor. C.2(b) translates to

$$0 \leq I_{\text{INFONCE}_L}(T_\phi, K) \leq I_{\text{S-INFONCE}_L}(K) \leq \log K \tag{55}$$

Proof. ***(a)*** We first show that the optimal energy function equals the log true importance weights up to a constant $c(\mathbf{x})$, $T(\mathbf{x},\mathbf{z}) = \log \frac{p(\mathbf{z}|\mathbf{x})}{q(\mathbf{z}|\mathbf{x})} + c(\mathbf{x})$. This implies that learning the energy function using GIWAE becomes unnecessary when this density ratio is available in closed form as in IWAE.

Using Prop. C.1, we can see that the gap in the GIWAE and IWAE bounds, which corresponds to the posterior KL divergence over the index variable s, will equal zero when

$$\begin{aligned} q^{\text{GIWAE}}_{\text{PROP}}(s|\mathbf{z}^{(1:K)},\mathbf{x}) = p^{\text{GIWAE}}_{\text{TGT}}(s|\mathbf{z}^{(1:K)},\mathbf{x}) \implies & \frac{e^{T(\mathbf{x},\mathbf{z}^{(s)})}}{\sum_{k=1}^{K} e^{T(\mathbf{x},\mathbf{z}^{(k)})}} = \frac{\frac{p(\mathbf{z}^{(s)}|\mathbf{x})}{q_\theta(\mathbf{z}^{(s)}|\mathbf{x})}}{\sum_{s=1}^{K} \frac{p(\mathbf{z}^{(s)}|\mathbf{x})}{q_\theta(\mathbf{z}^{(s)}|\mathbf{x})}} \\ \implies & T(\mathbf{x},\mathbf{z}) = \log \frac{p(\mathbf{z}|\mathbf{x})}{q_\theta(\mathbf{z}|\mathbf{x})} + c(\mathbf{x}) \end{aligned} \tag{56}$$

where the last line can include an arbitrary constant $c(\mathbf{x})$ which does not depend on $\mathbf{z}$, since this term will factor out of the denominator in Eq. 56.

For this choice of $T^*(\mathbf{x},\mathbf{z})$, we now show that the value of the GIWAE objective matches the IWAE lower bound on mutual information.

$$\begin{aligned} I_{\text{GIWAE}}(q,T^*) &= \mathbb{E}_{p(\mathbf{x},\mathbf{z}^{(1)})}\left[\log \frac{q(\mathbf{z}^{(1)}|\mathbf{x})}{p(\mathbf{z}^{(1)})}\right] + \mathbb{E}_{p(\mathbf{x})p(\mathbf{z}^{(1)}|\mathbf{x})q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \frac{e^{\log \frac{p(\mathbf{z}^{(1)}|\mathbf{x})}{q(\mathbf{z}^{(1)}|\mathbf{x})}} \cdot e^{c(\mathbf{x})}}{\left(\frac{1}{K}\sum_{i=1}^{K} e^{\log \frac{p(\mathbf{z}^{(k)}|\mathbf{x})}{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}}\right) \cdot e^{c(\mathbf{x})}}\right] \\ &= \mathbb{E}_{p(\mathbf{x},\mathbf{z}^{(1)})}\left[\log \frac{q(\mathbf{z}^{(1)}|\mathbf{x})}{p(\mathbf{z}^{(1)})} + \log \frac{p(\mathbf{z}^{(1)},\mathbf{x})}{q(\mathbf{z}^{(1)}|\mathbf{x})}\right] - \mathbb{E}_{p(\mathbf{x})p(\mathbf{z}^{(1)}|\mathbf{x})q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \frac{1}{K}\sum_{i=1}^{K} \frac{p(\mathbf{z}^{(k)},\mathbf{x})}{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}\right] \\ &= -\mathcal{H}(\mathbf{x}|\mathbf{z}) + \left(-\mathbb{E}_{p(\mathbf{x})p(\mathbf{z}^{(1)}|\mathbf{x})q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \frac{1}{K}\sum_{i=1}^{K} \frac{p(\mathbf{z}^{(k)},\mathbf{x})}{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}\right]\right) \\ &= I_{\text{IWAE}_L}(q_\theta, K) \end{aligned} \tag{57}$$

The second term is exactly the negative of the IWAE upper bound on $\log p(\mathbf{x})$ in Eq. (36), which provides a lower bound of $\mathcal{H}(\mathbf{x}) = -\mathbb{E}_{p(\mathbf{x})}[\log p(\mathbf{x})]$. Combined with the entropy $-\mathcal{H}(\mathbf{x}|\mathbf{z})$, we obtain the I_{IWAE} lower bound on MI. See Section 2.2.

(b) We begin by showing that $I_{\text{BA}_L}(q_\theta) \leq I_{\text{GIWAE}_L}(q_\theta, T_{\phi^*}, K)$. This follows from the fact that there exist ϕ_0 in the pearameter space of the neural network T_ϕ such that $T_{\phi_0} = 0$. With this ϕ_0, we would have $I_{\text{BA}_L}(q_\theta) = I_{\text{GIWAE}_L}(q_\theta, T_{\phi_0}, K)$. So the optimal ϕ^* in the parameter space, can only improve upon the BA bound. Also note that for any given ϕ including ϕ^*, we have: $I_{\text{GIWAE}_L}(q_\theta, T_{\phi^*}, K) \leq I_{\text{IWAE}_L}(q_\theta, K)$, as the T_{ϕ^*} is the optimal energy function. This relationship is visualized in Fig. 3b.

□

D GENERALIZED MUTUAL INFORMATION NEURAL ESTIMATION (GMINE)

In this section, we provide probabilistic interpretations for the MINE lower bounds on MI from Belghazi et al. (2018), which allows us to derive novel *Generalized MINE* bounds. In a similar spirit to GIWAE, these bounds use a base variational distribution $q_\theta(\mathbf{z}|\mathbf{x})$ to tighten the MINE-DV and MINE-F bounds and can be evaluated when $p(\mathbf{z})$ is available.

We begin by deriving a probabilistic interpretation for the IBAL lower bound in Sec. 4, which we will show is closely related to the probabilistic interpretations of MINE. Our MINE-AIS method is designed to optimize and evaluate the IBAL lower bound on MI, with detailed discussion in Sec. 4 and App. E. See Fig. 5 for a summary of Generalized MINE bounds and their relationships.

D.1 PROBABILISTIC INTERPRETATION OF IBAL

For the variational approximation

$$\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x}) = \frac{1}{\mathcal{Z}_{\theta,\phi}(\mathbf{x})} q_\theta(\mathbf{z}|\mathbf{x}) e^{T_\phi(\mathbf{x},\mathbf{z})} \quad \text{where} \quad \mathcal{Z}_{\theta,\phi}(\mathbf{x}) = \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right], \tag{58}$$

we consider the BA lower bound on MI,

$$\begin{aligned}
I_{\text{BA}_L}(\pi_{\theta,\phi}) &= I(\mathbf{x},\mathbf{z}) - \mathbb{E}_{p(\mathbf{x})}\left[D_{\text{KL}}[p(\mathbf{z}|\mathbf{x})||\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})]\right] \\
&= \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})p(\mathbf{z})}\right] - \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log p(\mathbf{z}|\mathbf{x})\right] + \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[T(\mathbf{x},\mathbf{z})\right] - \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \mathcal{Z}_{\theta,\phi}\right] \\
&= \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\theta(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right] + \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[T_\phi(\mathbf{x},\mathbf{z})\right] - \mathbb{E}_{p(\mathbf{x})}\left[\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})\right] \\
&= \underbrace{\mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\theta(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right]}_{I_{\text{BA}_L}(q_\theta)} + \underbrace{\mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{e^{T_\phi(\mathbf{x},\mathbf{z})}}{\mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right]}\right]}_{\text{contrastive term}} \\
&:= \text{IBAL}(q_\theta, T_\phi)
\end{aligned}$$

so that the gap in the mutual information bound is $\mathbb{E}_{p(\mathbf{x})}\left[D_{\text{KL}}[p(\mathbf{z}|\mathbf{x})||\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})]\right]$. Note that $\text{IBAL}(q_\theta, T_\phi)$ contains $I_{\text{BA}_L}(q_\theta)$ as one of its terms. We visualize the relationship between various energy-based bounds in Fig. 3b and Fig. 5.

D.2 PROBABILISTIC INTERPRETATION OF GENERALIZED MINE-DV

We can interpret the MINE-DV bound of Belghazi et al. (2018) as arising from a joint energy-based variational approximation $\pi_{\theta,\phi}(\mathbf{x},\mathbf{z})$ of the full joint distribution $p(\mathbf{x},\mathbf{z})$ as follows

$$\pi_{\theta,\phi}(\mathbf{x},\mathbf{z}) = \frac{1}{\mathcal{Z}} p(\mathbf{x}) q_\theta(\mathbf{z}|\mathbf{x}) e^{T_\phi(\mathbf{x},\mathbf{z})} \quad \text{where} \quad \mathcal{Z} = \mathbb{E}_{p(\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right]. \tag{59}$$

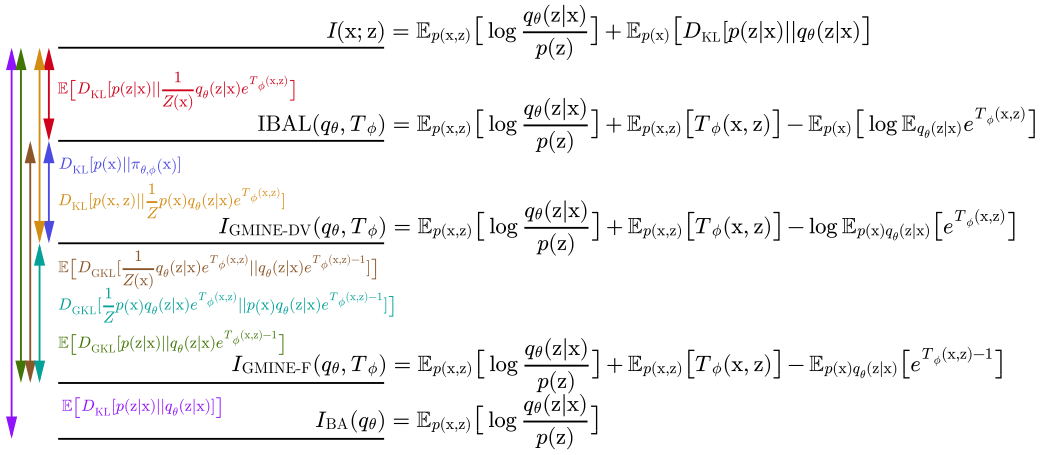


Figure 5: Generalized Energy Based Bounds. Arrows indicate the gaps in each MI lower bound or its relationship to other lower bounds. $D_{\mathrm{GKL}}(\cdot||\cdot)$ represents the *generalized* KL divergence between two unnormalized densities (see App. D.3). All bounds are written in terms of a base variational distribution $q_\theta(\mathbf{z}|\mathbf{x})$, which may be chosen to be the marginal $p(\mathbf{z})$ as in MINE-DV and MINE-F.

where $q_\theta(\mathbf{z}|\mathbf{x})$ is a base variational distribution, and T_ϕ is an energy function.

Note that for the induced marginal of the energy-based approximation, $\pi_{\theta,\phi}(\mathbf{x}) := \int \pi_{\theta,\phi}(\mathbf{x},\mathbf{z})d\mathbf{z} \neq p(\mathbf{x})$ due to the contribution of the energy term. Instead, we have

$$\pi_{\theta,\phi}(\mathbf{x}) = \frac{1}{\mathcal{Z}}p(\mathbf{x})\mathcal{Z}(\mathbf{x}) \tag{60}$$

where $\mathcal{Z}(\mathbf{x}) = \int q(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})}d\mathbf{z}$ and $\mathcal{Z} = \int p(\mathbf{x})q(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})}d\mathbf{x}d\mathbf{z} = \mathbb{E}_{p(\mathbf{x})}[\mathcal{Z}(\mathbf{x})]$.

Subtracting a joint KL divergence $D_{\mathrm{KL}}[p(\mathbf{x},\mathbf{z})||\pi_{\theta,\phi}(\mathbf{x},\mathbf{z})]$ from $I(\mathbf{x};\mathbf{z})$, we obtain the *Generalized* MINE-DV lower bound on MI

$$I(\mathbf{x};\mathbf{z}) \geq I_{\text{GMINE-DV}}(q_\theta, T_\phi) := I(\mathbf{x};\mathbf{z}) - D_{\mathrm{KL}}[p(\mathbf{x},\mathbf{z})||\frac{1}{\mathcal{Z}}p(\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})}] \tag{61}$$

$$= \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\theta(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right] + \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[T_\phi(\mathbf{x},\mathbf{z})\right] - \log \mathbb{E}_{p(\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right]. \tag{62}$$

By construction, Eq. (61) shows that the gap in Generalized MINE-DV is $D_{\mathrm{KL}}[p(\mathbf{x},\mathbf{z})||\frac{1}{\mathcal{Z}}p(\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})}]$, which has a probabilistic interpretation in terms of the approximate joint distribution $\pi_{\theta,\phi}(\mathbf{x},\mathbf{z})$.

Relationship with MINE-DV For $q_\theta(\mathbf{z}|\mathbf{x}) = p(\mathbf{z})$, we obtain the MINE-DV bound (Belghazi et al., 2018; Poole et al., 2019) as a special case with $I_{\text{MINE-DV}}(T_\phi) = I_{\text{GMINE-DV}}(p(\mathbf{z}), T_\phi)$. In particular, the joint base distribution in Eq. (59) corresponds to the product of marginals $p(\mathbf{x})p(\mathbf{z})$. Our probabilistic interpretation shows that the gap in MINE-DV corresponds to $D_{\mathrm{KL}}[p(\mathbf{x},\mathbf{z})||\frac{1}{\mathcal{Z}}p(\mathbf{x})p(\mathbf{z})e^{T_\phi(\mathbf{x},\mathbf{z})}]$.

We expect that the Generalized MINE, with a learned variational distribution $q_\theta(\mathbf{z}|\mathbf{x})$, to obtain tighter bounds than MINE-DV. We can guarantee that $I_{\text{GMINE-DV}}(q_\theta^*, T_\phi) \geq I_{\text{MINE-DV}}(T_\phi)$ for the optimal q_θ^* with a given T_ϕ, so long as $p(\mathbf{z})$ is in the variational family (i.e., $\exists\, \theta_0$ such that $q_{\theta_0}(\mathbf{z}|\mathbf{x}) = p(\mathbf{z})$).

Relationship with BA Bound Choosing a constant energy function $T_{\phi_0}(\mathbf{x},\mathbf{z}) = c$, we can see that $\pi_{\theta,\phi_0}(\mathbf{x},\mathbf{z}) = p(\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x})$ and $I_{\text{GMINE-DV}}(q_\theta, T_{\phi_0} = c) = I_{\mathrm{BA}_L}(q_\theta)$ for a given $q_\theta(\mathbf{z}|\mathbf{x})$.

Optimal Energy Function For a given $q_\theta(\mathbf{z}|\mathbf{x})$, the optimal energy function of Generalized MINE-DV corresponds to the log importance weights between the target $p(\mathbf{x},\mathbf{z})$ and the joint base distribution

$p(\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x})$, plus a constant

$$T^*(\mathbf{x},\mathbf{z}) = \log \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x})} + c\,. \tag{63}$$

For the optimal energy function associated with a given $q_\theta(\mathbf{z}|\mathbf{x})$, we have $I_{\text{MINE-DV}}(q_\theta, T^*) = I(\mathbf{x};\mathbf{z})$.

Relationship with IBAL We can use our probabilistic interpretation to show that the IBAL is tighter than Generalized MINE-DV (i.e., $\text{IBAL}(q_\theta, T_\phi) \geq I_{\text{GMINE-DV}}(q_\theta, T_\phi)$), and the difference corresponds to a marginal KL divergence in the $\mathbf{x}$ space

$$\text{IBAL}(q_\theta, T_\phi) = I_{\text{GMINE-DV}}(q_\theta, T_\phi) + D_{\text{KL}}[p(\mathbf{x})||\pi_{\theta,\phi}(\mathbf{x})] \geq I_{\text{GMINE-DV}}\,. \tag{64}$$

where $\pi_{\theta,\phi}(\mathbf{x})$ is defined in Eq. (60). As in Poole et al. (2019), we can also interpret the gap between the $\text{IBAL}(q_\theta, T_\phi)$ and $I_{\text{GMINE-DV}}(q_\theta, T_\phi)$ as an application of Jensen's inequality, with $D_{\text{KL}}[p(\mathbf{x})||\pi_{\theta,\phi}(\mathbf{x})] = \log \mathcal{Z} - \mathbb{E}_{p(\mathbf{x})}[\log \mathcal{Z}(\mathbf{x})] = \log \mathbb{E}_{p(\mathbf{x})}[\mathcal{Z}(\mathbf{x})] - \mathbb{E}_{p(\mathbf{x})}[\log \mathcal{Z}(\mathbf{x})]$ and $\log \mathbb{E}_{p(\mathbf{x})}[\mathcal{Z}(\mathbf{x})] \geq \mathbb{E}_{p(\mathbf{x})}[\log \mathcal{Z}(\mathbf{x})]$.

D.3 Probabilistic Interpretation of Generalized MINE-F

The BA lower bound and its corresponding EUBO upper bound on $\log p(\mathbf{x})$ are derived using a KL divergence between the true and approximate posterior, which are normalized conditional distributions over $\mathbf{z}$ given $\mathbf{x}$. In this section, we interpret the MINE-F bound (Belghazi et al., 2018) as arising from a generalized notion of the KL divergence between possibly *unnormalized* density functions. In particular, our probabilistic interpretation will involve an unnormalized $\tilde{\pi}_{\theta,\phi}(\mathbf{z}|\mathbf{x})$ which seeks to approximate the true (normalized) posterior $p(\mathbf{z}|\mathbf{x})$.

Generalized KL Divergence First, we recall the definition of the *generalized* KL divergence, which takes unnormalized measures as input arguments and corresponds to the limiting behavior of the α-divergence (Minka et al., 2005; Cichocki & Amari, 2010)

$$D_{\text{GKL}}[\tilde{r}(\mathbf{z})||\tilde{s}(\mathbf{z})] = \int \tilde{r}(\mathbf{z}) \log \frac{\tilde{r}(\mathbf{z})}{\tilde{s}(\mathbf{z})} d\mathbf{z} - \int \tilde{r}(\mathbf{z}) d\mathbf{z} + \int \tilde{s}(\mathbf{z}) d\mathbf{z}\,. \tag{65}$$

As in the case of the standard KL divergence, this quantity is nonnegative, convex in either argument, and vanishes for $\tilde{r}(\mathbf{z}) = \tilde{s}(\mathbf{z})$. It is also a member of both the family of Bregman divergences and f-divergences (Amari, 2009). If $r(\mathbf{z})$ and $s(\mathbf{z})$ are normalized, then $D_{\text{GKL}}[r(\mathbf{z})||s(\mathbf{z})] = D_{\text{KL}}[r(\mathbf{z})||s(\mathbf{z})]$. Finally, if $r(\mathbf{z})$ is normalized and $\tilde{s}(\mathbf{z})$ is unnormalized, one can easily confirm that

$$\begin{aligned} D_{\text{GKL}}[r(\mathbf{z})||\tilde{s}(\mathbf{z})] &= D_{\text{KL}}[r(\mathbf{z})||s(\mathbf{z})] + D_{\text{GKL}}[s(\mathbf{z})||\tilde{s}(\mathbf{z})] \\ &\geq D_{\text{KL}}[r(\mathbf{z})||s(\mathbf{z})]\,. \end{aligned} \tag{66}$$

Generalized EUBO Since the Generalized KL divergence is always nonnegative, we can define an Generalized EUBO by adding the Generalized KL divergence between the normalized true posterior and an unnormalized approximate posterior $\tilde{\pi}(\mathbf{z}|\mathbf{x})$.

$$\text{GEUBO}(\tilde{\pi}) := \log p(\mathbf{x}) + D_{\text{GKL}}[p(\mathbf{z}|\mathbf{x})||\tilde{\pi}(\mathbf{z}|\mathbf{x})] \tag{67}$$

$$= \mathbb{E}_{p(\mathbf{z}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{\tilde{\pi}(\mathbf{z}|\mathbf{x})}\right] - 1 + \mathcal{Z}_{\tilde{\pi}}(\mathbf{x}) \tag{68}$$

where we assume that $\int p(\mathbf{z}|\mathbf{x})d\mathbf{z} = 1$ and let $\mathcal{Z}_{\tilde{\pi}}(\mathbf{x}) := \int \tilde{\pi}(\mathbf{z}|\mathbf{x})d\mathbf{z}$. Note that for normalized $\pi(\mathbf{z}|\mathbf{x})$, we have $\text{EUBO}(\pi) = \text{GEUBO}(\pi)$. In general, if π is the normalized distribution of $\tilde{\pi}$, we have $\text{EUBO}(\pi) \leq \text{GEUBO}(\tilde{\pi})$.

Generalized BA Using the Generalized EUBO in place of $\log p(\mathbf{x})$, we obtain the following lower bound on MI, which we denote as the Generalized BA lower bound

$$\begin{aligned} I(\mathbf{x};\mathbf{z}) \geq I_{\text{GBA}_L}(\tilde{\pi}) &:= I(\mathbf{x};\mathbf{z}) - \mathbb{E}_{p(\mathbf{x})}\left[D_{\text{GKL}}[p(\mathbf{z}|\mathbf{x})||\tilde{\pi}(\mathbf{z}|\mathbf{x})]\right] \\ &= \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{\tilde{\pi}(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right] + 1 - \mathbb{E}_{p(\mathbf{x})}\left[\mathcal{Z}_{\tilde{\pi}}(\mathbf{x})\right]. \end{aligned} \tag{69}$$

Generalized MINE-F We now consider an unnormalized, energy-based approximation to the true posterior, involving a base variational distribution $q_\theta(\mathbf{z}|\mathbf{x})$ and a learned energy function $T_\phi(\mathbf{x},\mathbf{z})$

$$\tilde{\pi}_{\theta,\phi}(\mathbf{z}|\mathbf{x}) = q_\theta(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})-1}\,. \tag{70}$$

Using this unnormalized approximate posterior in the GBA lower bound in Eq. (69), we obtain the *Generalized* MINE-F (GMINE-F) lower bound

$$I(\mathbf{x};\mathbf{z}) \geq I_{\text{GMINE-F}}(q_\theta, T_\phi) := I(\mathbf{x};\mathbf{z}) - \mathbb{E}_{p(\mathbf{x})}\left[D_{\text{GKL}}[p(\mathbf{z}|\mathbf{x})||\tilde{\pi}_{\theta,\phi}(\mathbf{z}|\mathbf{x})]\right] \tag{71}$$

$$= \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log\frac{q_\theta(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right] + \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[T_\phi(\mathbf{x},\mathbf{z})\right] - \mathbb{E}_{p(\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})-1}\right] \tag{72}$$

By construction, we can see that the gap in $I_{\text{GMINE-F}}(q_\theta, T_\phi)$ is equal to the Generalized KL divergence $\mathbb{E}_{p(\mathbf{x})}\left[D_{\text{GKL}}[p(\mathbf{z}|\mathbf{x})||\tilde{\pi}_{\theta,\phi}(\mathbf{z}|\mathbf{x})]\right]$ in Eq. (71). As in the case of MINE-DV, we obtain the standard MINE-F lower bound $I_{\text{MINE-F}}(T_\phi) = I_{\text{GMINE-F}}(p(\mathbf{z}), T_\phi)$ when using the marginal $p(\mathbf{z})$ as the proposal.

Optimal Energy Function The optimal energy function of Generalized MINE-F is $T^*(\mathbf{x},\mathbf{z}) = 1 + \log\frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})q(\mathbf{z}|\mathbf{x})}$. In this case, we obtain $I_{\text{GMINE-F}}(q_\theta, T^*) = I(\mathbf{x};\mathbf{z})$ in Eq. (72) (Poole et al., 2019).

Relationship with IBAL We would now like to relate the gap in $I_{\text{GMINE-F}}(q_\theta, T_\phi)$ to the gap in $I_{\text{IBAL}}(q_\theta, T_\phi)$. First, note that the normalized distribution corresponding to $\tilde{\pi}_{\theta,\phi}(\mathbf{z}|\mathbf{x}) = q_\theta(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})-1}$ matches the the energy-based posterior in the IBAL, $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x}) = \frac{1}{\mathcal{Z}(\mathbf{x})}q_\theta(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})}$. Thus, using Eq. (66), we can write

$$\mathbb{E}_{p(\mathbf{x})}\left[D_{\text{GKL}}[p(\mathbf{z}|\mathbf{x})||\tilde{\pi}_{\theta,\phi}(\mathbf{z}|\mathbf{x})]\right] = \mathbb{E}_{p(\mathbf{x})}\left[D_{\text{KL}}[p(\mathbf{z}|\mathbf{x})||\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})]\right] + \mathbb{E}_{p(\mathbf{x})}\left[D_{\text{GKL}}[\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})||\tilde{\pi}_{\theta,\phi}(\mathbf{z}|\mathbf{x})]\right]$$

$$\underbrace{\mathbb{E}_{p(\mathbf{x})}\left[D_{\text{GKL}}[p(\mathbf{z}|\mathbf{x})||q_\theta(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})-1}]\right]}_{\text{gap of GMINE-F}} = \underbrace{\mathbb{E}_{p(\mathbf{x})}\left[D_{\text{KL}}[p(\mathbf{z}|\mathbf{x})||\frac{1}{\mathcal{Z}(\mathbf{x})}q_\theta(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})}]\right]}_{\text{gap of IBAL}} + \underbrace{\mathbb{E}_{p(\mathbf{x})}\left[D_{\text{GKL}}[\frac{1}{\mathcal{Z}(\mathbf{x})}q_\theta(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})}||q_\theta(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})-1}]\right]}_{\geq 0} \tag{73}$$

where the final term $D_{\text{GKL}}[\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})||\tilde{\pi}_{\theta,\phi}(\mathbf{z}|\mathbf{x})] \geq 0$ is nonnegative because it is a generalized KL divergence. After simplification, we can also see that $D_{\text{GKL}}[\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})||\tilde{\pi}_{\theta,\phi}(\mathbf{z}|\mathbf{x})] = \mathbb{E}_{p(\mathbf{x})}\left[\mathcal{Z}(\mathbf{x}) - 1 - \log\mathcal{Z}(\mathbf{x})\right] \geq 0$ since $\log u \leq u - 1$. We visualize the relationship between IBAL and Generalized MINE-DV in Fig. 5.

Relationship with Generalized MINE-DV To characterize the gap between Generalized MINE-DV and Generalized MINE-F, we can again use the equality from Eq. (66), but this time using divergences over joint distributions.

$$D_{\text{GKL}}[p(\mathbf{x},\mathbf{z})||\tilde{\pi}_{\theta,\phi}(\mathbf{x},\mathbf{z})] = D_{\text{KL}}[p(\mathbf{x},\mathbf{z})||\pi_{\theta,\phi}(\mathbf{x},\mathbf{z})] + D_{\text{GKL}}[\pi_{\theta,\phi}(\mathbf{x},\mathbf{z})||\tilde{\pi}_{\theta,\phi}(\mathbf{x},\mathbf{z})]$$

$$\underbrace{D_{\text{GKL}}[p(\mathbf{x},\mathbf{z})||p(\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})-1}]}_{\text{gap in GMINE-F}} = \underbrace{D_{\text{KL}}[p(\mathbf{x},\mathbf{z})||\frac{1}{\mathcal{Z}}p(\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})}]}_{\text{gap in GMINE-DV}} + \underbrace{D_{\text{GKL}}[\frac{1}{\mathcal{Z}}p(\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})}||p(\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})-1}]}_{\geq 0} \tag{74}$$

In this case, we have $D_{\text{GKL}}[\frac{1}{\mathcal{Z}}p(\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})}||p(\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x})e^{T_\phi(\mathbf{x},\mathbf{z})-1}] \geq 0$. Thus, Generalized MINE-DV is tighter than Generalized MINE-F (see Fig. 5), which generalizes the finding in Poole et al. (2019) that standard MINE-DV is tighter than standard MINE-F.

E MINE-AIS

E.1 Properties of MINE-AIS

Proposition E.1. *The lower bound* IBAL(q_θ, T_ϕ) *used in* MINE-AIS *has the following properties:*

(a) For a given $q_\theta(\mathbf{z}|\mathbf{x})$, the optimal energy function equals the true log importance weights up to an arbitrary constant: $T^*(\mathbf{x},\mathbf{z}) = \log \frac{p(\mathbf{x},\mathbf{z})}{q_\theta(\mathbf{z}|\mathbf{x})} + c(\mathbf{x})$. With this choice of $T^*(\mathbf{x},\mathbf{z})$,

$$\text{IBAL}(q_\theta, T^*) = I(\mathbf{x};\mathbf{z}) \tag{75}$$

and Cor. C.2(a) implies that

$$\lim_{K\to\infty} I_{\text{GIWAE}_L}(q_\theta, T^*, K) = \lim_{K\to\infty} I_{\text{IWAE}_L}(q_\theta, K) = \text{IBAL}(q_\theta, T^*) = I(\mathbf{x};\mathbf{z}) \tag{76}$$

(b) Suppose the energy function $T_\phi(\mathbf{x},\mathbf{z})$ is parameterized by ϕ, and that $\exists\, \phi_0 \; s.t. \forall\, (\mathbf{x},\mathbf{z}), T_{\phi_0}(\mathbf{x},\mathbf{z}) = const.$ For a given $q_\theta(\mathbf{z}|\mathbf{x})$, let $T_{\phi^*}(\mathbf{x},\mathbf{z})$ denote the energy function that maximizes the MINE-AIS *lower bound* $\text{IBAL}(q_\theta, T_\phi)$*, then we have*

$$I_{\text{BA}_L}(q_\theta) \le \text{IBAL}(q_\theta, T_{\phi^*}) \le I(\mathbf{x};\mathbf{z}) = I_{\text{BA}_L}(q_\theta) + D_{KL}[p(\mathbf{z}|\mathbf{x})||q_\theta(\mathbf{z}|\mathbf{x})] \tag{77}$$

In particular, the contrastive term in Eq. 22 is upper bounded by $D_{KL}[p(\mathbf{z}|\mathbf{x})||q_\theta(\mathbf{z}|\mathbf{x})]$. See Fig. 3b.

Proof. **(a)** Since the gap in the $\text{IBAL}(q_\theta, T_{\phi^*})$ is $\mathbb{E}_{p(\mathbf{x})}\big[D_{KL}[p(\mathbf{z}|\mathbf{x})||\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})]\big] = I(\mathbf{x};\mathbf{z}) - \text{IBAL}(q_\theta, T_{\phi^*})$. This implies that the bound will be optimal iff $p(\mathbf{z}|\mathbf{x}) = \pi_{\theta,\phi}(\mathbf{z}|\mathbf{x}) \propto p(\mathbf{x},\mathbf{z})$. We can easily show that the true log importance weights (plus a constant) satisfy this property and thus leads to the optimal energy function $T^*(\mathbf{x},\mathbf{z}) = \log p(\mathbf{x},\mathbf{z})/q_\theta(\mathbf{z}|\mathbf{x})$,

$$\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x}) = \frac{1}{\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})} q_\theta(\mathbf{z}|\mathbf{x}) e^{\log \frac{p(\mathbf{x},\mathbf{z})}{q_\theta(\mathbf{z}|\mathbf{x})} + c(\mathbf{x})} = \frac{e^{c(\mathbf{x})}}{\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})} p(\mathbf{x},\mathbf{z}) = p(\mathbf{z}|\mathbf{x}) \tag{78}$$

where $e^{c(\mathbf{x})}$ is absorbed into the normalization constant. At this value, the IBAL reduces to $I(\mathbf{x};\mathbf{z})$

$$\begin{aligned}
\text{IBAL}(q_\theta, T^*) &= \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\theta(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right] + \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[T_\phi(\mathbf{x},\mathbf{z})\right] - \mathbb{E}_{p(\mathbf{x})}\left[\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})\right] && (79)\\
&= \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{q_\theta(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right] + \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{q_\theta(\mathbf{z}|\mathbf{x})} + \log c(\mathbf{x}))\right] && (80)\\
&\quad - \mathbb{E}_{p(\mathbf{x})}\left[\log \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})} \frac{p(\mathbf{x},\mathbf{z})}{q_\theta(\mathbf{z}|\mathbf{x})} + \log c(\mathbf{x})\right] && (81)\\
&= \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{z})p(\mathbf{x})}\right] = I(\mathbf{x};\mathbf{z}) && (82)
\end{aligned}$$

(b) The BA bound is a special case of the $\text{IBAL}(q_\theta, T_\phi)$ with constant $T_{\phi_0} = c$. Thus, by learning T_ϕ, we can only improve upon $I_{\text{BA}_L}(q_\theta)$. The paramterized family of T_ϕ may not be expressive enough to match the true log importance weights, so $\text{IBAL}(q_\theta, T_{\phi^*}) \le \text{IBAL}(q_\theta, T^*) = I(\mathbf{x};\mathbf{z})$ by part (a). □

E.2 Proof of Proposition 4.1

Proposition 4.1. *For given $q_\theta(\mathbf{z}|\mathbf{x})$ and $T_\phi(\mathbf{x},\mathbf{z})$, $\lim_{K\to\infty} I_{\text{GIWAE}_L}(q_\theta, T_\phi, K) = \text{IBAL}(q_\theta, T_\phi)$.*

Proof. Inspecting the form of $I_{\text{GIWAE}_L}(q_\theta, T_\phi, K)$ in Eq. 9 to the $\text{IBAL}(q_\theta, T_\phi)$ in Eq. 22, we can see that for a fixed q_θ and T_ϕ, both bounds include the same first term $I_{\text{BA}_L}(q_\theta)$ and the same numerator of the contrastive term $\mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right]$. To prove the proposition, we can thus focus on characterizing the limiting behavior of the denominator

$$\lim_{K\to\infty} \underbrace{\mathbb{E}_{p(\mathbf{z}^{(1)}|\mathbf{x})q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \frac{1}{K}\sum_{k=1}^{K} e^{T_\phi(\mathbf{x},\mathbf{z})}\right]}_{:=\log \hat{\mathcal{Z}}_{\text{GIWAE}}(K,\mathbf{x})} = \underbrace{\log \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right]}_{:=\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})} \quad \forall \mathbf{x} \tag{83}$$

where we introduce the notation $\log \hat{\mathcal{Z}}_{\text{GIWAE}}(K,\mathbf{x})$ for convenience and the right hand side is the log partition function for the MINE-AIS distribution $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$. Intuitively, we expect Eq. 83 to hold

since the contribution of the single posterior sample $p(\mathbf{z}|\mathbf{x})$ in the GIWAE expectation will vanish as $K \to \infty$.

More formally, we consider the sequence of values $\log \hat{\mathcal{Z}}_{\text{GIWAE}}(K, \mathbf{x})$ as a function of K. We derive lower and upper bounds on the value of $\log \hat{\mathcal{Z}}_{\text{GIWAE}}(K, \mathbf{x})$ for fixed K and show that each of these sequences of lower and upper bounds converge to $\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})$ as $K \to \infty$. Using the squeeze theorem for sequences, this is sufficient to demonstrate the claim that $\lim_{K\to\infty} \log \hat{\mathcal{Z}}_{\text{GIWAE}}(K, \mathbf{x}) = \log \mathcal{Z}_{\theta,\phi}(\mathbf{x})$ in Eq. 83. Since the other terms in $I_{\text{GIWAE}_L}(q_\theta, T_\phi, K)$ and $\text{IBAL}(q_\theta, T_\phi)$ are identical, this is sufficient to prove the proposition.

Lower Bound: To lower bound $\log \hat{\mathcal{Z}}_{\text{GIWAE}}(K, \mathbf{x})$, we rely on the fact that the exponential function $e^{T_\phi(\mathbf{x},\mathbf{z})} \geq 0$ to simply ignore the contribution of the $p(\mathbf{z}|\mathbf{x})$ term.

$$\log \hat{\mathcal{Z}}_{\text{GIWAE}}(K, \mathbf{x}) = \mathbb{E}_{p(\mathbf{z}^{(1)}|\mathbf{x})q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \sum_{k=1}^{K} e^{T_\phi(\mathbf{x},\mathbf{z})}\right] - \log K \tag{84}$$

$$\leq \mathbb{E}_{q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \sum_{k=2}^{K} e^{T_\phi(\mathbf{x},\mathbf{z})}\right] - \log K \tag{85}$$

$$= \mathbb{E}_{q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \frac{1}{K-1} \sum_{k=2}^{K} e^{T_\phi(\mathbf{x},\mathbf{z})}\right] + \log \frac{K-1}{K} \tag{86}$$

$$\leq \log \frac{1}{K-1} \sum_{k=2}^{K} \mathbb{E}_{q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right] + \log \frac{K-1}{K} \tag{87}$$

$$= \log \mathbb{E}_{q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right] + \log \frac{K-1}{K} \tag{88}$$

$$:= \text{LB}\left(\log \hat{\mathcal{Z}}_{\text{GIWAE}}(K, \mathbf{x})\right) \tag{89}$$

As $K \to \infty$, we have $\log \frac{K-1}{K} \to 0$ and

$$\lim_{K\to\infty} \text{LB}\left(\log \hat{\mathcal{Z}}_{\text{GIWAE}}(K, \mathbf{x})\right) = \log \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right] = \log \mathcal{Z}_{\theta,\phi}(\mathbf{x}) \tag{90}$$

Upper Bound: To upper bound $\log \hat{\mathcal{Z}}_{\text{GIWAE}}(K, \mathbf{x})$, we separately consider terms arising from $p(\mathbf{z}|\mathbf{x})$ samples and $q_\theta(\mathbf{z}|\mathbf{x})$ samples. Noting that $p_{\text{TGT}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)}, s|\mathbf{x})$ in Eq. (41) is invariant to the index s, we can assume $\mathbf{z}^{(1)} \sim p(\mathbf{z}|\mathbf{x})$ and write

$$\log \hat{\mathcal{Z}}_{\text{GIWAE}}(K, \mathbf{x}) = \mathbb{E}_{p(\mathbf{z}^{(1)}|\mathbf{x})q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log\left(\frac{1}{K}\sum e^{T_\phi(\mathbf{x},\mathbf{z}^{(1)})} + \frac{1}{K}\sum_{k=2}^{K} e^{T_\phi(\mathbf{x},\mathbf{z}^{(k)})}\right)\right] \tag{91}$$

$$= \mathbb{E}_{p(\mathbf{z}^{(1)}|\mathbf{x})q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log\left(\frac{1}{K}e^{T_\phi(\mathbf{x},\mathbf{z}^{(1)})} + \frac{K-1}{K}\cdot\frac{1}{K-1}\sum_{k=2}^{K} e^{T_\phi(\mathbf{x},\mathbf{z}^{(k)})}\right)\right] \tag{92}$$

We consider the term inside the log as an expectation over a Bernoulli random variable with probabilities $\{\frac{1}{K}, \frac{K-1}{K}\}$. Using the concavity of log to bring this expectation outside, and the linearity of expectation to separate terms, we have

$$\log \hat{\mathcal{Z}}_{\text{GIWAE}}(K, \mathbf{x}) \leq \frac{1}{K}\mathbb{E}_{p(\mathbf{z}^{(1)})}\left[T_\phi(\mathbf{x}, \mathbf{z}^{(1)})\right] + \frac{K-1}{K}\mathbb{E}_{q_\theta(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \frac{1}{K-1}\sum_{k=2}^{K} e^{T_\phi(\mathbf{x},\mathbf{z}^{(k)})}\right] \tag{93}$$

$$:= \text{UB}\left(\log \hat{\mathcal{Z}}_{\text{GIWAE}}(K, \mathbf{x})\right) \tag{94}$$

As $K \to \infty$, we have $\frac{1}{K} \to 0$, $\frac{K-1}{K} \to 1$, and $\frac{1}{K-1}\sum_{k=2}^{K} e^{T_\phi(\mathbf{x},\mathbf{z}^{(k)})} \to \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right]$. Finally,

$$\lim_{K\to\infty} \text{UB}\left(\log \hat{\mathcal{Z}}_{\text{GIWAE}}(K, \mathbf{x})\right) = \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right] = \log \mathcal{Z}_{\theta,\phi}(\mathbf{x}) \tag{95}$$

As reasoned above, the convergence of both the upper and lower bound sequences to $\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})$ implies $\lim_{K\to\infty} \log \hat{\mathcal{Z}}_{\text{GIWAE}}(K, \mathbf{x}) = \log \mathcal{Z}_{\theta,\phi}(\mathbf{x})$ and $\lim_{K\to\infty} I_{\text{GIWAE}_L}(q_\theta, T_\phi) = \text{IBAL}(q_\theta, T_\phi)$, as desired. □

E.3 Convergence of GIWAE SNIS Distribution to MINE-AIS

Proposition E.2. *Let* ${q_{\text{PROP}}^{\text{GIWAE}}}_K(\mathbf{z}|\mathbf{x})$ *denote the marginal* SNIS *distribution of* GIWAE

$$q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}|\mathbf{x}; K) = \sum_{s=1}^{K} \int q_{\text{PROP}}^{\text{GIWAE}}(s, \mathbf{z}^{(1:K)}|\mathbf{x}) \delta(\mathbf{z} - \mathbf{z}^{(s)}) d\mathbf{z}^{(1:K)} \tag{96}$$

$$= \sum_{s=1}^{K} \int \delta(\mathbf{z} - \mathbf{z}^{(s)}) \left(\prod_{k=1}^{K} q(\mathbf{z}^{(k)}|\mathbf{x}) \right) \frac{e^{T_\phi(\mathbf{x}, \mathbf{z}^{(s)})}}{\sum_{k=1}^{K} e^{T_\phi(\mathbf{x}, \mathbf{z}^{(k)})}} d\mathbf{z}^{(1:K)} \tag{97}$$

Then as $K \to \infty$, the KL *divergence between the marginal* SNIS *distribution of the* GIWAE *and energy-based variational distribution of* MINE-AIS *goes to zero.*

$$\lim_{K\to\infty} D_{KL}[q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}|\mathbf{x}; K) || \pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})] = 0 \textit{ and } \lim_{K\to\infty} D_{KL}[\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x}) || q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}|\mathbf{x}; K)] = 0 \tag{98}$$

Proof. Note that the distribution $q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}|\mathbf{x}; K)$ is obtained by augmenting the standard (multi-sample) GIWAE proposal distribution $q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)}, s|\mathbf{x})$ using an additional variable $\mathbf{z} = \mathbf{z}^{(s)}$ which copies the sample in index $s \in [1, K]$. This index is chosen according to the SNIS weights induced by the energy function in Eq. (97), and we consider the distribution $q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}|\mathbf{x}; K)$ over $\mathbf{z}$ after marginalizing the joint $q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}, s, \mathbf{z}^{(1:K)})$ over $\mathbf{z}^{(1:K)}$ and s as in Eq. (96)-97.

To prove the proposition, we consider a mixture target distribution similar to the case of IWAE. However, in this case, we take a single sample from $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$ instead of $p(\mathbf{z}|\mathbf{x})$

$$p_{\text{TGT}}^{\pi_{\theta,\phi}}(\mathbf{z}, s, \mathbf{z}^{(1:K)}, \mathbf{x}) = \delta(\mathbf{z} - \mathbf{z}^{(s)}) \frac{1}{K} \pi_{\theta,\phi}(\mathbf{z}^{(s)}|\mathbf{x}) \prod_{k=1, k\neq s}^{K} q_\theta(\mathbf{z}^{(s)}|\mathbf{x}) \tag{99}$$

Note that marginalization over $\mathbf{z}$ will always yield a sample from $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$. Since marginalization can only reduce the KL divergence, we have

$$D_{KL}[q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}|\mathbf{x}; K) || \pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})] \leq D_{KL}[q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}, s, \mathbf{z}^{(1:K)}|\mathbf{x}) || p_{\text{TGT}}^{\pi_{\theta,\phi}}(\mathbf{z}, s, \mathbf{z}^{(1:K)}|\mathbf{x})] \tag{100}$$

$$= D_{KL}[q_{\text{PROP}}^{\text{GIWAE}}(s, \mathbf{z}^{(1:K)}|\mathbf{x}) || p_{\text{TGT}}^{\pi_{\theta,\phi}}(s, \mathbf{z}^{(1:K)}|\mathbf{x})] \tag{101}$$

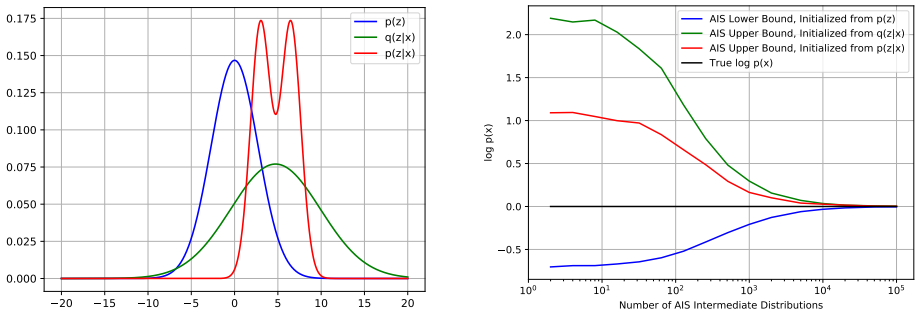
where the second line follows because $\mathbf{z}$ is a deterministic function of $\mathbf{z}^{(1:K)}$ and s. Now, we can show that this KL divergence goes to 0 as $K \to \infty$

$$D_{KL}[q_{\text{PROP}}^{\text{GIWAE}}(s, \mathbf{z}^{(1:K)}|\mathbf{x}) || p_{\text{TGT}}^{\pi_{\theta,\phi}}(s, \mathbf{z}^{(1:K)}|\mathbf{x})] = \mathbb{E}_{q_{\text{PROP}}^{\text{GIWAE}}} \left[\log \frac{\prod_{k=1}^{K} q(\mathbf{z}^{(k)}|\mathbf{x}) \frac{e^{T_\phi(\mathbf{x}, \mathbf{z}^{(s)})}}{\sum_{k=1}^{K} e^{T_\phi(\mathbf{x}, \mathbf{z}^{(k)})}}}{\frac{1}{K} \pi_{\theta,\phi}(\mathbf{z}^{(s)}|\mathbf{x}) \prod_{k=1, k\neq s}^{K} q(\mathbf{z}^{(k)}|\mathbf{x})} \right]$$

$$= \mathbb{E}_{q_{\text{PROP}}^{\text{GIWAE}}} \left[\log \frac{q(\mathbf{z}^{(s)}|\mathbf{x}) e^{T_\phi(\mathbf{x}, \mathbf{z}^{(s)})}}{q(\mathbf{z}^{(s)}|\mathbf{x}) e^{T_\phi(\mathbf{x}, \mathbf{z}^{(s)})}} \frac{\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})}{\frac{1}{K} \sum_{k=1}^{K} e^{T_\phi(\mathbf{x}, \mathbf{z}^{(k)})}} \right]$$

$$= \log \mathcal{Z}_{\theta,\phi}(\mathbf{x}) - \mathbb{E}_{q_{\text{PROP}}^{\text{GIWAE}}} \left[\log \frac{1}{K} \sum_{k=1}^{K} e^{T_\phi(\mathbf{x}, \mathbf{z}^{(k)})} \right] \tag{102}$$

Since $q_{\text{PROP}}^{\text{GIWAE}}(\mathbf{z}^{(1:K)}|\mathbf{x}) = \prod_{k=1}^{K} q_\theta(\mathbf{z}^{(k)}|\mathbf{x})$, it is clear that in the limit as $K \to \infty$, $\mathbb{E}_{q_{\text{PROP}}^{\text{GIWAE}}} \left[\log \frac{1}{K} \sum_{k=1}^{K} e^{T_\phi(\mathbf{x}, \mathbf{z}^{(k)})} \right] \to \log \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})} \left[e^{T_\phi(\mathbf{x}, \mathbf{z}^{(k)})} \right] = \log \mathcal{Z}_{\theta,\phi}(\mathbf{x})$. Thus, the difference



(a) KDE Plot (b) AIS Evaluation

Figure 6: Toy Example for changing the sampling distribution in MINE-AIS

in Eq. 102, the extended state space KL divergence in Eq. 101, and the marginal SNIS KL divergence in Eq. 100 all go to 0 as $K \to \infty$. As desired, $\lim_{K\to\infty} D_{KL}[q^{\text{GIWAE}}_{\text{PROP}}(\mathbf{z}|\mathbf{x}; K)||\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})] = 0$.

Similar reasoning as from Eq. 100-Eq. 102 holds for the KL divergence in the reverse direction, with $D_{KL}[p^{\pi_{\theta,\phi}}_{\text{TGT}}(s, \mathbf{z}^{(1:K)}|\mathbf{x})||q^{\text{GIWAE}}_{\text{PROP}}(s, \mathbf{z}^{(1:K)}|\mathbf{x})] = \mathbb{E}_{p^{\pi_{\theta,\phi}}_{\text{TGT}}}\left[\log \frac{1}{K}\sum_{k=1}^{K} e^{T_\phi(\mathbf{x},\mathbf{z}^{(k)})}\right] - \log \mathcal{Z}_{\theta,\phi}(\mathbf{x})$. The left term matches $\log \hat{\mathcal{Z}}_{\text{GIWAE}}(K, \mathbf{x})$ in Eq. 83 and the proof of Prop. E.3, apart from the fact that the 'positive' sample comes from $\pi_{\theta,\phi}$ instead of role of $p(\mathbf{z}|\mathbf{x})$. Using identical reasoning as above, the effect of this sample will vanish as $K \to \infty$. Thus, $\mathbb{E}_{p^{\pi_{\theta,\phi}}_{\text{TGT}}}\left[\log \frac{1}{K}\sum_{k=1}^{K} e^{T_\phi(\mathbf{x},\mathbf{z}^{(k)})}\right] \to \log \mathcal{Z}_{\theta,\phi}(\mathbf{x})$. Finally, this implies $\lim_{K\to\infty} D_{KL}[p^{\pi_{\theta,\phi}}_{\text{TGT}}(s, \mathbf{z}^{(1:K)}|\mathbf{x})||q^{\text{GIWAE}}_{\text{PROP}}(s, \mathbf{z}^{(1:K)}|\mathbf{x})] = 0$ and thus $\lim_{K\to\infty} D_{KL}[\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})||q^{\text{GIWAE}}_{\text{PROP}}(\mathbf{z}|\mathbf{x}; K)] = 0$ as desired.

□

E.4 Details for Training MINE-AIS

Eq. 24 indicates that in order to maximize the IBAL as a function of θ, ϕ, we need to increase the value of the energy function T_ϕ or score function $\log q_\theta(\mathbf{z}|\mathbf{x})$ on the *real* samples of the true joint $p(\mathbf{x}, \mathbf{z})$ and decrease the value on *fake* samples from the approximate $p(\mathbf{x})\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$. However, as is common in training energy-based models, it is difficult to draw samples from $\pi_{\theta,\phi}$.

In order to sample from $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$, a natural approach is to initialize MCMC chains from a sample of the base distribution $q_\theta(\mathbf{z}|\mathbf{x})$, using HMC transition kernels (Neal, 2011), for example. However, we may require infeasibly long MCMC chains when the base distribution is far from desired energy-based model $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$. Instead, we can choose to initialize the chains from the true posterior sample $\mathbf{z} \sim p(\mathbf{z}|\mathbf{x})$ for a simulated data point $\mathbf{x} \sim p(\mathbf{x})$. This initialization greatly reduces the computational cost and variance of the estimated gradient and enables training energy functions in high dimensional latent spaces, as shown in our experiments. This approach is in spirit similar to Contrastive Divergence learning of energy-based models (Hinton, 2002) where, in order to sample from the model distribution, we start an MCMC chain from the true data distribution to obtain a lower variance gradient estimate.

We use $T = 10$ MCMC steps in our experiments. After obtaining an approximate sample from $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$, we simply update the parameters θ using the gradients in Eq. 24.

E.5 Approximate Reverse Annealing for Evaluating MINE-AIS

Here, we argue that we can maintain an approximate upper bound on $\mathcal{Z}_{\theta^*,\phi^*}(\mathbf{x})$ and lower bound on IBAL by sampling from $p(\mathbf{z}|\mathbf{x})$ instead of $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$ in reverse importance sampling or reverse AIS.

Since these samples are unavailable, Prop. E.3 argues that we can preserve an approximate upper bound on $\log \mathcal{Z}_\theta(\mathbf{x})$ and lower bound on $\text{IBAL}(q_\theta, T_\phi)$ by sampling from $p(\mathbf{z}|\mathbf{x})$ instead of $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$ in reverse importance sampling.

Proposition E.3. *Suppose $T_\phi(\mathbf{x}, \mathbf{z})$ is a neural network parameterized θ and $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x}) = \frac{1}{\mathcal{Z}_{\theta,\phi}(\mathbf{x})} q_\theta(\mathbf{z}|\mathbf{x}) e^{T_\phi(\mathbf{x},\mathbf{z})}$. Suppose $\pi_{\theta^*,\phi^*}(\mathbf{z}|\mathbf{x})$ is the optimal* MINE-AIS *solution, within a variational family for $q_\theta(\mathbf{z}|\mathbf{x})$, and T_ϕ parameterized by a neural network. Then, we would have:*

$$\mathbb{E}_{p(\mathbf{z}|\mathbf{x})}\left[\log \frac{\pi_{\theta^*,\phi^*}(\mathbf{x}, \mathbf{z})}{q_{\theta^*}(\mathbf{z}|\mathbf{x})}\right] \geq \log \mathcal{Z}_{\theta^*,\phi^*}(\mathbf{x}) \tag{103}$$

Proof. Note that according to Prop. E.1(b), MINE-AIS optimal solution $\pi_{\theta^*,\phi^*}(\mathbf{z}|\mathbf{x})$ within a variational family improves upon the BA bound. This means, the gap of the MINE-AIS bound $D_{\text{KL}}[p(\mathbf{z}|\mathbf{x})||\pi_{\theta^*,\phi^*}(\mathbf{z}|\mathbf{x})]$ is less that the gap of the BA bound $D_{\text{KL}}[p(\mathbf{z}|\mathbf{x})||q_{\theta^*}(\mathbf{z}|\mathbf{x})]$. Thus we have:

$$\begin{aligned}\mathbb{E}_{p(\mathbf{z}|\mathbf{x})}\left[\log \frac{\pi_{\theta^*,\phi^*}(\mathbf{x}, \mathbf{z})}{q_{\theta^*}(\mathbf{z}|\mathbf{x})}\right] &= D_{\text{KL}}[p(\mathbf{z}|\mathbf{x})||q_{\theta^*}(\mathbf{z}|\mathbf{x})] - D_{\text{KL}}[p(\mathbf{z}|\mathbf{x})||\pi_{\theta^*,\phi^*}(\mathbf{z}|\mathbf{x})] + \log \mathcal{Z}_{\theta^*,\phi^*}(\mathbf{x}) \\ &\geq \log \mathcal{Z}_{\theta^*,\phi^*}(\mathbf{x}) \end{aligned} \tag{104}$$

□

As an illustrative example, Fig. 6a shows a Gaussian base distribution $q_{\theta^*}(\mathbf{z}|\mathbf{x})$, a multi modal true posterior distribution $p(\mathbf{z}|\mathbf{x})$, a posterior approximation $\pi_{\theta^*,\phi^*}(\mathbf{z}|\mathbf{x})$, whose energy function T_ϕ parameterized by μ_ϕ, σ_ϕ has been fit to minimize $D_{\text{KL}}[p(\mathbf{z}|\mathbf{x})||\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})]$. The standard reverse importance sampling, which take true samples from $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$, results in an upper bound on $\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})$. This is because the initial distribution $q_\theta(\mathbf{z}|\mathbf{x})$ is missing some of the modes of the target distribution $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$, and when we sample from $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$, particles from these modes will yield large importance weights since they have low density under $q_\theta(\mathbf{z}|\mathbf{x})$, resulting in an upper bound on $\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})$. Since MINE-AIS optimizes the reverse KL divergence, the learnt $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$ will conservatively allocate mass across the support of $p(\mathbf{z}|\mathbf{x})$. Prop. E.3 shows that sampling from $p(\mathbf{z}|\mathbf{x})$ instead of $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$ will still yield large importance weights and preserve the upper bound of reverse importance sampling.

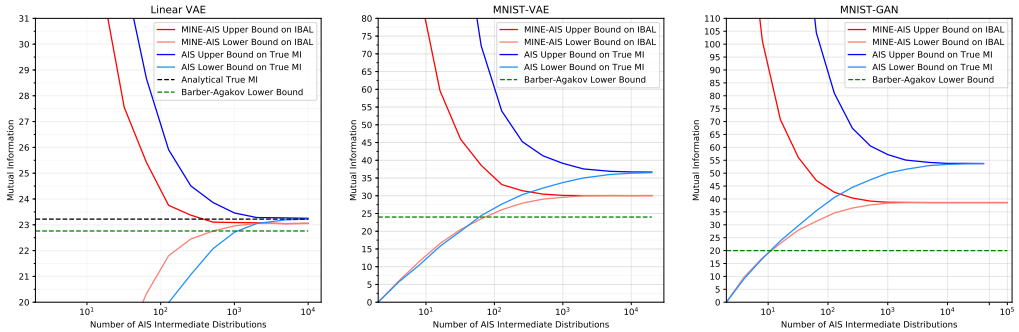
While Prop. E.3 shows that we can obtain a single-sample upper bound on $\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})$, these mathematical guarantees do not immediately translate to Multi-Sample AIS upper bounds, which includes additional transitions. However, we empirically observe in our experiments that this procedure does provide an upper bound on $\log \mathcal{Z}_\theta(\mathbf{x})$. For example, Fig. 6b shows the performance of the forward AIS starting from $q_\theta(\mathbf{z}|\mathbf{x})$ and reverse AIS starting from both $\pi_{\theta,\phi}(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z}|\mathbf{x})$. As expected, the forward AIS provides a lower-bound on $\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})$. We can also see that both reverse AIS chains provide an upper bound on the $\log \mathcal{Z}_{\theta,\phi}(\mathbf{x})$.

E.6 Multi-Sample AIS Convergence Results for MINE-AIS

In Fig. 7, we compare the performance of our MINE-AIS method, with its learned energy function T_ϕ, and evaluates Multi-Sample AIS lower and upper bounds on $\text{IBAL}(q_\theta, T_\phi) \leq I(\mathbf{x}; \mathbf{z})$ in cases where $p(\mathbf{x}|\mathbf{z})$ is unknown. As a comparison, we consider Multi-Sample AIS lower and upper bounds with known $p(\mathbf{x}|\mathbf{z})$, since this provides bounds on MI and represents the optimal energy function in MINE-AIS (Prop. E.3).

We also compare the convergence of the Multi-Sample AIS evaluation procedure as a function of T, by plotting our lower and upper bounds as a function of the number of intermediate T used for evaluation of $\text{IBAL}(p(\mathbf{z}), T_\phi)$. While the upper bound on IBAL may not actually lower bound the true MI for low T, we see that our bounds converge to the same estimate for high T.

At convergence, we can compare AIS estimates of MI and the IBAL for MINE-AIS to indicate the quality of the learned energy function. On the other hand, since the IBAL corresponds to the BA lower bound for a flexible, energy-based posterior approximation, we also compare against the BA



(a) Linear VAE (b) MNIST VAE (c) MNIST GAN

Figure 7: Estimating the mutual information of deep generative models with AIS and MINE-AIS.

lower bound for $q_\theta(\mathbf{z}|\mathbf{x})$ within a Gaussian variational family. We observe that IBAL lower bounds on MI fall between these two baselines, with MINE-AIS underestimating the value obtained from Multi-Sample AIS by 1%, 18%, and 28% on linear VAE, MNIST-VAE, and MNIST-GAN, respectively.

F Applications to Mutual Information Estimation without Known Marginals

While the focus in of our work is evaluating the mutual information $I(\mathbf{x};\mathbf{z})$ in settings where at least a single marginal is available, we are often interested in estimating or optimizing the mutual information where no marginal distribution is available. A natural setting where no marginal distribution is available is representation learning where the goal is to maximize the mutual information between the data distribution $q_{\text{data}}(\mathbf{x})$ and the representation induced by a stochastic mapping $q_\psi(\mathbf{z}|\mathbf{x})$ parameterized by ψ,

$$I(\mathbf{x};\mathbf{z}) = \mathbb{E}_{q_{\text{data}}(\mathbf{x})q_\psi(\mathbf{z}|\mathbf{x})}\left[\log \frac{q_\psi(\mathbf{x},\mathbf{z})}{q_{\text{data}}(\mathbf{x})q_\psi(\mathbf{z})}\right] = \mathbb{E}_{q_{\text{data}}(\mathbf{x})q_\psi(\mathbf{z}|\mathbf{x})}\left[\log \frac{q_\psi(\mathbf{x}|\mathbf{z})}{q_{\text{data}}(\mathbf{x})}\right] \tag{105}$$

where $q_\psi(\mathbf{z}) = \int q_{\text{data}}(\mathbf{x})q_\psi(\mathbf{z}|\mathbf{x})d\mathbf{x}$ represents the 'aggregated posterior' (Makhzani et al., 2015) or induced marginal distribution over $\mathbf{z}$.

F.1 BA Lower Bound

Using the notation of Eq. 105, we can write the BA lower bound as

$$\begin{aligned} I(\mathbf{x};\mathbf{z}) \geq I_{\text{BA}_L}(p_\theta(\mathbf{x}|\mathbf{z})) :&= I(\mathbf{x};\mathbf{z}) - \underbrace{\mathbb{E}_{q_\psi(\mathbf{z})}\left[D_{\text{KL}}[q_\psi(\mathbf{x}|\mathbf{z})||p_\theta(\mathbf{x}|\mathbf{z})]\right]}_{\text{gap}} \\ &= \mathbb{E}_{q_{\text{data}}(\mathbf{x})}\mathbb{E}_{q_\psi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}|\mathbf{z})}{q_{\text{data}}(\mathbf{x})}\right] \qquad (106) \\ &= \mathbb{E}_{q_{\text{data}}(\mathbf{x})}\underbrace{\mathbb{E}_{q_\psi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{negative reconstruction loss of } \mathbf{x}} + \underbrace{H_{\text{data}}(\mathbf{x})}_{\text{constant}} \qquad (107) \end{aligned}$$

where $p_\theta(\mathbf{x}|\mathbf{z})$ is a variational distribution, parameterized by θ, that tries to match to the *inverse encoding distribution* $q_\psi(\mathbf{x}|\mathbf{z}) \propto q_{\text{data}}(\mathbf{x})q_\psi(\mathbf{z}|\mathbf{x})$. The first term in Eq. 107 can be interpreted as the reconstruction term, and the second term is the entropy of the data distribution, which is constant.

Evaluating MI up to a Constant Note that the gradient of $I_{\text{BA}_L}(p_\theta(\mathbf{x}|\mathbf{z}))$ with respect to the parameters θ does not depend on the marginal $q_{\text{data}}(\mathbf{x})$, and thus we may still optimize the variational distribution when the data distribution is unknown. We can then use the resulting $p_\theta(\mathbf{x}|\mathbf{z})$ to estimate

the BA lower bound on mutual information *up to a constant*, $H_{\text{data}}(\mathbf{x})$. This is useful in comparing the MI induced by two different representations $q_{\psi_1}(\mathbf{z}|\mathbf{x})$ and $q_{\psi_2}(\mathbf{z}|\mathbf{x})$ of the same data distribution.

Optimizing MI with the BA Lower Bound The BA lower bound is also amenable to backpropagation through the parameters of the encoding distribution $q_\psi(\mathbf{z}|\mathbf{x})$, even when analytic marginal densities for $q_{\text{data}}(\mathbf{x})$ or $q_\psi(\mathbf{z})$ are not available.

Maximization of the BA lower bound appears in various settings, including in representation learning (Alemi et al., 2016; 2018), reinforcement learning (Mohamed & Rezende, 2015), and improving interpretability in GANs (Chen et al., 2016). In particular, in the information bottleneck method (Tishby et al., 2000), we are interested in maximizing the 'relevant information' $I(\mathbf{z};\mathbf{y})$ between a learned representation and a target variable $\mathbf{y}$. In supervised learning, $\mathbf{y}$ may represent classification labels (Alemi et al., 2016). When $\mathbf{y}$ is the original data as in unsupervised learning with β-VAEs (Alemi et al., 2018), the distortion or reconstruction loss corresponds to maximizing the BA lower bound on the encoding mutual information in Eq. (105).

F.2 GIWAE Lower Bound

In this section, we discuss the applicability of our GIWAE lower bound in mutual information maximization settings.

Rewriting the GIWAE lower bound in Eq. (9) for the setting of estimating $I(\mathbf{x};\mathbf{z})$ in Eq. (105), we have

$$I(\mathbf{x};\mathbf{z}) \geq I_{\text{GIWAE}_L}(p_\theta(\mathbf{x}|\mathbf{z}), T_\phi, K) = \mathbb{E}_{q_{\text{data}}(\mathbf{x})} \underbrace{\mathbb{E}_{q_\psi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big]}_{\text{negative reconstruction loss of } \mathbf{x}} + \underbrace{H_{\text{data}}(\mathbf{x})}_{\text{constant}} \tag{108}$$
$$+ \underbrace{\mathbb{E}_{q_\psi(\mathbf{x},\mathbf{z})p_\theta(\mathbf{x}^{(2:K)}|\mathbf{z})}\left[\log \frac{e^{T_\phi(\mathbf{x},\mathbf{z})}}{\frac{1}{K}\left(e^{T_\phi(\mathbf{x},\mathbf{z})} + \sum_{k=2}^{K} e^{T_\phi(\mathbf{x}^{(k)},\mathbf{z})}\right)}\right]}_{\text{contrastive term} \leq \log K}.$$

Note that we can use a single joint sample from $q_{\text{data}}(\mathbf{x})q_\psi(\mathbf{z}|\mathbf{x}) = q_\psi(\mathbf{z})q_\psi(\mathbf{x}|\mathbf{z})$ to obtain a positive sample from the inverse encoding distribution $\mathbf{x} \sim q_\psi(\mathbf{x}|\mathbf{z})$ for a particular $\mathbf{z} \sim q_\psi(\mathbf{z})$, in similar fashion to our ancestral sampling in Sec. 1.1. For a given $\mathbf{z}$, the negative samples correspond to $K-1$ samples from the stochastic decoder $\mathbf{x}^{(2:K)} \sim p_\theta(\mathbf{x}|\mathbf{z})$.

From the importance sampling perspective, we can view the GIWAE lower bound as corresponding to an upper bound on $\log q_\psi(\mathbf{z})$ for a particular $\mathbf{z} \sim q_\psi(\mathbf{z})$. The contrastive term in Eq. (108) arises from SNIS sampling of a single $\mathbf{x}^{(s)}$ in the extended state space proposal, with $q_{\text{PROP}}^{\text{GIWAE}}(s|\mathbf{z},\mathbf{x}^{1:K}) = e^{T_\phi(\mathbf{x}^{(s)},\mathbf{z})} / \sum_{k=1}^{K} e^{T_\phi(\mathbf{x}^{(k)},\mathbf{z})}$.

Finally, we note that optimization over the energy function $T_\phi(\mathbf{x},\mathbf{z})$ improves upon $I_{\text{BA}_L}(p_\theta(\mathbf{x}|\mathbf{z}))$ in Eq. (107) by at most $\log K$ nats using the contrastive term.

Evaluating MI up to a Constant Similar to the BA bound, the GIWAE lower bound can be used to evaluate the mutual information up to the constant data entropy when $q_{\text{data}}(\mathbf{x})$ is unknown. This may be useful in comparing the MI induced by two different representations $q_{\psi_1}(\mathbf{z}|\mathbf{x})$ and $q_{\psi_2}(\mathbf{z}|\mathbf{x})$ of the same data distribution. Note that the reconstruction and contrastive terms in Eq. (108) depend only on samples from $q_{\text{data}}(\mathbf{x})$, so that our ability to take gradients with respect to the parameters of the variational distribution $p_\theta(\mathbf{x}|\mathbf{z})$ or energy function $T_\phi(\mathbf{x},\mathbf{z})$ are not affected by the fact that the marginal distribution is unknown.

Optimizing MI with the GIWAE Lower Bound The GIWAE lower bound may also be used for mutual information maximization with respect to the parameters of $q_\psi(\mathbf{z}|\mathbf{x})$, since each term in Eq. (108) is amenable to backpropagation. Since GIWAE generalizes both the BA and INFO-NCE lower bounds, it can be used as a drop-in replacement for either of these bounds for optimizing MI.

Contrastive INFO-NCE bounds appear in representation learning across a variety of settings and modalities, including images (Hjelm et al., 2018; van den Oord et al., 2018; Bachman et al., 2019;

Chen et al., 2020), sequential data (van den Oord et al., 2018), reinforcement learning (Mazoure et al., 2020; van den Oord et al., 2018), or enforcing fairness constraints (Gupta et al., 2021).

F.3 MINE-AIS / IBAL LOWER BOUND

Recall that MINE-AIS estimation would involve an energy based variational approximation to the inverse encoding distribution $q_\psi(\mathbf{x}|\mathbf{z})$,

$$\pi_{\theta,\phi}(\mathbf{x}|\mathbf{z}) = \frac{1}{\mathcal{Z}(\mathbf{z})} p_\theta(\mathbf{x}|\mathbf{z}) e^{T_\phi(\mathbf{x},\mathbf{z})}. \tag{109}$$

Note that IBAL lower bound on $I(\mathbf{x};\mathbf{z})$ involves an intractable log partition function term $\mathbb{E}_{q_\psi(\mathbf{z})}[\log \mathcal{Z}(\mathbf{z})] = \mathbb{E}_{q_\psi(\mathbf{z})}\left[\log \mathbb{E}_{p_\theta(\mathbf{x}|\mathbf{z})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right]\right]$.

$$\begin{aligned} I(\mathbf{x};\mathbf{z}) \geq \text{IBAL}(p_\theta(\mathbf{x}|\mathbf{z}), T_\phi) = \mathbb{E}_{q_{\text{data}}(\mathbf{x})} & \underbrace{\mathbb{E}_{q_\psi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{negative reconstruction loss of } \mathbf{x}} + \underbrace{H_{\text{data}}(\mathbf{x})}_{\text{constant}} \\ & + \underbrace{\mathbb{E}_{q_\psi(\mathbf{x},\mathbf{z})}\left[\log \frac{e^{T_\phi(\mathbf{x},\mathbf{z})}}{\mathbb{E}_{p_\theta(\mathbf{x}|\mathbf{z})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right]}\right]}_{\text{contrastive term} \leq \mathbb{E}_{q_\psi(\mathbf{z})}[D_{\text{KL}}[q_\psi(\mathbf{x}|\mathbf{z})||p_\theta(\mathbf{x}|\mathbf{z})]]} \end{aligned} \tag{110}$$

where the term in the denominator is the partition function $\mathcal{Z}(\mathbf{z}) = \mathbb{E}_{p_\theta(\mathbf{x}|\mathbf{z})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right]$.

When taking gradients as in Eq. (24)-(25), we obtain

$$\frac{\partial}{\partial\theta}\text{IBAL}(p_\theta, T_\phi) = \mathbb{E}_{q_{\text{data}}(\mathbf{x})q_\psi(\mathbf{z}|\mathbf{x})}\left[\frac{\partial}{\partial\theta}\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \mathbb{E}_{q_\psi(\mathbf{z})\pi_{\theta,\phi}(\mathbf{x}|\mathbf{z})}\left[\frac{\partial}{\partial\theta}\log p_\theta(\mathbf{x}|\mathbf{z})\right]. \tag{111}$$

$$\frac{\partial}{\partial\phi}\text{IBAL}(p_\theta, T_\phi) = \mathbb{E}_{q_{\text{data}}(\mathbf{x})q_\psi(\mathbf{z}|\mathbf{x})}\left[\frac{\partial}{\partial\phi}T_\phi(\mathbf{x},\mathbf{z})\right] - \mathbb{E}_{q_\psi(\mathbf{z})\pi_{\theta,\phi}(\mathbf{x}|\mathbf{z})}\left[\frac{\partial}{\partial\phi}T_\phi(\mathbf{x},\mathbf{z})\right]. \tag{112}$$

To obtain approximate negative samples for a given $\mathbf{z}$, we can use MCMC transition kernels which only require access to the unnormalized target density $p_\theta(\mathbf{x}|\mathbf{z})e^{T_\phi(\mathbf{x},\mathbf{z})}$. This can be evaluated directly since both $p_\theta(\mathbf{x}|\mathbf{z})$ and $T_\phi(\mathbf{x},\mathbf{z})$ are tractable.

Evaluating MI up to a Constant Since only samples from $q_{\text{data}}(\mathbf{x})$ are required for the MINE-AIS training procedure in Eq. (111)-(112), we can learn the base variational distribution $p_\theta(\mathbf{x}|\mathbf{z})$ and energy function $T_\phi(\mathbf{x},\mathbf{z})$ in cases when the marginal distribution is unknown.

As in the case of GIWAE and BA lower bounds above, we can also evaluate the IBAL lower bound in Eq. (110) up to a constant. As in the main text, this involves using multi-sample AIS techniques to bound the intractable log partition function term $\log \mathcal{Z}(\mathbf{z}) = \log \mathbb{E}_{p_\theta(\mathbf{x}|\mathbf{z})}\left[e^{T_\phi(\mathbf{x},\mathbf{z})}\right]$.

Optimizing MI with the MINE-AIS Lower Bound If we are interested in optimizing the MINE-AIS lower bound with respect to the parameters of a stochastic encoder $q_\psi(\mathbf{z}|\mathbf{x})$, we would need to backpropagate through the multi-sample AIS procedure used to estimate the $\log \mathcal{Z}(\mathbf{z})$ term.

We thus conclude that, among our proposed methods, GIWAE is the most directly applicable in settings of mutual information maximization for representation learning.

G BACKGROUND ON AIS

AIS (Neal, 2001) constructs a sequence of intermediate distributions $\{\pi_t(\mathbf{z})\}_{t=0}^T$, which bridge between a normalized initial distribution $\pi_0(\mathbf{z}|\mathbf{x})$ and target distribution $\pi_T(\mathbf{z}|\mathbf{x}) = p(\mathbf{z}|\mathbf{x})$ with unnormalized density $\pi_T(\mathbf{x},\mathbf{z}) = p(\mathbf{x},\mathbf{z})$ and normalizing constant $\mathcal{Z}_T(\mathbf{x}) = p(\mathbf{x})$. A common choice for intermediate distributions is the geometric mixture path parameterized by $\{\beta_t\}_{t=0}^T$:

$$\pi_t(\mathbf{z}|\mathbf{x}) = \frac{\pi_0(\mathbf{z}|\mathbf{x})^{1-\beta_t}\,\pi_T(\mathbf{x},\mathbf{z})^{\beta_t}}{\mathcal{Z}_t(\mathbf{x})} \quad \text{where} \quad \mathcal{Z}_t(\mathbf{x}) = \int \pi_0(\mathbf{z}|\mathbf{x})^{1-\beta_t}\,\pi_T(\mathbf{x},\mathbf{z})^{\beta_t} d\mathbf{z}. \tag{113}$$

AIS involves a extended state space proposal distribution $q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}_{0:T}|\mathbf{x})$, obtained by sampling from the initial $\pi_0(\mathbf{z}|\mathbf{x})$ and constructing transitions $\mathcal{T}_f(\mathbf{z}_{t+1}|\mathbf{z}_t, \mathbf{x})$ which leave $\pi_t(\mathbf{z}|\mathbf{x})$ invariant. The target distribution $p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}_{0:T}|\mathbf{x})$ is given by running the reverse transitions $\mathcal{T}_r(\mathbf{z}_t|\mathbf{z}_{t+1}, \mathbf{x})$ starting from a target or posterior sample $\pi_T(\mathbf{z}|\mathbf{x})$, as shown in Fig. 2.

$$q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}_{0:T}|\mathbf{x}) = \pi_0(\mathbf{z}_0|\mathbf{x}) \prod_{t=0}^{T-1} \mathcal{T}_f(\mathbf{z}_{t+1}|\mathbf{z}_t) \qquad p^{\text{AIS}}_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{0:T}) = \pi_T(\mathbf{x}, \mathbf{z}_T) \prod_{t=0}^{T-1} \mathcal{T}_r(\mathbf{z}_t|\mathbf{z}_{t+1}, \mathbf{x}) \tag{114}$$

Since the forward and reverse kernels obey $\mathcal{T}_f(\mathbf{z}_{t+1}|\mathbf{z}_t, \mathbf{x})\pi_t(\mathbf{z}_t|\mathbf{x}) = \mathcal{T}_r(\mathbf{z}_t|\mathbf{z}_{t+1}, \mathbf{x})\pi_t(\mathbf{z}_{t+1}|\mathbf{x})$, we can write the importance weights $w := p^{\text{AIS}}_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{0:T})/q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}_{0:T}|\mathbf{x})$ in the extended state space as

$$w = \frac{p^{\text{AIS}}_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{0:T})}{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}_{0:T}|\mathbf{x})} = \frac{\pi_T(\mathbf{x}, \mathbf{z}_T)}{\pi_0(\mathbf{z}_0|\mathbf{x})} \prod_{t=0}^{T-1} \frac{\pi_t(\mathbf{x}, \mathbf{z}_t)}{\pi_t(\mathbf{x}, \mathbf{z}_{t+1})} = \prod_{t=1}^{T} \frac{\pi_t(\mathbf{x}, \mathbf{z}_t)}{\pi_{t-1}(\mathbf{x}, \mathbf{z}_t)} = \prod_{t=1}^{T} \left(\frac{\pi_T(\mathbf{x}, \mathbf{z}_t)}{\pi_0(\mathbf{z}_t|\mathbf{x})} \right)^{\beta_t - \beta_{t-1}} \tag{115}$$

Assuming the initial distribution is normalized $\mathcal{Z}_0 = 1$, AIS provides the following estimate of the partition function $\hat{\mathcal{Z}}_T(\mathbf{x})$

$$\hat{\mathcal{Z}}_T(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^{K} w^{(k)} \qquad \text{where} \quad w^{(k)} = \frac{p^{\text{AIS}}_{\text{TGT}}(\mathbf{x}, \mathbf{z}^{(k)}_{0:T})}{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(k)}_{0:T}|\mathbf{x})} \quad \text{and} \quad \mathbf{z}^{(k)}_{0:T} \sim q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}_{0:T}|\mathbf{x}), \tag{116}$$

which is unbiased since $\mathbb{E}[\hat{\mathcal{Z}}_T(\mathbf{x})] = \mathbb{E}_{\mathbf{z}_{0:T} \sim q^{\text{AIS}}_{\text{PROP}}}[w] = \mathcal{Z}_T(\mathbf{x})$.

Using the general approach in App. A, taking expectations of the log importance weights under the proposal and target yields the following lower and upper bounds on the log partition function.

$$\underbrace{\mathbb{E}_{\mathbf{z}_{0:T} \sim q^{\text{AIS}}_{\text{PROP}}} \left[\log \frac{p^{\text{AIS}}_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{0:T})}{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}_{0:T}|\mathbf{x})} \right]}_{\text{ELBO}_{\text{AIS}}(\pi_0, T)} \leq \log p(\mathbf{x}) \leq \underbrace{\mathbb{E}_{\mathbf{z}_{0:T} \sim p^{\text{AIS}}_{\text{TGT}}} \left[\log \frac{p^{\text{AIS}}_{\text{TGT}}(\mathbf{x}, \mathbf{z}_{0:T})}{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}_{0:T}|\mathbf{x})} \right]}_{\text{EUBO}_{\text{AIS}}(\pi_0, T)} \tag{117}$$

The lower bound is an extended state-space ELBO using the proposal and target distributions in Eq. (14), with gap equal to $D_{KL}[q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}_{0:T}|\mathbf{x})||p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}_{0:T}|\mathbf{x})]$. The upper bound is an extended state space EUBO, whose gap is equal to $D_{KL}[p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}_{0:T}|\mathbf{x})||q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}_{0:T}|\mathbf{x})]$. These single-chain lower and upper bounds on $\log p(\mathbf{x})$ translate to upper and lower bounds on MI, $I_{\text{AIS}_L}(\pi_0, T)$ and $I_{\text{AIS}_U}(\pi_0, T)$, which were suggested for MI estimation in the blog post of Sobolev (2019a).

Similar single-sample bounds also appear in variational inference methods (Salimans et al., 2015; Masrani et al., 2019), and backpropagation through these estimators in generative modeling constitutes an active area of research (Geffner & Domke, 2021; Zhang et al., 2021).

H Independent Multi-Sample AIS

H.1 Probabilistic Interpretation and Bounds

We can expand the probabilistic interpretation of Independent Multi-Sample AIS from the main text Eq. 120 and Fig. 2 using notation for the full AIS chains

$$q^{\text{IM-AIS}}_{\text{PROP}}(\mathbf{z}^{(1:K)}_{0:T}|\mathbf{x}) = \prod_{k=1}^{K} q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(k)}_{0:T}|\mathbf{x}) \tag{118}$$

$$= \prod_{k=1}^{K} \pi_0(\mathbf{z}^{(k)}_0|\mathbf{x}) \prod_{t=0}^{T-1} \mathcal{T}_f(\mathbf{z}^{(k)}_{t+1}|\mathbf{z}^{(k)}_t) \tag{119}$$

$$p^{\text{IM-AIS}}_{\text{TGT}}(\mathbf{z}^{(1:K)}_{0:T}, \mathbf{x}) = \frac{1}{K} \sum_{s=1}^{K} p^{\text{AIS}}_{\text{TGT}}(\mathbf{x}, \mathbf{z}^{(s)}_{0:T}) \prod_{\substack{k=1 \\ k \neq s}}^{K} q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(k)}_{0:T}|\mathbf{x}). \tag{120}$$

$$= \frac{1}{K} \sum_{k=1}^{K} \pi_T(\mathbf{z}^{(k)}_T|\mathbf{x}) \prod_{t=1}^{T} \mathcal{T}_r(\mathbf{z}^{(k)}_{t-1}|\mathbf{z}^{(k)}_t) \cdot \prod_{\substack{j=1 \\ j \neq k}}^{K} \pi_0(\mathbf{z}^{(k)}_0|\mathbf{x}) \prod_{t=0}^{T-1} \mathcal{T}_f(\mathbf{z}^{(k)}_{t+1}|\mathbf{z}^{(k)}_t)$$

The importance ratio $\log p^{\text{IM-AIS}}_{\text{TGT}}(\mathbf{x},\mathbf{z})/q^{\text{IM-AIS}}_{\text{PROP}}(\mathbf{x},\mathbf{z})$ simplifies to a mixture as in Eq. 30, although we are now averaging the importance weights along AIS chains

$$\log \frac{p^{\text{IM-AIS}}_{\text{TGT}}(\mathbf{z}^{(1:K)}_{0:T},\mathbf{x})}{q^{\text{IM-AIS}}_{\text{PROP}}(\mathbf{z}^{(1:K)}_{0:T},\mathbf{x})} = \log \frac{1}{K}\sum_{k=1}^{K}\frac{p^{\text{AIS}}_{\text{TGT}}(\mathbf{x},\mathbf{z}^{(k)}_{0:T})}{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(k)}_{0:T}|\mathbf{x})} = \log \frac{1}{K}\sum_{k=1}^{K}\prod_{t=0}^{T-1}\left(\frac{p(\mathbf{x},\mathbf{z}^{(k)}_t)}{q(\mathbf{z}^{(k)}_t|\mathbf{x})}\right)^{\beta_{t+1}-\beta_t} = \log \frac{1}{K}\sum_{k=1}^{K} w^{(k)} \tag{121}$$

Taking the expected log ratio under the proposal and target, respectively, yield lower and upper bounds on $\log p(\mathbf{x})$

$$\underbrace{\mathbb{E}_{\mathbf{z}^{(1:K)}_{0:T}\sim q^{\text{AIS}}_{\text{PROP}}}\left[\log \frac{1}{K}\sum_{k=1}^{K}\frac{p^{\text{AIS}}_{\text{TGT}}(\mathbf{x},\mathbf{z}^{(k)}_{0:T})}{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(k)}_{0:T}|\mathbf{x})}\right]}_{\text{ELBO}_{\text{IM-AIS}}(\pi_0,K,T)} \leq \log p(\mathbf{x}) \leq \underbrace{\mathbb{E}_{\substack{\mathbf{z}^{(1)}_{0:T}\sim p^{\text{AIS}}_{\text{TGT}}\\ \mathbf{z}^{(2:K)}_{0:T}\sim q^{\text{AIS}}_{\text{PROP}}}}\left[\log \frac{1}{K}\sum_{k=1}^{K}\frac{p^{\text{AIS}}_{\text{TGT}}(\mathbf{x},\mathbf{z}^{(k)}_{0:T})}{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(k)}_{0:T}|\mathbf{x})}\right]}_{\text{EUBO}_{\text{IM-AIS}}(\pi_0,K,T)}. \tag{122}$$

H.2 Proof of Proposition 3.1

Proposition 3.1 (Complexity in K). *For a given T, the bias of the Independent Multi-Sample* AIS *lower bound on* MI*, $I_{\text{IM-AIS}_L}(\pi_0,T)$, decreases (at most) logarithmically in K.*

$$I_{\text{AIS}_L}(\pi_0,T) \leq I_{\text{IM-AIS}_L}(\pi_0,K,T) \leq I_{\text{AIS}_L}(\pi_0,T) + \log K \tag{17}$$

To prove the proposition, we first analyze the $\text{EUBO}_{\text{IM-AIS}}(\pi_0,K,T)$

$$\log p(\mathbf{x}) \leq \text{EUBO}_{\text{IM-AIS}}(\pi_0,K,T) = \mathbb{E}_{p^{\text{IM-AIS}}_{\text{TGT}}(\mathbf{z}^{(1)}_{0:T}|\mathbf{x})q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(2:K)}_{0:T}|\mathbf{x})}\left[\log \frac{1}{K}\sum_{i=1}^{K}\frac{p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}^{(k)}_{0:T},\mathbf{x})}{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(k)}_{0:T}|\mathbf{x})}\right] \tag{123}$$

$$= \mathbb{E}_{p^{\text{IM-AIS}}_{\text{TGT}}(\mathbf{z}^{(1)}_{0:T}|\mathbf{x})q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(2:K)}_{0:T}|\mathbf{x})}\left[\log \sum_{i=1}^{K}\frac{p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}^{(k)}_{0:T},\mathbf{x})}{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(k)}_{0:T}|\mathbf{x})}\right] - \log K$$

$$\geq \mathbb{E}_{p^{\text{IM-AIS}}_{\text{TGT}}(\mathbf{z}^{(1)}_{0:T}|\mathbf{x})}\left[\log \frac{p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}^{(k)}_{0:T},\mathbf{x})}{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(k)}_{0:T}|\mathbf{x})}\right] - \log K \tag{124}$$

$$= \text{EUBO}_{\text{AIS}}(T) - \log K \tag{125}$$

so that $\text{EUBO}_{\text{IM-AIS}}(\pi_0,K,T) \geq \text{EUBO}_{\text{AIS}}(T) - \log K$ and the inequality bounds the improvement of the upper bound on $\log p(\mathbf{x})$. We use the non-negativity of the likelihood ratio for the inequality in the third line. Since $I_{\text{IM-AIS}_L}(\pi_0,K,T) = \mathbb{E}_{p(\mathbf{x}|\mathbf{z})}[\log p(\mathbf{x}|\mathbf{z})] - \text{EUBO}_{\text{IM-AIS}}(\pi_0,K,T)$ and $I_{\text{AIS}_L}(\pi_0,T) = \mathbb{E}_{p(\mathbf{x}|\mathbf{z})}[\log p(\mathbf{x}|\mathbf{z})] - \text{EUBO}_{\text{AIS}}(\pi_0,T)$, we have $I_{\text{IM-AIS}_L}(\pi_0,K,T) \leq I_{\text{AIS}_L}(\pi_0,T) + \log K$ as desired. $I_{\text{AIS}_L}(\pi_0,T) \leq I_{\text{IM-AIS}_L}(\pi_0,K,T)$ follows from the fact that multi-sample bounds are tighter than their single sample counterparts.

H.3 Proof of Proposition 3.2

Proposition 3.2 (Complexity in T). *Assuming perfect transitions and a geometric annealing path with linearly-spaced $\{\beta_t\}_{t=1}^{T}$, the sum of the gaps in the multi-sample* AIS *sandwich bounds on* MI*, $\delta^{T,K} = I_{\text{IM-AIS}_U}(\pi_0,K,T) - I_{\text{IM-AIS}_L}(\pi_0,K,T)$, reduces linearly with increasing T.*

Proof. We now prove the equality, which expresses the gap between the single sample AIS upper and lower bound for linear scheduling as a $1/T$ factor multiplied by the symmetrized KL divergence. We

first simplify the difference in the single-sample upper and lower bounds as

$$
\begin{aligned}
\delta_L^{T,K=1} + \delta_U^{T,K=1} &= \text{EUBO}_{\text{AIS}}(T, K=1) - \text{ELBO}_{\text{AIS}}(T, K=1) \\
&= \mathbb{E}_{\mathbf{z}_{0:T} \sim p_{\text{TGT}}^{\text{AIS}}} \left[\log \frac{q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T})}{p_{\text{TGT}}^{\text{AIS}}(\mathbf{x}, \mathbf{z}_{0:T})} \right] - \mathbb{E}_{\mathbf{z}_{0:T} \sim q_{\text{PROP}}^{\text{AIS}}} \left[\log \frac{p_{\text{TGT}}^{\text{AIS}}(\mathbf{x}, \mathbf{z}_{0:T})}{q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T})} \right] \\
&= \mathbb{E}_{\mathbf{z}_{0:T} \sim p_{\text{TGT}}^{\text{AIS}}} \left[\log \frac{\pi_T(\mathbf{z}_T|\mathbf{x}) \prod_{t=0}^{T-1} \mathcal{T}_r(\mathbf{z}_t|\mathbf{z}_{t+1})}{\pi_0(\mathbf{z}_0|\mathbf{x}) \prod_{t=0}^{T-1} \mathcal{T}_f(\mathbf{z}_{t+1}|\mathbf{z}_t)} \right] - \mathbb{E}_{\mathbf{z}_{0:T} \sim q_{\text{PROP}}^{\text{AIS}}} \left[\log \frac{\pi_T(\mathbf{z}_T|\mathbf{x}) \prod_{t=0}^{T-1} \mathcal{T}_r(\mathbf{z}_t|\mathbf{z}_{t+1})}{\pi_0(\mathbf{z}_0|\mathbf{x}) \prod_{t=0}^{T-1} \mathcal{T}_f(\mathbf{z}_{t+1}|\mathbf{z}_t)} \right] \\
&= \mathbb{E}_{\mathbf{z}_{0:T} \sim p_{\text{TGT}}^{\text{AIS}}} \left[\log \prod_{t=1}^{T} \left(\frac{\tilde{\pi}_T(\mathbf{x}, \mathbf{z}_t)}{\tilde{\pi}_0(\mathbf{z}_t)} \right)^{\beta_t - \beta_{t-1}} \right] - \mathbb{E}_{\mathbf{z}_{0:T} \sim q_{\text{PROP}}^{\text{AIS}}} \left[\log \prod_{t=1}^{T} \left(\frac{\tilde{\pi}_T(\mathbf{x}, \mathbf{z}_t)}{\tilde{\pi}_0(\mathbf{z}_t)} \right)^{\beta_t - \beta_{t-1}} \right] \\
&= \mathbb{E}_{\mathbf{z}_{0:T} \sim p_{\text{TGT}}^{\text{AIS}}} \left[(\beta_t - \beta_{t-1}) \log \frac{\tilde{\pi}_T(\mathbf{x}, \mathbf{z}_t)}{\tilde{\pi}_0(\mathbf{z}_t|\mathbf{x})} \right] - \mathbb{E}_{\mathbf{z}_{0:T} \sim q_{\text{PROP}}^{\text{AIS}}} \left[(\beta_t - \beta_{t-1}) \log \frac{\tilde{\pi}_T(\mathbf{x}, \mathbf{z}_t)}{\tilde{\pi}_0(\mathbf{z}_t|\mathbf{x})} \right]
\end{aligned}
\tag{126}
$$

We now assume perfect transitions, which is common in analysis of AIS (Neal, 2001; Grosse et al., 2013). Since each forward transition $\mathcal{T}_f(\mathbf{z}_{t+1}|\mathbf{z}_t)$ leaves $\pi_t(\mathbf{z})$ invariant, perfect transitions results in independent exact samples from

$$
\mathbf{z}_{0:T}^{(1:K)} \sim q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(1:K)}|\mathbf{x}) \implies \mathbf{z}_{0:T}^{(1:K)} \sim \pi_0(\mathbf{z}_0) \prod_{t=1}^{T} \pi_{\beta_{t-1}}(\mathbf{z}_t) \tag{127}
$$

so that $\mathbf{z}_1 \sim \pi_0(\mathbf{z})$, $\mathbf{z}_2 \sim \pi_{\beta_1}(\mathbf{z})$, $\mathbf{z}_3 \sim \pi_{\beta_2}(\mathbf{z})$, etc. Similarly, perfect transitions in the reverse direction $\mathcal{T}_f(\mathbf{z}_t|\mathbf{z}_{t+1})$ results in exact samples from $\mathbf{z}_t \sim \pi_{\beta_t}(\mathbf{z})$

$$
\mathbf{z}_{0:T}^{(1:K)} \sim p_{\text{TGT}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(1:K)}|\mathbf{x}) \implies \mathbf{z}_{0:T}^{(1:K)} \sim \pi_T(\mathbf{z}_T) \prod_{t=1}^{T} \pi_{\beta_{t-1}}(\mathbf{z}_{t-1}) \tag{128}
$$

so that $\mathbf{z}_{T-1} \sim \pi_{\beta_{T-1}}(\mathbf{z})$ and so on down to $\mathbf{z}_0 \sim \pi_0(\mathbf{z})$.

To put this all together, we substitute back into Eq. 126 using a linear annealing schedule $\beta_t - \beta_{t-1} = \frac{1}{T}\ \forall t$. In particular, since Eq. 127 and Eq. 128 factorize, the expectations over the extended state space simplify to a sum of expectations at each $\mathbf{z}_t$, with $\mathbf{z}_t \sim \pi_{\beta_{t-1}}(\mathbf{z})$ for the forward chain and $\mathbf{z}_t \sim \pi_{\beta_t}(\mathbf{z})$ for the reverse chain

$$
\begin{aligned}
\delta_L^{T,K=1} + \delta_U^{T,K=1} &= \text{EUBO}_{\text{AIS}}(T, K=1) - \text{ELBO}_{\text{AIS}}(T, K=1) \\
&= \mathbb{E}_{\mathbf{z}_{0:T} \sim p_{\text{TGT}}^{\text{AIS}}} \left[(\beta_t - \beta_{t-1}) \log \frac{\tilde{\pi}_T(\mathbf{x}, \mathbf{z}_t)}{\tilde{\pi}_0(\mathbf{z}_t|\mathbf{x})} \right] - \mathbb{E}_{\mathbf{z}_{0:T} \sim q_{\text{PROP}}^{\text{AIS}}} \left[(\beta_t - \beta_{t-1}) \log \frac{\tilde{\pi}_T(\mathbf{x}, \mathbf{z}_t)}{\tilde{\pi}_0(\mathbf{z}_t|\mathbf{x})} \right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\pi_{\beta_t}(\mathbf{z})} \left[(\beta_t - \beta_{t-1}) \log \frac{\tilde{\pi}_T(\mathbf{x}, \mathbf{z})}{\tilde{\pi}_0(\mathbf{z}|\mathbf{x})} \right] - \sum_{t=1}^{T} \mathbb{E}_{\pi_{\beta_{t-1}}(\mathbf{z})} \left[(\beta_t - \beta_{t-1}) \log \frac{\tilde{\pi}_T(\mathbf{x}, \mathbf{z})}{\tilde{\pi}_0(\mathbf{z}|\mathbf{x})} \right] \\
&= \frac{1}{T} \mathbb{E}_{\pi_T(\mathbf{z})} \left[\log \frac{\tilde{\pi}_T(\mathbf{x}, \mathbf{z})}{\tilde{\pi}_0(\mathbf{z}|\mathbf{x})} \right] - \frac{1}{T} \mathbb{E}_{\pi_0(\mathbf{z})} \left[\log \frac{\tilde{\pi}_T(\mathbf{x}, \mathbf{z})}{\tilde{\pi}_0(\mathbf{z}|\mathbf{x})} \right] \\
&= \frac{1}{T} \left(D_{KL}(\pi_0 || \pi_T) + D_{KL}(\pi_T || \pi_0) \right).
\end{aligned}
$$

In moving from the third to the fourth line, we see that intermediate terms cancel since each term has the same weight under linear scheduling. In contrast to Grosse et al. (2013), this allows us to derive linear scheduling result which holds for finite T.

The above proves the proposition for the case of single sample AIS. Using additional chains and averaging inside the logarithm, the gap in the *multi-sample* AIS bounds is less than or equal to that of the single sample bounds by similar arguments which are used to show IWAE bounds are tighter than the ELBO or EUBO (Burda et al., 2016; Sobolev & Vetrov, 2019). Thus, the gap in the Independent Multi-Sample AIS sandwich bounds also decreases linearly in T. □

I REVERSE IWAE

In this section, we propose Reverse IWAE (RIWAE), which is an impractical alternative to standard IWAE. However, we use this as the basis for our Independent Reverse and Coupled Reverse Multi-Sample AIS bounds in Sec. 3.3.

I.1 PROBABILISTIC INTERPRETATION AND BOUNDS

Similar to reverse importance sampling, we consider K independent posterior samples in the extended state space proposal distribution. The target distribution for our importance sampling scheme is a mixture of $K-1$ posterior distributions and one variational distribution,

$$q_{\text{PROP}}^{\text{RIWAE}}(\mathbf{z}^{(1:K)}|\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} q_\theta(\mathbf{z}^{(k)}|\mathbf{x}) \prod_{j=1, j\neq k}^{K} p(\mathbf{z}^{(j)}|\mathbf{x}) \qquad p_{\text{TGT}}^{\text{RIWAE}}(\mathbf{z}^{(1:K)}|\mathbf{x}) = \prod_{k=1}^{K} p(\mathbf{z}^{(k)}|\mathbf{x})\,. \tag{129}$$

We visualize this sampling scheme in Fig. 4.

Similarly to Eq. 30, the log-ratio simplifies to

$$\log \frac{p_{\text{TGT}}^{\text{RIWAE}}(\mathbf{z}^{(1:K)}, \mathbf{x})}{q_{\text{PROP}}^{\text{RIWAE}}(\mathbf{z}^{(1:K)}, \mathbf{x})} = \log \frac{\prod_{k=1}^{K} p(\mathbf{x}, \mathbf{z}^{(k)})}{\frac{1}{K}\sum_{k=1}^{K} q_\theta(\mathbf{z}^{(k)}|\mathbf{x}) \prod_{\substack{j=1\\ j\neq k}}^{K} p(\mathbf{x}, \mathbf{z}^{(k)})} \tag{130}$$

$$= -\log\left(\frac{1}{K}\sum_{k=1}^{K} q_\theta(\mathbf{z}^{(k)}|\mathbf{x}) \prod_{\substack{j=1\\ j\neq k}}^{K} p(\mathbf{x}, \mathbf{z}^{(k)}) \cdot \prod_{k=1}^{k} p(\mathbf{x}, \mathbf{z}^{(k)})^{-1}\right)$$

$$= -\log \frac{1}{K}\sum_{k=1}^{K} \frac{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}{p(\mathbf{x}, \mathbf{z}^{(k)})} \tag{131}$$

Taking the expected log ratio under the proposal and target distributions yield lower and upper bounds on $\log p(\mathbf{x})$. These translate to upper and lower bounds on $\log p(\mathbf{x})$ which are different that standard IWAE bounds in Eq. 5,

$$-\mathbb{E}_{q_\theta(\mathbf{z}^{(1)}|\mathbf{x})p(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}{p(\mathbf{x}, \mathbf{z}^{(k)})}\right] \leq \log p(\mathbf{x}) \leq -\mathbb{E}_{p(\mathbf{z}^{(1:K)}|\mathbf{x})}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}{p(\mathbf{x}, \mathbf{z}^{(k)})}\right]$$

Note these bounds are *impractical* as they would need more than one true posterior sample from $p(\mathbf{z}|\mathbf{x})$. However, we will use them in Sec. 3.3 to derive practical multi-sample AIS bounds.

I.2 PROOF OF LOGARITHMIC IMPROVEMENT OF RIWAE

Proposition I.1. *The bias of the* RIWAE *lower bound on* $\log p(\mathbf{x})$ *decreases logarithmically in* K*. In particular, the* K*-sample lower bound* $\text{ELBO}_{\text{RIWAE}}(q, K)$ *can improve upon the single-sample* ELBO *by at most* $\log K$ *nats.*

$$\text{ELBO}(q) \leq \text{ELBO}_{\text{RIWAE}}(q, K) \leq \text{ELBO}(q) + \log K \tag{132}$$

Proof.

$$\log p(\mathbf{x}) \geq \text{ELBO}_{\text{RIWAE}}(q_\theta, K) = -\mathbb{E}_{q_\theta(\mathbf{z}^{(1)}|\mathbf{x})p(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \frac{1}{K}\sum_{i=1}^{K} \frac{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}{p(\mathbf{x}, \mathbf{z}^{(k)})}\right] \tag{133}$$

$$= -\mathbb{E}_{q_\theta(\mathbf{z}^{(1)}|\mathbf{x})p(\mathbf{z}^{(2:K)}|\mathbf{x})}\left[\log \sum_{i=1}^{K} \frac{q_\theta(\mathbf{z}^{(k)}|\mathbf{x})}{p(\mathbf{x}, \mathbf{z}^{(k)})}\right] + \log K$$

$$\leq -\mathbb{E}_{q_\theta(\mathbf{z}^{(1)}|\mathbf{x})}\left[\log \frac{q_\theta(\mathbf{z}^{(1)}|\mathbf{x})}{p(\mathbf{x}, \mathbf{z}^{(1)})}\right] + \log K \tag{134}$$

$$= \text{ELBO}(q_\theta) + \log K \tag{135}$$

where, in the third line, we have used the fact that the likelihood ratio is non-negative. Thus, the sum of $\log$ likelihood ratios is greater than a single term and the negative sign results in a less-than or equal-to inequality. □

The logarithmic improvement on the RIWAE lower bound in Prop. I.1 is analogous to the logarithmic improvement on the IWAE upper bound in Prop. B.2, and similarly translates to bounds on MI.

J INDEPENDENT REVERSE MULTI-SAMPLE AIS

In similar fashion to Reverse IWAE, we now use K independent reverse AIS chains to form an extended state space target distribution $p_{\text{TGT}}^{(}\mathbf{z}_{0:T}^{(1:K)}, \mathbf{x})$, and a mixture proposal distribution which contains a single forward AIS chain (see Fig. 2).

$$q_{\text{PROP}}^{\text{IR-AIS}}(\mathbf{z}_{0:T}^{(1:K)}, \mathbf{x}) = \frac{1}{K}\sum_{s=1}^{K} q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(s)}|\mathbf{x}) \prod_{k=1,k\neq s}^{K} p_{\text{TGT}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(k)}|\mathbf{x}) \qquad p_{\text{TGT}}^{\text{IR-AIS}}(\mathbf{z}_{0:T}^{(1:K)}|\mathbf{x}) = \prod_{k=1}^{K} p_{\text{TGT}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(k)}|\mathbf{x})$$

As in Eq. 130-131, the unnormalized log ratio becomes

$$\log \frac{p_{\text{TGT}}^{\text{IR-AIS}}(\mathbf{z}_{0:T}^{(1:K)}, \mathbf{x})}{q_{\text{PROP}}^{\text{IR-AIS}}(\mathbf{z}_{0:T}^{(1:K)}, \mathbf{x})} = \log \frac{\prod_{k=1}^{K} p_{\text{TGT}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(k)}|\mathbf{x})}{\frac{1}{K}\sum_{k=1}^{K} q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(k)}, \mathbf{x}) \prod_{t=1}^{T} p_{\text{TGT}}^{\text{AIS}}(\mathbf{x}, \mathbf{z}_{0:T}^{(k)})} \tag{136}$$

$$= -\log\left(\frac{1}{K}\sum_{k=1}^{K} q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(k)}, \mathbf{x}) \prod_{\substack{j=1\\ j\neq k}}^{K} p_{\text{TGT}}^{\text{AIS}}(\mathbf{x}, \mathbf{z}_{0:T}^{(k)}) \cdot \prod_{k=1}^{k} p_{\text{TGT}}^{\text{AIS}}(\mathbf{x}, \mathbf{z}_{0:T}^{(k)})^{-1}\right)$$

$$= -\log \frac{1}{K}\sum_{k=1}^{K} \frac{q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(k)}, \mathbf{x})}{p_{\text{TGT}}^{\text{AIS}}(\mathbf{x}, \mathbf{z}_{0:T}^{(k)})} \tag{137}$$

Taking expectations of the log unnormalized density ratio under the proposal and target, respectively, yield lower and upper bounds on $\log p(\mathbf{x})$

$$\underbrace{-\mathbb{E}_{\substack{\mathbf{z}_{0:T}^{(1)}\sim q_{\text{PROP}}^{\text{AIS}}\\ \mathbf{z}_{0:T}^{(2:K)}\sim p_{\text{TGT}}^{\text{AIS}}}}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(k)}|\mathbf{x})}{p_{\text{TGT}}^{\text{AIS}}(\mathbf{x}, \mathbf{z}_{0:T}^{(k)})}\right]}_{\text{ELBO}_{\text{IR-AIS}}(\pi_0, K, T)} \leq \log p(\mathbf{x}) \leq \underbrace{-\mathbb{E}_{\mathbf{z}_{0:T}^{(1:K)}\sim p_{\text{TGT}}^{\text{AIS}}}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(k)}|\mathbf{x})}{p_{\text{TGT}}^{\text{AIS}}(\mathbf{x}, \mathbf{z}_{0:T}^{(k)})}\right]}_{\text{EUBO}_{\text{IR-AIS}}(\pi_0, K, T)}. \tag{138}$$

However, reverse Multi-Sample AIS may be impractical in common settings such as posterior inference or energy-based modeling where the normalization constant is unknown, since is infeasible to obtain more than single sample from the target distribution.

Proposition J.1. *For the Independent Reverse Multi-Sample* AIS *log partition lower bound, increasing K leads to (at most) logarithmic reduction of bias compared with*

$$\text{ELBO}_{\text{IR-AIS}}(K, T) - \text{ELBO}_{\text{AIS}}(T) \leq \log K \tag{139}$$

Proof.

$$\log p(\mathbf{x}) \geq \text{ELBO}_{\text{IR-AIS}}(K, T) = -\mathbb{E}_{q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(1)}|\mathbf{x})p_{\text{TGT}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(2:K)}|\mathbf{x})}\left[\log \frac{1}{K}\sum_{i=1}^{K} \frac{q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(k)}|\mathbf{x})}{p_{\text{TGT}}^{\text{AIS}}(\mathbf{x}, \mathbf{z}_{0:T}^{(k)})}\right] \tag{140}$$

$$= -\mathbb{E}_{q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(1)}|\mathbf{x})p_{\text{TGT}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(2:K)}|\mathbf{x})}\left[\log \sum_{i=1}^{K} \frac{q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(k)}|\mathbf{x})}{p_{\text{TGT}}^{\text{AIS}}(\mathbf{x}, \mathbf{z}_{0:T}^{(k)})}\right] + \log K$$

$$\leq -\mathbb{E}_{q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(1)}|\mathbf{x})}\left[\log \frac{q_{\text{PROP}}^{\text{AIS}}(\mathbf{z}_{0:T}^{(k)}|\mathbf{x})}{p_{\text{TGT}}^{\text{AIS}}(\mathbf{x}, \mathbf{z}_{0:T}^{(k)})}\right] + \log K \tag{141}$$

$$= \text{ELBO}_{\text{AIS}}(T) + \log K \tag{142}$$

where, in the third line, we have used the fact that the likelihood ratio is non-negative. □

K COUPLED REVERSE MULTI-SAMPLE AIS

K.1 PROBABILISTIC INTERPRETATION AND BOUNDS

The Coupled Reverse Multi-Sample AIS extended state space target distribution in Fig. 2 initializes K backward chains from a single target sample $\mathbf{z}_T \sim \pi_T(\mathbf{z}|\mathbf{x})$, with remaining transitions denoted $p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}_{0:T-1}|\mathbf{z}_T)$ since they are identical to standard AIS in Eq. 14.

$$p^{\text{CR-AIS}}_{\text{TGT}}(\mathbf{z}^{(1:K)}_{0:T-1}, \mathbf{z}_T, \mathbf{x}) = \pi_T(\mathbf{z}_T, \mathbf{x}) \prod_{k=1}^{K} p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}^{(k)}_{0:T-1}|\mathbf{z}_T, \mathbf{x}). \tag{143}$$

We can construct log-partition bounds using the unnormalized importance ratio

$$\log \frac{p^{\text{CR-AIS}}_{\text{TGT}}(\mathbf{z}^{(1:K)}_{0:T}, \mathbf{x})}{q^{\text{CR-AIS}}_{\text{PROP}}(\mathbf{z}^{(1:K)}_{0:T}|\mathbf{x})} = \log \frac{p(\mathbf{z}_T, \mathbf{x}) \prod_{k=1}^{K} p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}^{(k)}_{0:T-1}|\mathbf{z}_T)}{\frac{1}{K}\sum_{j=1}^{K} q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(k)}_{0:T-1}, \mathbf{z}_T|\mathbf{x}) \prod_{\substack{k=1\\ j\neq k}}^{K} p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}^{(k)}_{0:T-1}|\mathbf{z}_T)} \tag{144}$$

$$= -\log \frac{1}{K} \sum_{k=1}^{K} \frac{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(k)}_{0:T}, \mathbf{z}_T|\mathbf{x})}{p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}^{(k)}_{0:T-1}|\mathbf{z}_T) p(\mathbf{z}_T, \mathbf{x})} \tag{145}$$

Taking the expected log ratio under the proposal and target yields lower and upper bounds on $\log p(\mathbf{x})$,

$$\text{ELBO}_{\text{CR-AIS}}(\pi_0, K, T) := -\mathbb{E}_{\substack{\mathbf{z}^{(1)}_{0:T-1}, \mathbf{z}_T \sim q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}_{0:T}|\mathbf{x}) \\ \mathbf{z}^{(2:K)}_{0:T-1} \sim p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}_{0:T-1}|\mathbf{z}_T, \mathbf{x})}} \left[\log \frac{1}{K} \sum_{k=1}^{K} \frac{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(k)}_{0:T-1}, \mathbf{z}_T|\mathbf{x})}{p^{\text{AIS}}_{\text{TGT}}(\mathbf{x}, \mathbf{z}^{(k)}_{0:T-1}, \mathbf{z}_T)} \right] \leq \log p(\mathbf{x}) \tag{146}$$

$$\text{EUBO}_{\text{CR-AIS}}(\pi_0, K, T) := -\mathbb{E}_{\substack{\mathbf{z}_T \sim \pi_T(\mathbf{z}_T|\mathbf{x}) \\ \mathbf{z}^{(1:K)}_{0:T-1} \sim p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}_{0:T-1}|\mathbf{z}_T, \mathbf{x})}} \left[\log \frac{1}{K} \sum_{k=1}^{K} \frac{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(k)}_{0:T-1}, \mathbf{z}_T|\mathbf{x})}{p^{\text{AIS}}_{\text{TGT}}(\mathbf{x}, \mathbf{z}^{(k)}_{0:T-1}, \mathbf{z}_T)} \right] \geq \log p(\mathbf{x}) \tag{147}$$

K.2 PROOF OF PROPOSITION 3.3

Proposition 3.3. *For a given T, the bias of the Coupled Reverse Multi-Sample* AIS *upper bound on* MI*, $I_{\text{CR-AIS}_U}(\pi_0, K, T)$, decreases (at most) logarithmically in K.*

$$I_{\text{AIS}_U}(\pi_0, T) - \log K \leq I_{\text{CR-AIS}_U}(\pi_0, K, T) \leq I_{\text{AIS}_U}(\pi_0, T). \tag{20}$$

Proof.

$$\begin{aligned}
\log p(\mathbf{x}) \geq & \text{ELBO}_{\text{CR-AIS}}(K, T) = -\mathbb{E}_{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(1)}_{0:T-1}, \mathbf{z}_T|\mathbf{x}) p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}^{(2:K)}_{0:T-1}, \mathbf{z}_T|\mathbf{x})} \left[\log \frac{1}{K} \sum_{i=1}^{K} \frac{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(k)}_{0:T-1}, \mathbf{z}_T|\mathbf{x})}{p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}^{(k)}_{0:T-1}|\mathbf{z}_T) p(\mathbf{z}_T, \mathbf{x})} \right] \\
&= -\mathbb{E}_{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(1)}_{0:T-1}, \mathbf{z}_T|\mathbf{x}) p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}^{(2:K)}_{0:T-1}, \mathbf{z}_T|\mathbf{x})} \left[\log \sum_{i=1}^{K} \frac{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(k)}_{0:T-1}, \mathbf{z}_T|\mathbf{x})}{p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}^{(k)}_{0:T-1}|\mathbf{z}_T) p(\mathbf{z}_T, \mathbf{x})} \right] + \log K \\
&\leq -\mathbb{E}_{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(1)}_{0:T-1}, \mathbf{z}_T|\mathbf{x})} \left[\log \frac{q^{\text{AIS}}_{\text{PROP}}(\mathbf{z}^{(k)}_{0:T-1}, \mathbf{z}_T|\mathbf{x})}{p^{\text{AIS}}_{\text{TGT}}(\mathbf{z}^{(k)}_{0:T-1}|\mathbf{z}_T) p(\mathbf{z}_T, \mathbf{x})} \right] + \log K \\
&= \text{ELBO}_{\text{AIS}}(T) + \log K
\end{aligned} \tag{148}$$

where, in the third line, we have used the fact that the likelihood ratio is non-negative. Since $I_{\text{CR-AIS}_U}(\pi_0, K, T) = \mathbb{E}_{p(\mathbf{x}|\mathbf{z})}[\log p(\mathbf{x}|\mathbf{z})] - \text{ELBO}_{\text{CR-AIS}}(\pi_0, K, T)$ and $I_{\text{AIS}_L}(\pi_0, T) = \mathbb{E}_{p(\mathbf{x}|\mathbf{z})}[\log p(\mathbf{x}|\mathbf{z})] - \text{ELBO}_{\text{AIS}}(\pi_0, T)$, then Eq. 148 implies $I_{\text{CR-AIS}_U}(\pi_0, K, T) \geq I_{\text{AIS}_U}(\pi_0, T) - \log K$ as desired. □

L EXPERIMENTAL DETAILS

L.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 (Table 1), the encoder's architecture is $784 - 1024 - 1024 - 1024 - z$ where z is the latent code size shown in the row header of Table 1. The decoder architecture is the reverse of the encoder architecture. The decoder variance is a global parameter learned during the training.

For the GAN experiments on MNIST (Table 1), we used the same decoder architecture as our VAEs. In order to stabilize the training dynamic, we used the gradient penalty (GP) (Salimans et al., 2016).

The network was trained for 300 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 evaluation.

CIFAR-10. For the CIFAR-10 experiments (Table 1), we experimented with the DCGAN (Radford et al., 2015) with Gradient Penalty (GP-GAN) (Gulrajani et al., 2017). The number at the end of each model name in Table 1 indicates the latent code size.

L.2 MULTI-SAMPLE AIS SETTINGS

We conducted more extensive experimentation for upper bounding the mutual information (lower bounding the partition function $\log p(\mathbf{x})$) and lower bounding the Mutual Information (upper bounding the partition function $\log p(\mathbf{x})$) using coupled reverse multi-sample AIS (See Fig. 4 for the definitions, results are shown in Table 1) For all these experiments, we used 20 leap frog steps for HMC. We used 48 independent chains for each data point. The step size of HMC was adaptive tuned to achieve an average acceptance probability of 65% as detailed in Neal (2001).

For all MNIST experiments in Table 1, we evaluated on a single batch size of 128 simulated data. We used a sigmoid temperature schedule as used in Wu et al. (2016) but with radius parameter set to 10 instead of 4. For all CIFAR experiments in Table 1 we used a single batch of 32 simulated data. It is much harder to get a tight upper and lower bound for CIFAR models, and we discovered that the tightness of the bound is sensitive to the radius parameter, so we used the optimal radius for each setting. For GAN100 models on CIFAR10, the radius parameter for the temperature is 6 when the initial distribution is $q(\mathbf{z}|\mathbf{x})$ and 8 when it is $\log p(\mathbf{z})$.

L.3 RUNTIME COMPARISON

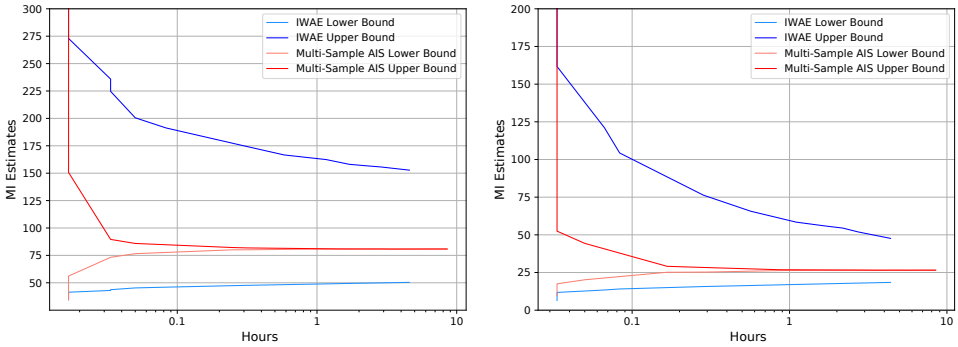
For MNIST, it took about 20 mins to run IWAE for $K = 1M$, and it took about 8 minutes to run AIS for $T = 500$ and about 8 hours for $T = 30k$. For CIFAR, it took about 70 mins to run AIS for $T = 500$ steps and about nine days for $T = 100k$.

In Fig. 8, we evaluate the tradeoff between runtime and bound tightness for evaluating the generative MI for VAE and GAN models with 100-dimensional latent codes trained on the MNIST dataset. We compare IWAE and multi-sample AIS evaluation, with the same experimental settings as in App. L.2 and an initial distribution $q_\theta(\mathbf{z}|\mathbf{x})$. We plot wall clock time on the x-axis, where increasing runtime reflects increasing K for IWAE and increasing T for AIS.

As predicted by Prop. 2.1, the improvement of the IWAE lower bound on MI is logarithmic with increasing K, which appears linear with the logarithmic scale of the x-axis. We can also confirm the result in Prop. 3.2 that the gap in the multi-sample AIS lower and upper bounds reduces linearly with increasing T, as each bound appears to improve exponentially in the logarithmic scale of the x-axis.

L.4 INITIAL DISTRIBUTION OF AIS

We can use any distribution that is tractable to sample and evaluate as the initial distribution of AIS. In practice, it is important to use a distribution that is mode-covering for the posterior distribution so



(a) VAE-100 (b) GAN-100

Figure 8: Runtime vs. Bound Tightness for VAE (left) and GAN (right) models on MNIST.

that the AIS particles do not get stuck in a few modes. A typical option for the initial distribution is the prior $p(\mathbf{z})$. In this case, the intermediate distributions will be $q_\beta(\mathbf{z}|\mathbf{x}) \propto p(\mathbf{z})p(\mathbf{x}|\mathbf{z})^\beta$.

Another possibility is a $q_\theta(\mathbf{z}|\mathbf{x})$ distribution which has been trained to maximize the importance sampling or BA lower bound on mutual information:

$$I(\mathbf{x};\mathbf{z}) \geq I(\mathbf{x};\mathbf{z}) - D_{\mathrm{KL}}[p(\mathbf{z}|\mathbf{x})||q(\mathbf{z}|\mathbf{x})] = \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{q(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right]$$

As this bound optimizes a reverse KL divergence, the learned $q_\theta(\mathbf{z}|\mathbf{x})$ will be a mode covering distribution for $p(\mathbf{z}|\mathbf{x})$ and thus a good candidate from which to start the AIS chains. In this case, the intermediate distributions for single-sample AIS chains will be $q_\beta(\mathbf{z}|\mathbf{x}) \propto q(\mathbf{z}|\mathbf{x})^{(1-\beta)}p(\mathbf{x},\mathbf{z})^\beta$.

L.5 Analytical Solution of the Mutual Information on the Linear VAE

In order to verify our implementations, we have derived the mutual information analytically for Linear VAEs trained on MNIST.

For simplicity, we assume a fixed identity covariance matrix I at the output of the conditional likelihood of the linear VAE decoder, i.e., 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 $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{W}$ 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 is: $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \mathbf{b}, \mathbf{I})$ and thus we can solve for the marginal:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{x} = \mathbf{b}, \boldsymbol{\Sigma}_\mathbf{x} = \mathbf{I} + \mathbf{W}\mathbf{W}^\intercal) \tag{149}$$

Thus the differential entropy of $\mathbf{x}$ is:

$$h(\mathbf{x}) = \frac{k}{2} + \frac{k}{2}\log(2\pi) + \frac{1}{2}\log(\det \boldsymbol{\Sigma}_\mathbf{x}) \tag{150}$$

Where k is the dimension of the observation. The conditional entropy:

$$h(\mathbf{x}|\mathbf{z}) = \frac{k}{2} + \frac{k}{2}\log(2\pi) + \frac{1}{2}\log(\det \mathbf{I}) \tag{151}$$

$$= \frac{k}{2} + \frac{k}{2}\log(2\pi) \tag{152}$$

Thus the mutual information:

$$I(\mathbf{x};\mathbf{z}) = h(\mathbf{x}) - h(\mathbf{x}|\mathbf{z}) \tag{153}$$

$$= \frac{1}{2}\log(\det \boldsymbol{\Sigma}_\mathbf{x}) \tag{154}$$

M ADDITIONAL EXPERIMENTS

M.1 COMPARISON OF MULTI-SAMPLE AIS BOUNDS

Fig. 9 compares performance of our various Multi-Sample AIS bounds of Linear VAE, VAE, and GAN models with 20 latent variables trained on MNIST. We perform this comparison with the goal of making recommendations for which Multi-Sample AIS bounds to use in practice.

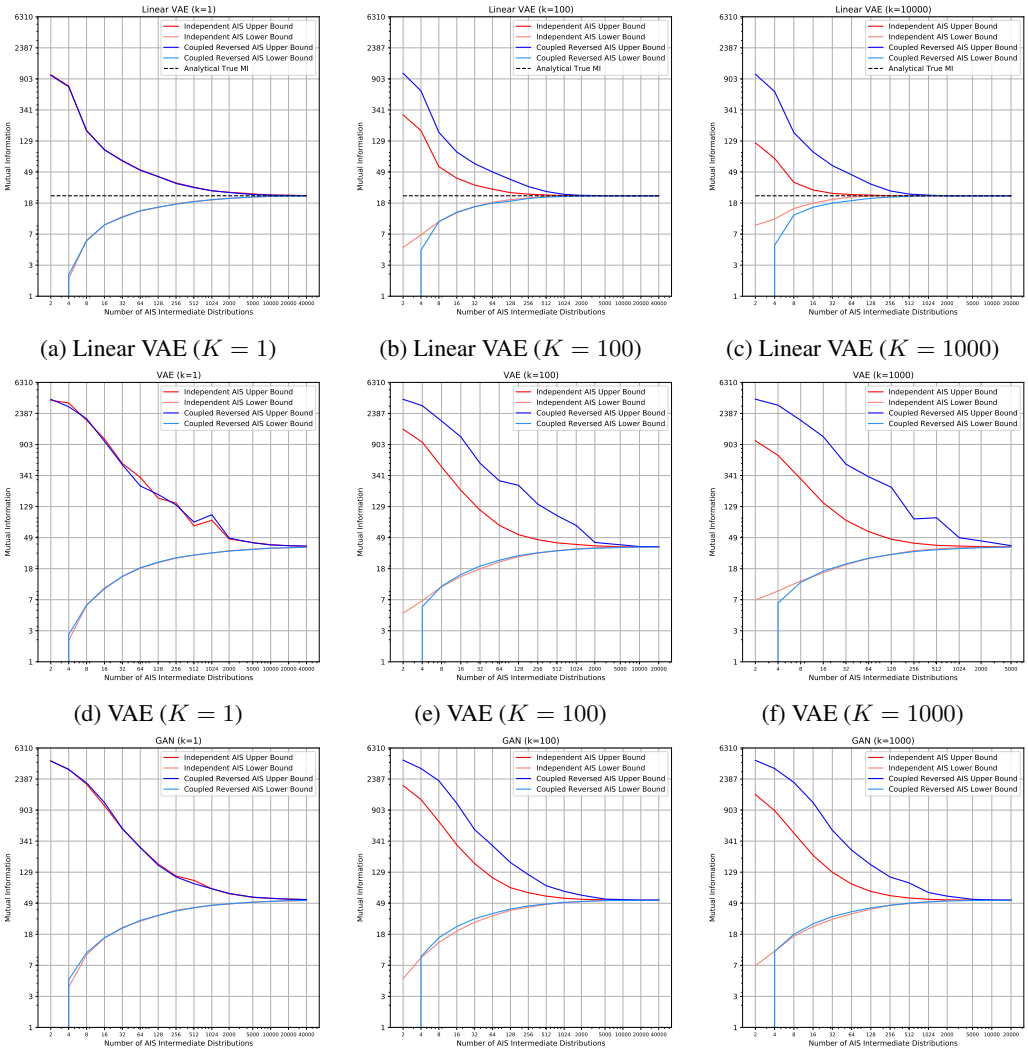
First, the MI upper bound using the Independent Multi-Sample AIS lower bound on $\log p(\mathbf{x})$, which corresponds to the forward direction of BDMC, outperforms its counterparts. This bound uses independent samples and is not limited to only $\log K$ improvement, compared with the Coupled Reverse Multi-Sample AIS MI upper bound (Prop. 3.3). We recommend using the Independent Multi-Sample AIS MI upper bound (LHS of Eq. 16) in all cases.

The results are less conclusive for the multi-sample AIS lower bounds on MI, where either the Independent Multi-Sample AIS EUBO or Coupled Reverse AIS EUBO may be preferable for log partition function estimation. Recall that these bounds have different sources of stochasticity that provide improvement over single-chain AIS. The stochasticity in the Independent Multi-Sample AIS EUBO comes from $K-1$ independent 'negative' forward chains which, by Prop. 3.1, can only lead to $\log K$ improvement over single-sample AIS. However, these gains are easily attained for low values of T. For example, with zero intermediate distributions ($T=1$), and $\pi_0(\mathbf{z}) = p(\mathbf{z})$, $I_{\text{IM-AIS}_L}(p(\mathbf{z}), K, T=1)$ reduces to Structured INFO-NCE. This may be useful for quickly estimating values of MI at a similar order of magnitude as $\log K$.

The stochasticity in the Coupled Reverse AIS EUBO is induced by MCMC transitions in $K-1$ coupled backward chains. While this does not formally limit the improvement over single-sample AIS, we see in Fig. 9 that at least moderate values of T may be needed to match or marginally improve upon Independent Multi-Sample AIS. These observations suggest that the preferred lower MI bounds may vary based on the scale of the true MI and the amount of computation available.

M.2 CONFIDENCE INTERVALS FOR MULTI-SAMPLE AIS EXPERIMENTS

In Table 4 and Table 5, we have conducted more extensive experimentation for upper bounding the mutual information (lower bounding the partition function $\log p(\mathbf{x})$) and lower bounding the Mutual Information (upper bounding the partition function $\log p(\mathbf{x})$) using Coupled Reverse Multi-Sample AIS. We have also computed the 95% confidence intervals for those experiments. For MNIST experiments, we computed the confidence interval over 8 batches of 16 data points. For CIFAR experiments, due to computational constraints, we computed the confidence intervals over 8 batches of 4 data points. Results for MNIST are shown in Table 4 and results for CIFAR are shown in Table 5.



(a) Linear VAE ($K = 1$) (b) Linear VAE ($K = 100$) (c) Linear VAE ($K = 1000$)

(d) VAE ($K = 1$) (e) VAE ($K = 100$) (f) VAE ($K = 1000$)

(g) GAN ($K = 1$) (h) GAN ($K = 100$) (i) GAN ($K = 1000$)

Figure 9: Comparing Multi-Sample AIS sandwich bounds and evaluating the effect of K and T for MI estimation in deep generative models.

Method	Proposal		VAE2	VAE10	VAE100	GAN2	GAN10	GAN100
AIS T=1	$p(\mathbf{z})$	UB	$(193.57, 254.28)$	$(1794.64, 2005.89)$	$(5669.46, 6202.81)$	$(698.52, 836.26)$	$(752.24, 950.16)$	$(812.11, 879.82)$
		LB	$(0, 0)$	$(0, 0)$	$(0, 0)$	$(0, 0)$	$(0, 0)$	$(0, 0)$
	$q(\mathbf{z}\|\mathbf{x})$	UB	$(8.86, 9.46)$	$(60.41, 67.88)$	$(351.43, 382.23)$	$(6.96, 7.49)$	$(283.23, 397.36)$	$(476.06, 549.21)$
		LB	$(7.04, 7.94)$	$(20.70, 21.50)$	$(31.71, 36.59)$	$(6.96, 7.49)$	$(3.43, 4.22)$	$(2.34, 2.80)$
AIS T=500	$p(\mathbf{z})$	UB	$(14.12, 20.12)$	$(100.34, 114.38)$	$(448.16, 539.69)$	$(68.70, 127.48)$	$(137.67, 197.84)$	$(193.06, 264.36)$
		LB	$(7.93, 8.53)$	$(33.11, 33.95)$	$(77.34, 82.42)$	$(7.71, 8.74)$	$(20.97, 22.07)$	$(23.75, 25.15)$
	$q(\mathbf{z}\|\mathbf{x})$	UB	$(8.71, 9.24)$	$(51.62, 55.00)$	$(237.56, 271.13)$	$(10.55, 11.71)$	$(80.34, 108.39)$	$(139.34, 190.78)$
		LB	$(8.70, 9.24)$	$(33.68, 34.69)$	$(77.34, 82.42)$	$(10.07, 10.83)$	$(21.60, 22.60)$	$(24.30, 25.82)$
AIS (T=30K)	$p(\mathbf{z})$	UB	$(8.73, 9.23)$	$(33.84, 34.80)$	$(78.48, 83.58)$	$(10.13, 10.96)$	$(22.13, 23.04)$	$(25.65, 27.56)$
		LB	$(8.59, 9.13)$	$(33.80, 34.75)$	$(78.37, 83.48)$	$(8.59, 9.13)$	$(22.07, 22.96)$	$(25.54, 27.33)$
	$q(\mathbf{z}\|\mathbf{x})$	UB	$(8.72, 9.25)$	$(33.84, 34.80)$	$(78.42, 83.52)$	$(10.18, 10.95)$	$(22.13, 23.05)$	$(25.62, 27.51)$
		LB	$(8.72, 9.24)$	$(33.83, 34.77)$	$(78.41, 83.46)$	$(10.17, 10.95)$	$(22.11, 22.98)$	$(25.56, 27.42)$
IWAE K=1	$p(\mathbf{z})$	UB	$(735.75, 864.75)$	$(3751.24, 4185.62)$	$(10152.39, 11687.02)$	$(1586.98, 1769.75)$	$(1599.69, 1777.36)$	$(1525.07, 1680.19)$
		LB	$(0, 0)$	$(0, 0)$	$(0, 0)$	$(0, 0)$	$(0, 0)$	$(0, 0)$
	$q(\mathbf{z}\|\mathbf{x})$	UB	$(12.83, 16.16)$	$(132.91, 162.64)$	$(647.144, 758.28)$	$(68.33, 152.98)$	$(1021.08, 1175.48)$	$(1228.66, 1463.56)$
		LB	$(7.48, 7.80)$	$(20.81, 21.42)$	$(31.83, 37.28)$	$(6.93, 7.43)$	$(3.43, 4.21)$	$(2.46, 2.71)$
IWAE K=1K	$p(\mathbf{z})$	UB	$(23.19, 35.38)$	$(1127.70, 1274.83)$	$(3985.98, 4502.03)$	$(84.67, 128.39)$	$(409.2, 480.08)$	$(435.03, 531.65)$
		LB	$(6.67, 6.87)$	$(6.91, 6.91)$	$(6.91, 6.91)$	$(6.84, 6.90)$	$(6.91, 6.91)$	$(6.91, 6.91)$
	$q(\mathbf{z}\|\mathbf{x})$	UB	$(8.74, 9.29)$	$(44.43, 47.61)$	$(253.62, 280.20)$	$(10.33, 11.21)$	$(124.53, 176.76)$	$(257.41, 325.35))$
		LB	$(8.74, 9.27)$	$(27.70, 28.29)$	$(38.75, 44.19)$	$(10.14, 10.85)$	$(10.34, 11.12)$	9.48
IWAE K=1M	$p(\mathbf{z})$	UB	$(8.75, 9.25)$	$(358.34, 400.33)$	$(2083.056, 2435.14)$	$(10.51, 11.07)$	$(77.13, 92.64)$	$(104.75, 137.96)$
		LB	$(8.73, 9.24)$	$(13.82, 13.82)$	$(13.82, 13.82)$	$(10.05, 10.87)$	$(13.76, 13.82)$	$(13.82, 13.82)$
	$q(\mathbf{z}\|\mathbf{x})$	UB	$(8.74, 9.27)$	$(34.24, 35.51)$	$(161.67, 179.07)$	$(10.20, 10.96)$	$(28.22, 35.76)$	$(62.95, 75.82)$
		LB	$(8.74, 9.27)$	$(32.83, 33.48)$	$(45.65, 51.10)$	$(10.20, 10.96)$	$(16.87, 17.87)$	$(16.17, 16.41)$

Table 4: Confidence intervals of AIS and IWAE estimates of Mutual Information on MNIST. UB stands for Upper Bound, and LB stands for Lower Bound.

Model	Proposal		GAN100
AIS (T=1)	$p(\mathbf{z})$	UB	$(10193.76, 13563.06)$
		LB	$(0, 0)$
	$q(\mathbf{z}\|\mathbf{x})$	UB	$(7075.05, 10428.11)$
		LB	$(5.60, 9.51)$
AIS T=500	$p(\mathbf{z})$	UB	$(5102.42, 7106.76)$
		LB	$(69.96, 90.60)$
	$q(\mathbf{z}\|\mathbf{x})$	UB	$(4604.80, 6028.27)$
		LB	$(70.96, 95.73)$
AIS T=100K	$p(\mathbf{z})$	UB	$(376.09, 595.069)$
		LB	$(163.97, 226.43)$
	$q(\mathbf{z}\|\mathbf{x})$	UB	$(201.71, 227.14)$
		LB	$(151.72, 202.27)$
IWAE K=1	$p(\mathbf{z})$	UB	$(21916.21, 31105.44)$
		LB	$(0, 0)$
	$q(\mathbf{z}\|\mathbf{x})$	UB	$(4709.99, 15771.13)$
		LB	$(6.52, 8.80)$
IWAE K=1K	$p(\mathbf{z})$	UB	$(7674.19, 10944.53)$
		LB	$(6.91, 6.91)$
	$q(\mathbf{z}\|\mathbf{x})$	UB	$(4201.98, 5882.21)$
		LB	$(16.48, 20.33)$
IWAE K=20K	$p(\mathbf{z})$	UB	$(6245.65, 9150.71)$
		LB	$(9.90, 9.90)$
	$q(\mathbf{z}\|\mathbf{x})$	UB	$(3307.19, 4595.68)$
		LB	$(19.47, 23.32)$

Table 5: Confidence intervals of AIS and IWAE estimates of Mutual Information on CIFAR. UB stands for Upper Bound, and LB stands for Lower Bound.