Amortized Inference Regularization

Rui Shu
Stanford University
ruishu@stanford.edu

Hung H. Bui
DeepMind
buih@google.com

Shengjia Zhao
Stanford University
sjzhao@stanford.edu

Mykel J. Kochenderfer
Stanford University
mykel@stanford.edu

Stefano Ermon
Stanford University
ermon@cs.stanford.edu

Abstract

The variational autoencoder (VAE) is a popular model for density estimation and representation learning. Canonically, the variational principle suggests to prefer an expressive inference model so that the variational approximation is accurate. However, it is often overlooked that an overly-expressive inference model can be detrimental to the test set performance of both the amortized posterior approximator and, more importantly, the generative density estimator. In this paper, we leverage the fact that VAEs rely on *amortized* inference and propose techniques for *amortized inference regularization* (AIR) that control the smoothness of the inference model. We demonstrate that, by applying AIR, it is possible to improve VAE generalization on both inference and generative performance. Our paper challenges the belief that amortized inference is simply a mechanism for approximating maximum likelihood training and illustrates that regularization of the amortization family provides a new direction for understanding and improving generalization in VAEs.

1 Introduction

Variational autoencoders are a class of generative models with widespread applications in density estimation, semi-supervised learning, and representation learning [1, 2, 3, 4]. A popular approach for the training of such models is to maximize the log-likelihood of the training data. However, maximum likelihood is often intractable due to the presence of latent variables. Variational Bayes resolves this issue by constructing a tractable lower bound of the log-likelihood and maximizing the lower bound instead. Classically, Variational Bayes introduces per-sample approximate proposal distributions that need to be optimized using a process called variational inference. However, per-sample optimization incurs a high computational cost. A key contribution of the variational autoencoding framework is the observation that the cost of variational inference can be amortized by using an amortized inference model that learns an efficient mapping from samples to proposal distributions. This perspective portrays amortized inference as a tool for efficiently approximating maximum likelihood training. Many techniques have since been proposed to expand the expressivity of the amortized inference model in order to better approximate maximum likelihood training [5, 6, 7, 8].

In this paper, we challenge the conventional role that amortized inference plays in variational autoencoders. For datasets where the generative model is prone to overfitting, we show that having an amortized inference model actually provides a new and effective way to regularize maximum likelihood training. Rather than making the amortized inference model more expressive, we propose instead to restrict the capacity of the amortization family. Through amortized inference regularization (AIR), we show that it is possible to reduce the inference gap and increase the log-likelihood performance on the test set. We propose several techniques for AIR and provide extensive theoretical and empirical analyses of our proposed techniques when applied to the variational autoencoder and the

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

importance-weighted autoencoder. By rethinking the role of the amortized inference model, amortized inference regularization provides a new direction for studying and improving the generalization performance of latent variable models.

2 Background and Notation

2.1 Variational Inference and the Evidence Lower Bound

Consider a joint distribution $p_\theta(x, z)$ parameterized by θ, where $x \in \mathcal{X}$ is observed and $z \in \mathcal{Z}$ is latent. Given a uniform distribution $\hat{p}(x)$ over the dataset $\mathcal{D} = \{x^{(i)}\}$, maximum likelihood estimation performs model selection using the objective

$$\max_\theta \mathbb{E}_{\hat{p}(x)} \ln p_\theta(x) = \max_\theta \mathbb{E}_{\hat{p}(x)} \ln \int_z p_\theta(x, z) dz. \tag{1}$$

However, marginalization of the latent variable is often intractable; to address this issue, it is common to employ the variational principle to maximize the following lower bound

$$\max_\theta \mathbb{E}_{\hat{p}(x)} \left[\ln p_\theta(x) - \min_{q \in \mathcal{Q}} D(q(z) \parallel p_\theta(z \mid x)) \right] = \max_\theta \mathbb{E}_{\hat{p}(x)} \left[\max_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \ln \frac{p_\theta(x, z)}{q(z)} \right], \tag{2}$$

where D is the Kullback-Leibler divergence and $\mathcal{Q}$ is a variational family. This lower bound, commonly called the evidence lower bound (ELBO), converts log-likelihood estimation into a tractable optimization problem. Since the lower bound holds for any q, the variational family $\mathcal{Q}$ can be chosen to ensure that $q(z)$ is easily computable, and the lower bound is optimized to select the best proposal distribution $q_x^*(z)$ for each $x \in \mathcal{D}$.

2.2 Amortization and Variational Autoencoders

[1, 9] proposed to construct $p(x \mid z)$ using a parametric function $g_\theta \in \mathcal{G}(\mathcal{P}) : \mathcal{Z} \to \mathcal{P}$, where $\mathcal{P}$ is some family of distributions over x, and $\mathcal{G}$ is a family of functions indexed by parameters θ. To expedite training, they observed that it is possible to amortize the computational cost of variational inference by framing the per-sample optimization process as a *regression* problem; rather than solving for the optimal proposal $q_x^*(z)$ directly, they instead use a recognition model $f_\phi \in \mathcal{F}(\mathcal{Q}) : \mathcal{X} \to \mathcal{Q}$ to predict $q_x^*(z)$. The functions (f_ϕ, g_θ) can be concisely represented as conditional distributions, where

$$p_\theta(x \mid z) = g_\theta(z)(x) \tag{3}$$
$$q_\phi(z \mid x) = f_\phi(x)(z). \tag{4}$$

The use of amortized inference yields the variational autoencoder, which is trained to maximize the variational autoencoder objective

$$\max_{\theta, \phi} \mathbb{E}_{\hat{p}(x)} \left[\mathbb{E}_{q_\phi(z|x)} \ln \frac{p(z) p_\theta(x \mid z)}{q_\phi(z \mid x)} \right] = \max_{f \in \mathcal{F}(\mathcal{Q}), g \in \mathcal{G}(\mathcal{P})} \mathbb{E}_{\hat{p}(x)} \left[\mathbb{E}_{z \sim f(x)} \ln \frac{p(z) g(z)(x)}{f(x)(z)} \right]. \tag{5}$$

We omit the dependency of $(p(z), g)$ on θ and f on ϕ for notational simplicity. In addition to the typical presentation of the variational autoencoder objective (LHS), we also show an alternative formulation (RHS) that reveals the influence of the model capacities $\mathcal{F}, \mathcal{G}$ and distribution family capacities $\mathcal{Q}, \mathcal{P}$ on the objective function. In this paper, we use (q_ϕ, f) interchangeably, depending on the choice of emphasis. To highlight the relationship between the ELBO in Eq. (2) and the standard variational autoencoder objective in Eq. (5), we shall also refer to the latter as the amortized ELBO.

2.3 Amortized Inference Suboptimality

For a fixed generative model, the optimal unamortized and amortized inference models are

$$q_x^* = \arg\max_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[\ln \frac{p_\theta(x, z)}{q(z)} \right], \text{for each } x \in \mathcal{D} \tag{6}$$

$$f^* = \arg\max_{f \in \mathcal{F}} \mathbb{E}_{\hat{p}(x)} \left[\mathbb{E}_{z \sim f(x)} \ln \frac{p_\theta(x, z)}{f(x)(z)} \right]. \tag{7}$$

A notable consequence of using an amortization family to approximate variational inference is that Eq. (5) is a lower bound of Eq. (2). This naturally raises the question of whether the learned inference model can accurately approximate the mapping $x \mapsto q_x^*(z)$. To address this question, [10] defined the inference, approximation, and amortization gaps as

$$\Delta_{\text{inf}}(\hat{p}) = \mathbb{E}_{\hat{p}(x)} D(f^*(x) \parallel p_\theta(z \mid x)) \tag{8}$$

$$\Delta_{\text{ap}}(\hat{p}) = \mathbb{E}_{\hat{p}(x)} D(q_x^*(z) \parallel p_\theta(z \mid x)) \tag{9}$$

$$\Delta_{\text{am}}(\hat{p}) = \Delta_{\text{inf}}(\hat{p}) - \Delta_{\text{ap}}(\hat{p}), \tag{10}$$

Studies have found that the inference gap is non-negligible [11] and primarily attributable to the presence of a large amortization gap [10].

The amortization gap raises two critical considerations. On the one hand, we wish to reduce the training amortization gap $\Delta_{\text{am}}(\hat{p}_{\text{train}})$. If the family $\mathcal{F}$ is too low in capacity, then it is unable to approximate $x \mapsto q_x^*$ and will thus increase the amortization gap. Motivated by this perspective, [5, 12] proposed to reduce the training amortization gap by performing stochastic variational inference on top of amortized inference. In this paper, we take the opposing perspective that an over-expressive $\mathcal{F}$ hurts generalization (see Appendix A) and that restricting the capacity of $\mathcal{F}$ is a form of regularization that can prevent both the inference *and* generative models from overfitting to the training set.

3 Amortized Inference Regularization in Variational Autoencoders

Many methods have been proposed to expand the variational and amortization families in order to better approximate maximum likelihood training [5, 6, 7, 8, 13, 14]. We argue, however, that achieving a better approximation to maximum likelihood training is not necessarily the best training objective, even if the end goal is test set density estimation. In general, it may be beneficial to regularize the maximum likelihood training objective.

Importantly, we observe that the evidence lower bound in Eq. (2) admits a natural interpretation as implicitly regularizing maximum likelihood training

$$\max_\theta \Big(\overbrace{\mathbb{E}_{\hat{p}(x)} \left[\ln p_\theta(x)\right]}^{\text{log-likelihood}} - \overbrace{\mathbb{E}_{\hat{p}(x)} \min_{q \in \mathcal{Q}} D(q(z) \parallel p_\theta(z \mid x))}^{\text{regularizer } R(\theta; \mathcal{Q})} \Big). \tag{11}$$

This formulation exposes the ELBO as a *data-dependent regularized* maximum likelihood objective. For infinite capacity $\mathcal{Q}$, $R(\theta\,;\,\mathcal{Q})$ is zero for all $\theta \in \Theta$, and the objective reduces to maximum likelihood. When $\mathcal{Q}$ is the set of Gaussian distributions (as is the case in the standard VAE), then $R(\theta\,;\,\mathcal{Q})$ is zero only if $p_\theta(z \mid x)$ is Gaussian for all $x \in \mathcal{D}$. In other words, a Gaussian variational family regularizes the true posterior $p_\theta(z \mid x)$ toward being Gaussian [10]. Careful selection of the variational family to encourage $p_\theta(z \mid x)$ to adopt certain properties (e.g. unimodality, fully-factorized posterior, etc.) can thus be considered a special case of *posterior regularization* [15, 16].

Unlike traditional variational techniques, the variational autoencoder introduces an amortized inference model $f \in \mathcal{F}$ and thus a new source of posterior regularization.

$$\max_\theta \Big(\overbrace{\mathbb{E}_{\hat{p}(x)} \left[\ln p_\theta(x)\right]}^{\text{log-likelihood}} - \overbrace{\min_{f \in \mathcal{F}(\mathcal{Q})} \mathbb{E}_{\hat{p}(x)} \left[D(f(x) \parallel p_\theta(z \mid x))\right]}^{\text{regularizer } R(\theta; \mathcal{Q}, \mathcal{F})} \Big). \tag{12}$$

In contrast to unamortized variational inference, the introduction of the amortization family $\mathcal{F}$ forces the inference model to consider the *global structure* of how $\mathcal{X}$ maps to $\mathcal{Q}$. We thus define *amortized inference regularization* as the strategy of restricting the inference model capacity $\mathcal{F}$ to satisfy certain desiderata. In this paper, we explore a special case of AIR where a candidate model $f \in \mathcal{F}$ is penalized if it is not sufficiently smooth. We propose two models that encourage inference model smoothness and demonstrate that they can reduce the inference gap and increase log-likelihood on the test set.

3.1 Denoising Variational Autoencoder

In this section, we propose using random perturbation training for amortized inference regularization. The resulting model—the denoising variational autoencoder (DVAE)—modifies the variational

autoencoder objective by injecting ε noise into the inference model

$$\max_{\theta} \left(\mathbb{E}_{\hat{p}(x)} \left[\ln p_\theta(x) \right] - \min_{f \in \mathcal{F}(\mathcal{Q})} \mathbb{E}_{\hat{p}(x)} \mathbb{E}_{\varepsilon} \left[D(f(x+\varepsilon) \parallel p_\theta(z \mid x)) \right] \right). \tag{13}$$

Note that the noise term only appears in the regularizer term. We consider the case of zero-mean isotropic Gaussian noise $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma \mathbf{I})$ and denote the denoising regularizer as $R(\theta\,;\sigma)$. At this point, we note that the DVAE was first described in [17]. However, our treatment of DVAE differs from [17]'s in both theoretical analysis and underlying motivation. We found that [17] incorrectly stated the tightness of the DVAE variational lower bound (see Appendix B). In contrast, our analysis demonstrates that the denoising objective smooths the inference model and necessarily lower bounds the original variational autoencoder objective (see Theorem 1 and Proposition 1).

We now show that 1) the optimal DVAE amortized inference model is a kernel regression model and that 2) the variance of the noise ε controls the smoothness of the optimal inference model.

Lemma 1. *For fixed $(\theta, \sigma, \mathcal{Q})$ and infinite capacity $\mathcal{F}$, the inference model that optimizes the DVAE objective in Eq. (13) is the kernel regression model*

$$f^*_\sigma(x) = \arg\min_{q \in \mathcal{Q}} \sum_{i=1}^{n} w_\sigma(x, x^{(i)}) \cdot D(q(z) \parallel p_\theta(z \mid x^{(i)})), \tag{14}$$

where $w_\sigma(x, x^{(i)}) = \frac{K_\sigma(x, x^{(i)})}{\sum_j K_\sigma(x, x^{(j)})}$ and $K_\sigma(x, y) = \exp\left(-\frac{\|x-y\|}{2\sigma^2}\right)$ is the RBF kernel.

Lemma 1 shows that the optimal denoising inference model f^*_σ is dependent on the noise level σ. The output of $f^*_\sigma(x)$ is the proposal distribution that minimizes the weighted Kullback-Leibler (KL) divergence from $f^*_\sigma(x)$ to each $p_\theta(z \mid x^{(i)})$, where the weighting $w_\sigma(x, x^{(i)})$ depends on the distance $\|x - x^{(i)}\|$ and the bandwidth σ. When $\sigma > 0$, the amortized inference model forces neighboring points $(x^{(i)}, x^{(j)})$ to have similar proposal distributions. Note that as σ increases, $w_\sigma(x, x^{(i)}) \to \frac{1}{n}$, where n is the number of training samples. Controlling σ thus modulates the smoothness of f^*_σ (we say that f^*_σ is smooth if it maps similar inputs to similar outputs under some suitable measure of similarity). Intuitively, the denoising regularizer $R(\theta\,;\sigma)$ approximates the true posteriors with a "σ-smoothed" inference model and penalizes generative models whose posteriors cannot easily be approximated by such an inference model. This intuition is formalized in Theorem 1.

Theorem 1. *Let $\mathcal{Q}$ be a minimal exponential family with corresponding natural parameter space Ω. With a slight abuse of notation, consider $f \in \mathcal{F} : \mathcal{X} \to \Omega$. Under the simplifying assumption that $p_\theta(z \mid x^{(i)})$ is contained within $\mathcal{Q}$ and parameterized by $\eta^{(i)} \in \Omega$, and that $\mathcal{F}$ has infinite capacity, then the optimal inference model in Lemma 1 returns $f^*_\sigma(x) = \eta \in \Omega$, where*

$$\eta = \sum_{i=1}^{n} w_\sigma(x, x^{(i)}) \cdot \eta^{(i)} \tag{15}$$

and Lipschitz constant of f^*_σ is bounded by $O(1/\sigma^2)$.

We wish to address Theorem 1's assumption that the true posteriors lie in the variational family. Note that for sufficiently large exponential families, this assumption is likely to hold. But even in the case where the variational family is Gaussian (a relatively small exponential family), the small approximation gap observed in [10] suggests that it is plausible that posterior regularization would encourage the true posteriors to be approximately Gaussian.

Given that σ modulates the smoothness of the inference model, it is natural to suspect that a larger choice of σ results in a stronger regularization. To formalize this notion of regularization strength, we introduce a way to partially order a set of regularizers $\{R_i(\theta)\}$.

Definition 1. *Suppose two regularizers $R_1(\theta)$ and $R_2(\theta)$ share the same minimum $\min_\theta R_1(\theta) = \min_\theta R_2(\theta)$. We say that R_1 is a stronger regularizer than R_2 if $R_1(\theta) \geq R_2(\theta)$ for all $\theta \in \Theta$.*

Note that any two regularizers can be modified via scalar addition to share the same minimum. Furthermore, if R_1 is stronger than R_2, then R_1 and R_2 share at least one minimizer. We now apply Definition 1 to characterize the regularization strength of $R(\theta\,;\sigma)$ as σ increases.

Definition 2. *We say that $\mathcal{F}$ is closed under input translation if $f \in \mathcal{F} \implies f_a \in \mathcal{F}$ for all $a \in \mathcal{X}$, where $f_a(x) = f(x + a)$.*

Proposition 1. *Consider the denoising regularizer $R(\theta\,;\sigma)$. Suppose $\mathcal{F}$ is closed under input translation and that, for any $\theta \in \Theta$, there exists $f \in \mathcal{F}$ such that $f(x)$ maps to the prior $p_\theta(z)$ all $x \in \mathcal{X}$. Furthermore, assume that there exists $\theta \in \Theta$ such that $p_\theta(x, z) = p_\theta(z)p_\theta(x)$. Then $R(\theta\,;\sigma_1)$ is stronger $R(\theta\,;\sigma_2)$ when $\sigma_1 \geq \sigma_2$; i.e., $\min_\theta R(\theta\,;\sigma_1) = \min_\theta R(\theta\,;\sigma_2) = 0$ and $R(\theta\,;\sigma_1) \geq R(\theta\,;\sigma_2)$ for all $\theta \in \Theta$.*

Lemma 1 and Proposition 1 show that as we increase σ, the optimal inference model is forced to become smoother and the regularization strength increases. Figure 1 is consistent with this analysis, showing the progression from under-regularized to over-regularized models as we increase σ.

It is worth noting that, in addition to adjusting the denoising regularizer strength via σ, it is also possible to adjust the strength by taking a convex combination of the VAE and DVAE objectives. In particular, we can define the *partially* denoising regularizer $R(\theta\,;\sigma,\alpha)$ as

$$\min_{f\in\mathcal{F}(\mathcal{Q})} \mathbb{E}_{\hat{p}(x)}\Big(\alpha\cdot\mathbb{E}_\varepsilon\left[D(f(x+\varepsilon)\parallel p_\theta(z\mid x))\right] + (1-\alpha)\cdot D(f(x)\parallel p_\theta(z\mid x))\Big) \tag{16}$$

Importantly, we note that $R(\theta\,;\sigma,\alpha)$ is still strictly non-negative and, when combined with the log-likelihood term, still yields a tractable variational lower bound.

3.2 Weight-Normalized Amortized Inference

In addition to DVAE, we propose an alternative method that directly restricts $\mathcal{F}$ to the set of smooth functions. To do so, we consider the case where the inference model is a neural network encoder parameterized by weight matrices $\{W_i\}$ and leverage [18]'s weight normalization technique, which proposes to reparameterize the columns w_i of each weight matrix W as

$$w_i = \frac{v_i}{\|v_i\|}\cdot s_i, \tag{17}$$

where $v_i \in \mathbb{R}^d, s_i \in \mathbb{R}$ are trainable parameters. Since it is possible to modulate the smoothness of the encoder by capping the magnitude of s_i, we introduce a new parameter $u_i \in \mathbb{R}$ and define

$$s_i = \min\left\{\|v_i\|, \left(\frac{H}{1+\exp(-u_i)}\right)\right\}. \tag{18}$$

The norm $\|w_i\|$ is thus bounded by the hyperparameter H. We denote the weight-normalized regularizer as $R(\theta\,;\mathcal{F}_H)$, where $\mathcal{F}_H$ is the amortization family induced by a H-weight-normalized encoder. Under similar assumptions as Proposition 1, it is easy to see that $\min_\theta R(\theta\,;\mathcal{F}_H) = 0$ for any $H \geq 0$ and that $R(\theta\,;\mathcal{F}_{H_1}) \geq R(\theta\,;\mathcal{F}_{H_2})$ for all $\theta \in \Theta$ when $H_1 \leq H_2$ (since $\mathcal{F}_{H_1} \subseteq \mathcal{F}_{H_2}$). We refer to the resulting model as the weight-normalized inference VAE (WNI-VAE) and show in Table 1 that weight-normalized amortized inference can achieve similar performance as DVAE.

3.3 Experiments

We conducted experiments on statically binarized MNIST, statically binarized OMNIGLOT, and the Caltech 101 Silhouettes datasets. These datasets have a relatively small amount of training data and are thus susceptible to model overfitting. For each dataset, we used the same decoder architecture across all four models (VAE, DVAE ($\alpha = 0.5$), DVAE ($\alpha = 1.0$), WNI-VAE) and only modified the encoder, and trained all models using Adam [19] (see Appendix E for more details). To approximate the log-likelihood, we proposed to use importance-weighted stochastic variational inference (IW-SVI), an extension of SVI [20] which we describe in detail in Appendix C. Hyperparameter tuning of DVAE's σ and WNI-VAE's $\mathcal{F}_H$ is described in Table 7.

Table 1 shows the performance of VAE, DVAE, and WNI-VAE. Regularizing the inference model consistently improved the test set log-likelihood performance. On the MNIST and Caltech 101 Silhouettes datasets, the results also show a consistent reduction of the test set inference gap when the inference model is regularized. We observed differences in the performance of DVAE versus WNI-VAE on the Caltech 101 Silhouettes dataset, suggesting a difference in how denoising and weight normalization regularizes the inference model; an interesting consideration would thus be to combine DVAE and WNI. As a whole, Table 1 demonstrates that AIR benefits the generative model.

The denoising and weight normalization regularizers have respective hyperparameters σ and H that control the regularization strength. In Figure 1, we performed an ablation analysis of how adjusting

Table 1: Test set evaluation of VAE, DVAE, and WNI-VAE. The performance metrics are log-likelihood $\ln p_\theta(x)$, the amortized ELBO $\mathcal{L}(x)$, and the inference gap $\Delta_{\text{inf}} = \ln p_\theta(x) - \mathcal{L}(x)$. All three proposed models out-perform VAE across most metrics.

	MNIST			OMNIGLOT			CALTECH		
	$-\ln p_\theta(x)$	Δ_{inf}	$-\mathcal{L}(x)$	$-\ln p_\theta(x)$	Δ_{inf}	$-\mathcal{L}(x)$	$-\ln p_\theta(x)$	Δ_{inf}	$-\mathcal{L}(x)$
VAE	86.93 $\pm$0.04	8.54 $\pm$0.14	95.48 $\pm$0.07	110.32 $\pm$0.16	12.03 $\pm$0.25	122.35 $\pm$0.33	109.14 $\pm$0.28	28.90 $\pm$0.42	138.05 $\pm$0.15
DVAE ($\alpha = 0.5$)	86.46 $\pm$0.02	**6.34** $\pm$0.05	**92.80** $\pm$0.07	109.31 $\pm$0.19	12.56 $\pm$0.18	121.87 $\pm$0.37	**108.64** $\pm$0.19	**23.40** $\pm$0.19	**132.04** $\pm$0.37
DVAE ($\alpha = 1.0$)	86.51 $\pm$0.02	6.83 $\pm$0.04	93.35 $\pm$0.06	110.12 $\pm$0.18	12.44 $\pm$0.16	122.56 $\pm$0.34	108.66 $\pm$0.23	23.94 $\pm$0.15	132.60 $\pm$0.15
WNI-VAE	**86.42** $\pm$0.01	6.68 $\pm$0.01	93.10 $\pm$0.02	**109.16** $\pm$0.12	**11.39** $\pm$0.10	**120.55** $\pm$0.20	108.94 $\pm$0.31	28.88 $\pm$0.29	137.82 $\pm$0.25

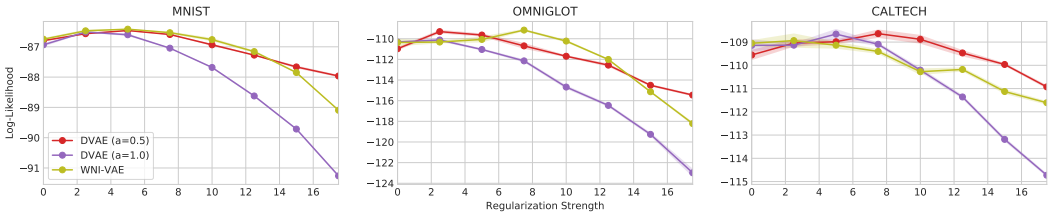


Figure 1: Evaluation of the log-likelihood performance of all three proposed models as we vary the regularization parameter value. The regularization parameter is defined in Table 7. When the parameter value is too small, the model overfits and the test set performance degrades. When the parameter value is too high, the model underfits.

the regularization strength impacts the test set log-likelihood. In almost all cases, we see a transition from overfitting to underfitting as we adjust the strength of AIR. For well-chosen regularization strength, however, it is possible to increase the test set log-likelihood performance by $0.5 \sim 1.0$ nats—a non-trivial improvement.

3.4 How Does Amortized Inference Regularization Affect the Generator?

Table 1 shows that regularizing the inference model empirically benefits the generative model. We now provide some initial theoretical characterization of how a smoothed amortized inference model affects the generative model. Our analysis rests on the following proposition.

Proposition 2. *Let $\mathcal{P}$ be an exponential family with corresponding mean parameter space $\mathcal{M}$ and sufficient statistic function $T(\cdot)$. With a slight abuse of notation, consider $g \in \mathcal{G} : \mathcal{Z} \to \mathcal{M}$. Define $q(x, z) = \hat{p}(x)q(z \mid x)$, where $q(z \mid x)$ is a fixed inference model. Supposing $\mathcal{G}$ has infinite capacity, then the optimal generative model in Eq. (5) returns $g^*(z) = \mu \in \mathcal{M}$, where*

$$\mu = \sum_{i=1}^{n} q(x^{(i)} \mid z) \cdot T(x^{(i)}) = \sum_{i=1}^{n} \left(\frac{q(z \mid x^{(i)})}{\sum_j q(z \mid x^{(j)})} \cdot T(x^{(i)}) \right). \tag{19}$$

Proposition 2 generalizes the analysis in [21] which determined the optimal generative model when $\mathcal{P}$ is Gaussian. The key observation is that the optimal generative model outputs a convex combination of $\{\phi(x^{(i)})\}$, weighted by $q(x^{(i)} \mid z)$. Furthermore, the weights $q(x^{(i)} \mid z)$ are simply density ratios of the proposal distributions $\{q(z \mid x^{(i)})\}$. As we increase the smoothness of the amortized inference model, the weight $q(x^{(i)} \mid z)$ should tend toward $\frac{1}{n}$ for all $z \in \mathcal{Z}$. This suggests that a smoothed inference model provides a natural way to smooth (and thus regularize) the generative model.

4 Amortized Inference Regularization in Importance-Weighted Autoencoders

In this section, we extend AIR to importance-weighted autoencoders (IWAE-k). Although the application is straightforward, we demonstrate a noteworthy relationship between the number of importance samples k and the effect of AIR. To begin our analysis, we consider the IWAE-k objective

$$\max_{\theta,\phi} \mathbb{E}_{z_1 \ldots z_k \sim q_\phi(z|x)} \left[\ln \frac{1}{k} \sum_{i=1}^{k} \frac{p_\theta(x, z_i)}{q_\phi(z_i \mid x)} \right], \tag{20}$$

where $\{z_1 \dots z_k\}$ are k samples from the proposal distribution $q_\phi(z \mid x)$ to be used as importance-samples. Analysis by [22] allows us to rewrite it as a regularized maximum likelihood objective

$$\max_\theta \mathbb{E}_{\hat{p}(x)} \left[\ln p_\theta(x)\right] - \overbrace{\min_{f \in \mathcal{F}(\mathcal{Q})} \mathbb{E}_{\hat{p}(x)} \mathbb{E}_{z_2 \dots z_k \sim f(x)} \tilde{D}(\tilde{f}_k(x, z_1 \dots z_k) \parallel p_\theta(z \mid x))}^{R_k(\theta)}, \tag{21}$$

where $\tilde{f}_k$ (or equivalently $\tilde{q}_k$) is the unnormalized distribution

$$\tilde{f}_k(x, z_2 \dots z_k)(z_1) = \frac{p_\theta(x, z_1)}{\frac{1}{k}\sum_i \frac{p_\theta(x, z_i)}{f(x)(z_i)}} = \tilde{q}_k(z_1 \mid x, z_2 \dots z_k) \tag{22}$$

and $\tilde{D}(q \parallel p) = \int q(z) \left[\ln q(z) - \ln p(z)\right] dz$ is the Kullback-Leibler divergence extended to unnormalized distributions. For notational simplicity, we omit the dependency of $\tilde{f}_k$ on $(z_2 \dots z_k)$. Importantly, [22] showed that the IWAE with k importance samples drawn from the amortized inference model f is, on expectation, equivalent to a VAE with 1 importance sample drawn from the more expressive inference model $\tilde{f}_k$.

4.1 Importance Sampling Attenuates Amortized Inference Regularization

We now consider the interaction between importance sampling and AIR. We introduce the regularizer $R_k(\theta\,; \sigma, \mathcal{F}_H)$ as follows

$$R_k(\theta\,; \sigma, \mathcal{F}_H) = \min_{f \in \mathcal{F}_H(\mathcal{Q})} \mathbb{E}_{\hat{p}(x)} \mathbb{E}_\varepsilon \mathbb{E}_{z_2 \dots z_k \sim f(x+\varepsilon)} \tilde{D}(\tilde{f}_k(x + \varepsilon) \parallel p_\theta(z \mid x)), \tag{23}$$

which corresponds to a regularizer where weight normalization, denoising, and importance sampling are simultaneously applied. By adapting Theorem 1 from [8], we can show that

Proposition 3. *Consider the regularizer $R_k(\theta\,; \sigma, \mathcal{F}_H)$. Under similar assumptions as Proposition 1, then R_{k_1} is stronger than R_{k_2} when $k_1 \leq k_2$; i.e., $\min_\theta R_{k_1}(\theta\,; \sigma, \mathcal{F}_H) = \min_\theta R_{k_2}(\theta\,; \sigma, \mathcal{F}_H) = 0$ and $R_{k_1}(\theta\,; \sigma, \mathcal{F}_H) \leq R_{k_2}(\theta\,; \sigma, \mathcal{F}_H)$ for all $\theta \in \Theta$.*

A notable consequence of Proposition 3 is that as k increases, AIR exhibits a weaker regularizing effect on the posterior distributions $\{p_\theta(z \mid x^{(i)})\}$. Intuitively, this arises from the phenomenon that although AIR is applied to f, the subsequent importance-weighting procedure can still create a flexible $\tilde{f}_k$. Our analysis thus predicts that AIR is less likely to cause *underfitting* of IWAE-k's generative model as k increases, which we demonstrate in Figure 2. In the limit of infinite importance samples, we also predict AIR to have zero regularizing effect since $\tilde{f}_\infty$ (under some assumptions) can always approximate any posterior. However, for practically feasible values of k, we show in Tables 2 and 3 that AIR is a highly effective regularizer.

4.2 Experiments

Table 2: Test set evaluation of the four models when trained with 8 importance samples. $\mathcal{L}_8(x)$ denotes the amortized ELBO using 8 importance samples. $\Delta_{\text{inf}} = \ln p_\theta(x) - \mathcal{L}_8(x)$.

	MNIST			OMNIGLOT			CALTECH		
	$-\ln p_\theta(x)$	Δ_{inf}	$-\mathcal{L}_8(x)$	$-\ln p_\theta(x)$	Δ_{inf}	$-\mathcal{L}_8(x)$	$-\ln p_\theta(x)$	Δ_{inf}	$-\mathcal{L}_8(x)$
IWAE	86.21 $\pm$0.01	6.13 $\pm$0.03	92.34 $\pm$0.02	108.18 $\pm$0.24	8.69 $\pm$0.39	116.87 $\pm$0.16	108.65 $\pm$0.11	21.52 $\pm$0.13	130.17 $\pm$0.09
DIWAE ($\alpha = 0.5$)	**85.78** $\pm$0.02	4.47 $\pm$0.02	90.25 $\pm$0.03	**107.01** $\pm$0.11	8.64 $\pm$0.07	**115.66** $\pm$0.17	**107.34** $\pm$0.17	17.61 $\pm$0.18	124.96 $\pm$0.14
DIWAE ($\alpha = 1.0$)	**85.78** $\pm$0.03	**4.21** $\pm$0.03	**90.00** $\pm$0.06	107.47 $\pm$0.06	**8.57** $\pm$0.14	116.04 $\pm$0.18	107.54 $\pm$0.11	**17.06** $\pm$0.35	**124.60** $\pm$0.29
WNI-IWAE	85.81 $\pm$0.01	4.33 $\pm$0.03	90.14 $\pm$0.04	107.15 $\pm$0.08	8.78 $\pm$0.17	115.93 $\pm$0.10	107.98 $\pm$0.19	22.18 $\pm$0.33	130.16 $\pm$0.14

Table 3: Test set evaluation of the four models when trained with 64 importance samples. $\Delta_{\text{inf}} = \ln p_\theta(x) - \mathcal{L}_{64}(x)$.

	MNIST			OMNIGLOT			CALTECH		
	$-\ln p_\theta(x)$	Δ_{inf}	$-\mathcal{L}_{64}(x)$	$-\ln p_\theta(x)$	Δ_{inf}	$-\mathcal{L}_{64}(x)$	$-\ln p_\theta(x)$	Δ_{inf}	$-\mathcal{L}_{64}(x)$
IWAE	86.06 $\pm$0.03	4.41 $\pm$0.10	90.48 $\pm$0.07	107.31 $\pm$0.14	**6.66** $\pm$0.22	113.97 $\pm$0.10	108.89 $\pm$0.35	16.51 $\pm$0.32	125.40 $\pm$0.25
DIWAE ($\alpha = 0.5$)	**85.55** $\pm$0.02	**3.01** $\pm$0.01	**88.56** $\pm$0.02	**106.02** $\pm$0.01	6.98 $\pm$0.06	113.00 $\pm$0.07	**106.94** $\pm$0.11	**12.28** $\pm$0.14	**119.22** $\pm$0.11
DIWAE ($\alpha = 1.0$)	**85.55** $\pm$0.02	3.15 $\pm$0.02	88.70 $\pm$0.04	106.15 $\pm$0.03	6.70 $\pm$0.05	**112.85** $\pm$0.07	106.96 $\pm$0.11	12.94 $\pm$0.22	119.87 $\pm$0.16
WNI-IWAE	85.64 $\pm$0.03	3.10 $\pm$0.01	88.74 $\pm$0.03	106.17 $\pm$0.07	7.11 $\pm$0.07	113.28 $\pm$0.13	108.15 $\pm$0.11	14.42 $\pm$0.20	122.57 $\pm$0.10

Tables 2 and 3 extends the model evaluation to IWAE-8 and IWAE-64. We see that the denoising IWAE (DIWAE) and weight-normalized inference IWAE (WNI-IWAE) consistently out-perform the standard IWAE on test set log-likelihood evaluations. Furthermore, the regularized models frequently reduced the inference gap as well. Our results demonstrate that AIR is a highly effective regularizer even when a large number of importance samples are used.

Our main experimental contribution in this section is the verification that increasing the number of importance samples results in less underfitting when the inference model is over-regularized. In contrast to $k = 1$, where aggressively increasing the regularization strength can cause considerable underfitting, Figure 2 shows that increasing the number of importance samples to $k = 8$ and $k = 64$ makes the models much more robust to mis-specified choices of regularization strength. Interestingly, we also observed that the optimal regularization strength (determined using the validation set) increases with k (see Table 7 for details). The robustness of importance sampling when paired with amortized inference regularization makes AIR an effective and practical way to regularize IWAE.

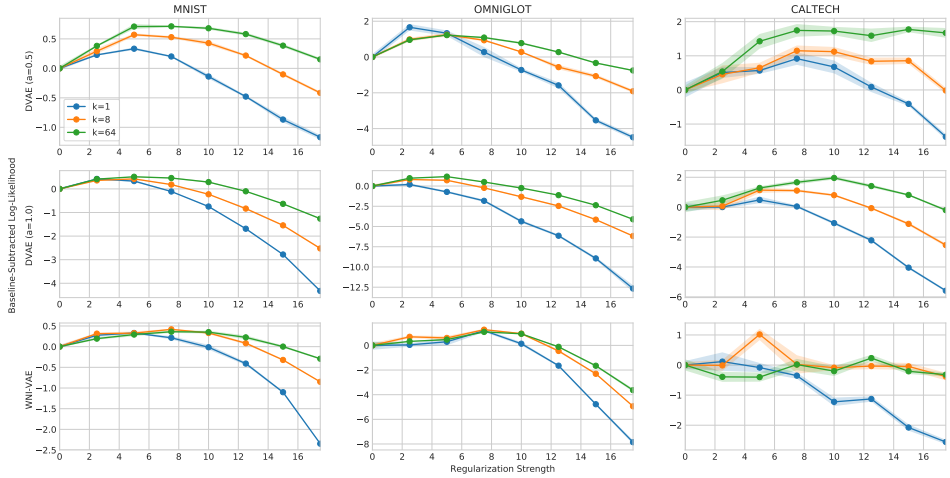


Figure 2: Evaluation of the log-likelihood performance of all three proposed models as we vary the regularization parameter (see Table 7 for definition) and number of importance samples k. To compare across different k's, the performance without regularization (IWAE-k baseline) is subtracted. We see that IWAE-64 is the least likely to underfit when the regularization parameter value is high.

4.3 Are High Signal-to-Noise Ratio Gradients Necessarily Better?

We note the existence of a related work [23] that also concluded that approximating maximum likelihood training is not necessarily better. However, [23] focused on increasing the signal-to-noise ratio of the gradient updates and analyzed the trade-off between importance sampling and Monte Carlo sampling under budgetary constraints. An in-depth discussion of these two works within the context of generalization is provided in Appendix D.

5 Conclusion

In this paper, we challenged the conventional role that amortized inference plays in training deep generative models. In addition to expediting variational inference, amortized inference introduces new ways to regularize maximum likelihood training. We considered a special case of amortized inference regularization (AIR) where the inference model must learn a smoothed mapping from $\mathcal{X} \to \mathcal{Q}$ and showed that the denoising variational autoencoder (DVAE) and weight-normalized inference (WNI) are effective instantiations of AIR. Promising directions for future work include replacing denoising with adversarial training [24] and weight normalization with spectral normalization [25]. Furthermore, we demonstrated that AIR plays a crucial role in the regularization of IWAE, and that higher levels of regularization may be necessary due to the attenuating effects of importance sampling on AIR. We believe that variational family expansion by Monte Carlo methods [26] may exhibit the same attenuating effect on AIR and recommend this as an additional research direction.

Acknowledgements

This research was supported by TRI, NSF (#1651565, #1522054, #1733686), ONR, Sony, and FLI. Toyota Research Institute provided funds to assist the authors with their research but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity.

References

[1] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[2] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-Supervised Learning With Deep Generative Models. In *Advances In Neural Information Processing Systems*, pages 3581–3589, 2014.

[3] Hyunjik Kim and Andriy Mnih. Disentangling By Factorising. *arXiv preprint arXiv:1802.05983*, 2018.

[4] Tian Qi Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating Sources Of Disentanglement In Variational Autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

[5] Yoon Kim, Sam Wiseman, Andrew C Miller, David Sontag, and Alexander M Rush. Semi-Amortized Variational Autoencoders. *arXiv preprint arXiv:1802.02550*, 2018.

[6] Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved Variational Inference With Inverse Autoregressive Flow. In *Advances In Neural Information Processing Systems*, pages 4743–4751, 2016.

[7] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder Variational Autoencoders. In *Advances In Neural Information Processing Systems*, pages 3738–3746, 2016.

[8] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance Weighted Autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[9] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation And Approximate Inference In Deep Generative Models. *arXiv preprint arXiv:1401.4082*, 2014.

[10] Chris Cremer, Xuechen Li, and David Duvenaud. Inference Suboptimality In Variational Autoencoders. *arXiv preprint arXiv:1801.03558*, 2018.

[11] Yuhuai Wu, Yuri Burda, Ruslan Salakhutdinov, and Roger Grosse. On The Quantitative Analysis Of Decoder-Based Generative Models. *arXiv preprint arXiv:1611.04273*, 2016.

[12] Rahul G Krishnan, Dawen Liang, and Matthew Hoffman. On the challenges of learning with inference networks on sparse, high-dimensional data. *arXiv preprint arXiv:1710.06085*, 2017.

[13] Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. Auxiliary Deep Generative Models. *arXiv preprint arXiv:1602.05473*, 2016.

[14] Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical Variational Models. In *International Conference On Machine Learning*, pages 324–333, 2016.

[15] Kuzman Ganchev, Jennifer Gillenwater, Ben Taskar, et al. Posterior Regularization For Structured Latent Variable Models. *Journal of Machine Learning Research*, 11(Jul):2001–2049, 2010.

[16] Jun Zhu, Ning Chen, and Eric P Xing. Bayesian Inference With Posterior Regularization And Applications To Infinite Latent Svms. *The Journal of Machine Learning Research*, 15(1):1799–1847, 2014.

[17] Daniel Jiwoong Im, Sungjin Ahn, Roland Memisevic, Yoshua Bengio, et al. Denoising Criterion For Variational Auto-Encoding Framework. In *AAAI*, pages 2059–2065, 2017.

[18] Tim Salimans and Diederik P Kingma. Weight Normalization: A Simple Reparameterization To Accelerate Training Of Deep Neural Networks. In *Advances In Neural Information Processing Systems*, pages 901–909, 2016.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method For Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic Variational Inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[21] Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl-Johann Simon-Gabriel, and Bernhard Schoelkopf. From Optimal Transport To Generative Modeling: The VEGAN Cookbook. *arXiv preprint arXiv:1705.07642*, 2017.

[22] Chris Cremer, Quaid Morris, and David Duvenaud. Reinterpreting Importance-Weighted Autoencoders. *arXiv preprint arXiv:1704.02916*, 2017.

[23] Tom Rainforth, Adam R Kosiorek, Tuan Anh Le, Chris J Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter Variational Bounds Are Not Necessarily Better. *arXiv preprint arXiv:1802.04537*, 2018.

[24] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining And Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572*, 2014.

[25] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral Normalization For Generative Adversarial Networks. *arXiv preprint arXiv:1802.05957*, 2018.

[26] Matthew D Hoffman. Learning Deep Latent Gaussian Models With Markov Chain Monte Carlo. In *International Conference On Machine Learning*, pages 1510–1519, 2017.

[27] Yingzhen Li and Richard E Turner. Rényi Divergence Variational Inference. In *Advances In Neural Information Processing Systems*, pages 1073–1081, 2016.

[28] Jakub M Tomczak and Max Welling. VAE With A Vampprior. *arXiv preprint arXiv:1705.07120*, 2017.

[29] Samuel L. Smith and Quoc V. Le. A bayesian Perspective On Generalization And Stochastic Gradient Descent. In *International Conference On Learning Representations*, 2018.

[30] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp Minima Can Generalize For Deep Nets. *arXiv preprint arXiv:1703.04933*, 2017.

[31] Dominic Masters and Carlo Luschi. Revisiting Small Batch Training For Deep Neural Networks. *arXiv preprint arXiv:1804.07612*, 2018.

[32] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding Deep Learning Requires Rethinking Generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[33] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering With Bregman Divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

A Overly Expressive Amortization Family Hurts Generalization

In the experiments by [10], they observed that an overly expressive amortization family increases the test set inference gap, but does not impact the test set log-likelihood. We show in Table 4 that [10]'s observation is not true in general, and that an overly expressive amortization family can in fact hurt test set log-likelihood. Details regarding the architectures are provided in Appendix E.

Table 4: Performance evaluation when an over-expressive amortization family is used (i.e. a larger encoder). Comparison is made against models that use a smaller encoder. The results show that using a large encoder consistently hurts generalization by over 1 nat.

	MNIST ($k=1$)			MNIST ($k=8$)			MNIST ($k=64$)		
	$-\ln p_\theta(x)$	Δ_{inf}	$-\mathcal{L}_1(x)$	$-\ln p_\theta(x)$	Δ_{inf}	$-\mathcal{L}_8(x)$	$-\ln p_\theta(x)$	Δ_{inf}	$-\mathcal{L}_64(x)$
IWAE (Large Encoder)	87.43 ± 0.05	11.32 ± 0.21	98.74 ± 0.25	86.98 ± 0.07	8.00 ± 0.18	94.98 ± 0.17	86.70 ± 0.06	5.91 ± 0.11	92.61 ± 0.10
IWAE	86.93 ± 0.04	8.54 ± 0.14	95.48 ± 0.07	86.21 ± 0.01	6.13 ± 0.03	92.34 ± 0.02	86.06 ± 0.03	4.41 ± 0.10	90.48 ± 0.07
DIWAE ($\alpha = 0.5$)	86.46 ± 0.02	**6.34** ± 0.05	**92.80** ± 0.07	**85.78** ± 0.02	4.47 ± 0.02	90.25 ± 0.03	**85.55** ± 0.02	**3.01** ± 0.01	**88.56** ± 0.02
DIWAE ($\alpha = 1.0$)	86.51 ± 0.02	6.83 ± 0.04	93.35 ± 0.06	**85.78** ± 0.03	**4.21** ± 0.03	**90.00** ± 0.06	**85.55** ± 0.02	3.15 ± 0.02	88.70 ± 0.04
WNI-IWAE	**86.42** ± 0.01	6.68 ± 0.01	93.10 ± 0.02	85.81 ± 0.01	4.33 ± 0.03	90.14 ± 0.04	85.64 ± 0.03	3.10 ± 0.01	88.74 ± 0.03

B Revisiting [17]'s Denoising Variational Autoencoder Analysis

In [17]'s Lemma 1, they considered a joint distribution $p_\theta(x, z)$. They introduced an auxiliary variable z' into their inference model (here z' takes on the role of the perturbed input $\tilde{x} = x + \varepsilon$. To avoid confusion, we stick to the notation used in their Lemma) and considered the inference model

$$q_\varphi(z \mid z')q_\psi(z' \mid x). \tag{24}$$

They considered two ways to use this inference model. The first approach is to marginalize the auxiliary latent variable z'. This defines the resulting inference model

$$q_\phi(z \mid x) = \int q_\varphi(z \mid z')q_\psi(z' \mid x)dz'. \tag{25}$$

This yields the lower bound

$$\mathcal{L}_a = \mathbb{E}_{q_\phi(z|x)}\left[\ln \frac{p_\theta(x,z)}{q_\phi(z \mid x)}\right]. \tag{26}$$

Next, they considered an alternative lower bound

$$\mathcal{L}_b = \mathbb{E}_{q_\varphi(z|z')q_\psi(z'|x)}\left[\ln \frac{p_\theta(x,z)}{q_\varphi(z \mid z')}\right]. \tag{27}$$

[17]'s Lemma 1 claims that

1. $\mathcal{L}_a$ and $\mathcal{L}_b$ are valid lower bounds of $\ln p_\theta(x)$
2. $\mathcal{L}_b \geq \mathcal{L}_a$.

Using Lemma 1, [17] motivated the denoising variational autoencoder by concluding that it provides a tighter bound than marginalization of the noise variable. Although statement 1 is correct, statement 2 is not. Their proof of statement 2 is presented as follows

$$\begin{aligned}
\mathbb{E}_{q_\varphi(z|z')q_\psi(z'|x)}\left[\ln \frac{p_\theta(x,z)}{q_\varphi(z \mid z')}\right] &\overset{?}{=} \mathbb{E}_{q_\phi(z|x)}\left[\ln \frac{p_\theta(x,z)}{q_\varphi(z \mid z')}\right] && (28)\\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x,z)\right] - \mathbb{E}_{q_\phi(z|x)}\left[\ln q_\phi(z \mid z')\right] && (29)\\
&\geq \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x,z)\right] - \mathbb{E}_{q_\phi(z|x)}\left[\ln q_\phi(z \mid z')\right] && (30)\\
&= \mathbb{E}_{q_\phi(z|x)}\left[\ln \frac{p_\theta(x,z)}{q_\phi(z \mid x)}\right] && (31)
\end{aligned}$$

We indicate the mistake with $\overset{?}{=}$; their proof of statement 2 relied on the assumption that

$$\mathbb{E}_{q_\varphi(z|z')q_\psi(z'|x)}\left[\ln \frac{p_\theta(x,z)}{q_\varphi(z \mid z')}\right] = \mathbb{E}_{q_\phi(z|x)}\left[\ln \frac{p_\theta(x,z)}{q_\varphi(z \mid z')}\right]. \tag{32}$$

Crucially, the RHS is ill-defined since it does not take the expectation over z', whereas the LHS explicitly specifies an expectation over $z' \sim q_\psi(z' \mid x)$. This difference, while subtle, invalidates the subsequent steps. If we fix Eq. (28) and attempt to see if the rest of the proof still follows, we will find that

$$\mathbb{E}_{q_\varphi(z|z')q_\psi(z'|x)}\left[\ln \frac{p_\theta(x,z)}{q_\varphi(z \mid z')}\right] = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x,z)\right] - \mathbb{E}_{q_\varphi(z|z')q_\psi(z'|x)}\left[\ln q_\psi(z \mid z')\right] \tag{33}$$

$$\not\geq \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x,z)\right] - \mathbb{E}_{q_\varphi(z|z')q_\psi(z'|x)}\left[\ln q_\phi(z \mid x)\right] \tag{34}$$

$$= \mathbb{E}_{q_\phi(z|x)}\left[\ln \frac{p_\theta(x,z)}{q_\phi(z \mid x)}\right]. \tag{35}$$

Indeed, the inequality will point the other way since

$$\mathbb{E}_{q_\varphi(z|z')q_\psi(z'|x)}\left[\ln q_\psi(z \mid z') - \ln q_\phi(z \mid x)\right] = \mathbb{E}_{q_\psi(z'|x)}\mathbb{E}_{q_\varphi(z|z')} \ln \frac{q_\varphi(z \mid z')}{q_\phi(z \mid x)} \tag{36}$$

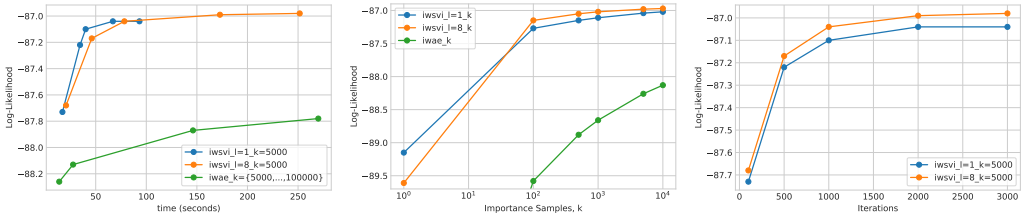
$$= \mathbb{E}_{q_\psi(z'|x)} D(q_\varphi(z \mid z') \parallel q_\phi(z \mid x)) \tag{37}$$

$$\geq 0 \implies \tag{38}$$

$$-\mathbb{E}_{q_\varphi(z|z')q_\psi(z'|x)}\left[\ln q_\psi(z \mid z')\right] \leq -\mathbb{E}_{q_\varphi(z|z')q_\psi(z'|x)}\left[\ln q_\phi(z \mid x)\right]. \tag{39}$$

Their conclusion that marginalizing over the noise variable results in a looser bound is thus incorrect. In the text (beneath [17] Eq. (11)), they further implied that the denoising VAE and standard VAE objectives are not comparable. We show in Proposition 1 that the denoising VAE objective is in fact a lower bound of the standard VAE objective.

C Importance-Weighted Stochastic Variational Inference



(a) (b) (c)

Figure 3: Evaluation of IW-SVI versus IWAE-k for a fixed generative model. IW-SVI out-performss IWAE-k on both computation time and number of importance samples needed. Similar to [11], we conclude that IWAE-k's poor approximation of the log-likelihood is attributable to an overfit amortized inference model. Fig. 3a) IW-SVI computation time depends on the number of gradient update steps. IWAE-k computation time depends on the number of importance samples k. IWAE-100000 still under-performs IW-SVI ($k = 5000, \ell = 1, T = 100$), demonstrating the efficacy of IW-SVI. Fig. 3b) Comparison of IWAE and IW-SVI ($T = 3000$) for different values of k. Fig. 3c Comparison of IW-SVI ($k = 5000$) for different values of T.

We propose a simple method to approximate the marginal $\ln p_\theta(x)$. A common approach for approximating the log marginal is the IWAE-5000 [7, 8, 27, 28], which proposes to compute $\mathcal{L}_{5000}(x; \theta, \phi)$ where

$$\ln p_\theta(x) \geq \mathcal{L}_k(x\,; \theta, \phi) = \mathbb{E}_{z_1 \ldots z_k \sim q_\phi(z|x)}\left(\ln \frac{1}{k}\sum_{i=1}^{k} \frac{p_\theta(x, z^{(i)})}{q_\phi(z^{(i)} \mid x)}\right). \tag{40}$$

However, this approach relies on the learned inference model $q_\phi(z \mid x)$, which might overfit to the training set. To address this issue, we propose to perform importance-weighted stochastic variational

inference (IW-SVI)

$$\ln p_\theta(x) \geq \mathcal{L}_k(x\,;\theta,q^*_{x,\ell}) = \mathbb{E}_{z_1\ldots z_k\sim q^*_{x,\ell}(z)}\left(\ln\frac{1}{k}\sum_{i=1}^{k}\frac{p_\theta(z\mid x)}{q^*_{x,\ell}(z)}\right), \tag{41}$$

$$\text{where } q^*_{x,\ell} = \arg\max_{q\in\mathcal{Q}} \mathcal{L}_\ell(x\,;\theta,q). \tag{42}$$

The optimization in Eq. (42) is approximate with T gradient steps. As k and ℓ increase, the approximation will approach the true log-likelihood. We approximate log-likelihood over the entire test set using $\mathbb{E}_{\hat{p}_{\text{test}}(x)}\mathcal{L}_k(p_\theta, q^*_{x,\ell}\,; x)$. To reduce speed and memory cost during the per-sample optimization in Eq. (42), we use a large $k = 5000$ but smaller $\ell = 8$, and approximately solved the optimization problem using $T = 3000$ gradient steps. In comparison to IWAE-5000, we consistently observe significant improvement in the log-likelihood approximation. IW-SVI provides a simple alternative to Annealed Importance Sampling, requiring minimal modification to any existing IWAE-k implementation.

D Are High Signal-to-Noise Ratio Gradients Necessarily Better?

Our paper shares a similar high level message with a recent study by [23]: that approximating maximum likelihood training is not necessarily better. However, we approach this message in very different ways. [23] observed that importance sampling weakens the signal-to-noise ratio of the gradients used to update the amortized inference model. In response, they proposed to increase this ratio by increasing the number of Monte Carlo samples m used to estimate the expectation in Eq. (5). Under a fixed budget of $T \geq mk$ (where k is the number of importance samples and m is the number of Monte Carlo samples), they observed that it may be desirable to trade off k in order to increase m. Given an infinite budget, however, [23]'s hypothesis would still conclude to increase k as much as possible in order to approximate maximum likelihood training.

In contrast, we argue that it may be inherently desirable to regularize the maximum likelihood objective, and that amortized inference regularization is an effective means of doing so. From the perspective of generalization, it is also worth wondering whether high signal-to-noise ratio gradients are necessarily better. The desirability of noisy gradients for improving generalization is an active area of research [29, 30, 31, 32], and an extensive investigation of the role of gradient stochasticity in regularizing the amortized inference model is beyond the scope of our paper. To encourage future exploration in this direction, we show in Figure 4 that the effect of gradient stochasticity is non-negligible. For the standard VAE, we observed that increasing m can cause the model to overfit (on the amortized ELBO objective) over the course of training. Interestingly, we observed that DVAE does not experience this overfitting effect, suggesting that AIR is robust to larger values of m.

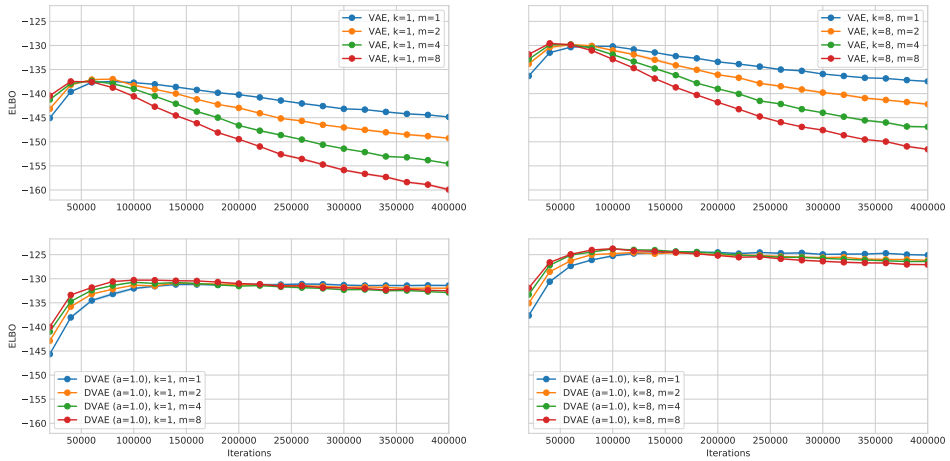


Figure 4: Comparison of the test set amortized ELBO during training for VAE and DVAE as we vary the number of importance samples k and the number of Monte Carlo samples m. In contrast to DVAE, VAE is susceptible to overfitting when m is increased.

E Experimental Details

Datasets. We carried out experiments on the static MNIST, static OMNIGLOT, and Caltech 101 Silhouettes datasets. OMNIGLOT was statically binarized at the beginning of training via random sampling using the pixel real-values as Bernoulli parameters. Training, validation, and test split sizes are provided in Table 5. The MNIST validation set was created by randomly holding out 10000 samples from the original 60000-sample training set. The OMNIGLOT validation set was similarly created by randomly holding out 1345 samples from the original 24345-sample training set.

Table 5: Training, validation and test splits for each dataset.

Dataset	Training Split	Validation Split	Test Split
MNIST	50000	10000	10000
OMNIGLOT	23000	1345	8070
CALTECH	4100	2264	2307

Training parameters. Important training parameters are provided in Table 6. We used the Adam optimizer and exponentially decayed the initial learning rate according to the formula

$$\alpha_t = \alpha_0 \cdot (0.1)^{\frac{t}{T-1}}, \tag{43}$$

where $t \in \{0, \ldots, T-1\}$ is the current iteration and T is the total number of iterations. Early-stopping is applied according to IWAE-5000 evaluation on the validation set.

Table 6: Training parameters used for each dataset. The same architecture is used for all models, with minor modification for WNI-VAE (to account for the weight-normalization implementation). In all cases, we use a Bernoulli decoder and a Gaussian encoder. Notation: d300 denotes a dense layer with ELU activation and 300 output units. z64 denotes 1) a dense layer with 64 output units (represents the mean of z) and 2) a dense layer with softplus activation and 64 output units (represents the variance of z). x784 denotes a dense layer with 784 output units (represents the logits for x)

Dataset	Encoder Architecture	Decoder Architecture	Initial Learning Rate	Training Iterations	Batch Size
MNIST (Appendix A)	d1000-d1000-d1000-z64	d300-d300-x784	10^{-3}	1.5×10^6	100
MNIST	d300-d300-z64	d300-d300-x784	10^{-3}	1.5×10^6	100
OMNIGLOT	d200-d200-z64	d200-d200-x784	10^{-3}	1.5×10^6	100
CALTECH	d500-z64	d500-x784	10^{-4}	4×10^5	10

Regularization strength tuning. The denoising and weight normalization regularizers have hyperparameters σ and H respectively. See Table 7 for hyperparameter search space details. We performed a basic grid search and tuned the regularization strength hyperparameters based on the validation set.

Table 7: The regularization parameter is chosen applied based on hyperparameter tuning on the validation set. Rather than selecting for σ or H directly, we reparameterized the search space as $\sigma \cdot \sqrt{d}$ and $\frac{10}{H}$, where d denotes the sample dimensionality, i.e., $\mathcal{X} = \mathbb{R}^d$. Coincidentally, we found that this reparameterization allowed us to use the same search space for both DIWAE and WNI-IWAE. We introduce the convention that setting $\frac{10}{H}$ to zero indicates setting $H = +\infty$. Via this convention, setting $\sigma \cdot \sqrt{d} = \frac{10}{H} = 0$ corresponds to the standard VAE. We restricted the search space to the set $\{2.5, 5.0, \ldots, 17.5\}$, deliberately omitting $\{0.0\}$ to not encompass the baseline (standard VAE).

		MNIST		OMNIGLOT		CALTECH	
	k	$\sigma \cdot \sqrt{d}$	$\frac{10}{H}$	$\sigma \cdot \sqrt{d}$	$\frac{10}{H}$	$\sigma \cdot \sqrt{d}$	$\frac{10}{H}$
DIWAE ($\alpha = 0.5$)	1	5.0	-	2.5	-	7.5	-
	8	5.0	-	5.0	-	7.5	-
	64	7.5	-	5.0	-	15.0	-
DIWAE ($\alpha = 1.0$)	1	2.5	-	2.5	-	5.0	-
	8	5.0	-	5.0	-	7.5	-
	64	5.0	-	5.0	-	10.0	-
WNI-IWAE	1	-	5.0	-	7.5	-	2.5
	8	-	7.5	-	7.5	-	5.0
	64	-	10.0	-	7.5	-	12.5

F Proofs

Remark. Some of the proofs mention the notion of an infinite capacity $\mathcal{F}$, $\mathcal{G}$ or $\mathcal{Q}$. To clarify, we say that $\mathcal{F}$ has infinite capacity if it is the set of all possible functions that map from $\mathcal{X}$ to $\mathcal{Q}$. Analogously, $\mathcal{G}$ has infinite capacity if it is the set of all possible functions that map from $\mathcal{Z}$ to $\mathcal{P}$. We say that $\mathcal{Q}$ has infinite capacity if it is the set of all possible distributions over the space $\mathcal{Z}$.

Lemma 1. *For fixed $(\theta, \sigma, \mathcal{Q})$ and infinite capacity $\mathcal{F}$, the inference model that optimizes the DVAE objective in Eq. (13) is the kernel regression model*

$$f_\sigma^*(x) = \arg\min_{q \in \mathcal{Q}} \sum_{i=1}^{n} w_\sigma(x, x^{(i)}) \cdot D(q(z) \parallel p_\theta(z \mid x^{(i)})), \tag{14}$$

where $w_\sigma(x, x^{(i)}) = \frac{K_\sigma(x, x^{(i)})}{\sum_j K_\sigma(x, x^{(j)})}$ and $K_\sigma(x, y) = \exp\left(-\frac{\|x-y\|}{2\sigma^2}\right)$ is the RBF kernel.

Proof. Define $\tilde{x} = x + \varepsilon$ and $\hat{p}(x, \tilde{x}) = \hat{p}(x)\mathcal{N}(\tilde{x} \mid x, \sigma\mathbf{I})$. Rewrite the objective as

$$R(\theta\,; \sigma) = \min_{f \in \mathcal{F}(\mathcal{Q})} \mathbb{E}_{\hat{p}(x,\tilde{x})} \left[D(f(\tilde{x}) \parallel p_\theta(z \mid x)) \right] \tag{44}$$

$$\geq \mathbb{E}_{\hat{p}(\tilde{x})} \min_{q \in \mathcal{Q}} \mathbb{E}_{\hat{p}(x \mid \tilde{x})} \left[D(q(z) \parallel p_\theta(z \mid x)) \right]. \tag{45}$$

Recall that $\mathcal{F}$ has infinite capacity. This lower bound is tight since we can select $f_\sigma^* \in \mathcal{F}$ such that

$$f_\sigma^*(\tilde{x}) = \arg\min_{q \in \mathcal{Q}} \mathbb{E}_{\hat{p}(x \mid \tilde{x})} D(q(z) \parallel p_\theta(z \mid x)). \tag{46}$$

Reexpressing Eq. (46) by expanding $\hat{p}(x \mid \tilde{x})$ yields Eq. (14). □

Theorem 1. *Let $\mathcal{Q}$ be a minimal exponential family with corresponding natural parameter space Ω. With a slight abuse of notation, consider $f \in \mathcal{F} : \mathcal{X} \to \Omega$. Under the simplifying assumption that $p_\theta(z \mid x^{(i)})$ is contained within $\mathcal{Q}$ and parameterized by $\eta^{(i)} \in \Omega$, and that $\mathcal{F}$ has infinite capacity, then the optimal inference model in Lemma 1 returns $f_\sigma^*(x) = \eta \in \Omega$, where*

$$\eta = \sum_{i=1}^{n} w_\sigma(x, x^{(i)}) \cdot \eta^{(i)} \tag{15}$$

and Lipschitz constant of f_σ^* is bounded by $O(1/\sigma^2)$.

Proof. Proof provided in two parts.

Part 1. The Kullback-Leibler divergence can be represented as a Bregman divergence associated with the strictly convex log-partition function A of the minimal exponential family as follows

$$D(\eta \parallel \eta^{(i)}) = d_A(\eta^{(i)}, \eta) = A(\eta^{(i)}) - A(\eta) - \nabla A(\eta)^\top (\eta^{(i)} - \eta). \tag{47}$$

Proposition 1 from [33] shows that that for any convex combination weights $\{w_i\}, \sum_{i=1}^n w_i = 1$, the minimizer of a weighted average of Bregman divergences is

$$\sum_{i=1}^{n} w_i x_i = \arg\min_{y \in \Omega} \sum_{i=1}^{n} w_i d_A(x_i, y). \tag{48}$$

It thus follows that

$$f_\sigma^*(x) = \arg\min_{\eta \in \Omega} \sum_{i=1}^{n} w_\sigma(x, x^{(i)}) \cdot D(\eta \parallel \eta^{(i)}) \tag{49}$$

$$= \arg\min_{\eta \in \Omega} \sum_{i=1}^{n} w_\sigma(x, x^{(i)}) \cdot d_A(\eta^{(i)}, \eta) \tag{50}$$

$$= \sum_{i=1}^{n} w_\sigma(x, x^{(i)}) \cdot \eta^{(i)}. \tag{51}$$

Part 2. We will write the derivative $\nabla_x f^*_\sigma(x)$ in matrix form by the following notation

$$\nabla_x W_\sigma(x) = \begin{pmatrix} \nabla_x w_\sigma(x, x^{(1)}) & \cdots & \nabla_x w_\sigma(x, x^{(m)}) \end{pmatrix}$$
$$M = \begin{pmatrix} \eta^{(1)} & \cdots & \eta^{(m)} \end{pmatrix}$$

where we also suppose input space x is n-dimensional, latent parameter space Ω is d-dimensional, and there are m training examples. Then

$$\nabla_x f^*_\sigma(x) = M \nabla_x W_\sigma(x)^T$$

Let $\|\cdot\|_1$ be the induced 1-norm for matrices, then by the sub-multiplicative property

$$\|\nabla_x f^*(x)\|_1 \leq \|M\|_1 \|\nabla_x W_\sigma(x)^T\|_1$$

Since $\|M\|_1$ is a constant with respect to σ, we only have to bound $\|\nabla_x W_\sigma(x)^T\|_1$. To do this we study the derivative of $\nabla_x w_\sigma(x, x^{(i)})$, where

$$\begin{aligned}
\nabla_x w_\sigma(x, x^{(i)}) &= \nabla_x \frac{K_\sigma(x, x^{(i)})}{\sum_j K_\sigma(x, x^{(j)})} \\
&= \frac{K(x, x^{(i)}) \frac{x^{(i)} - x}{\sigma^2} \sum_j K_\sigma(x, x^{(j)}) + K(x, x^{(i)}) \sum_j K(x, x^{(j)}) \frac{x - x^{(j)}}{\sigma^2}}{(\sum_j K_\sigma(x, x^{(j)}))^2} \\
&= \frac{K(x, x^{(i)}) \sum_j K_\sigma(x, x^{(j)}) \frac{x^{(i)} - x^{(j)}}{\sigma^2}}{(\sum_j K_\sigma(x, x^{(j)}))^2}
\end{aligned}$$

Let $|\cdot|$ denote taking element-wise absolute value, and $x \leq^* y$ denotes for all elements of the vector $|x_i| \leq |y_i|$. By Cauchy inequality and $\|\cdot\|_2 \leq \|\cdot\|_1$ we have

$$\nabla_x w_\sigma(x, x^{(i)}) \leq^* \frac{K(x, x^{(i)}) \sum_j K(x, x^{(j)}) \sum_j |x^{(i)} - x^{(j)}|}{\sigma^2 (\sum_j K_\sigma(x, x^{(j)}))^2} \leq^* \frac{1}{\sigma^2} \sum_j |x^{(i)} - x^{(j)}|$$

Therefore

$$\sup_x \|\nabla_x w_\sigma(x, x^{(i)})\|_1 = O(1/\sigma^2)$$

This gives us a bound on the matrix 1-norm

$$\sup_x \|\nabla_x W_\sigma(x)^T\|_1 \leq \sup_x \sqrt{mn} \|\nabla_x W_\sigma(x)^T\|_\infty = \sqrt{mn} \sup_x \max_{i=1}^n \|\nabla_x w_\sigma(x, x^{(i)})\|_1 = O(1/\sigma^2)$$

Because both Ω and $\mathcal{X}$ are convex sets, this implies the following Lipschitz property

$$\frac{\|f^*(x_1) - f^*(x_2)\|_1}{\|x_1 - x_2\|_1} \leqslant \sup_x \|\nabla_x f^*(x)\|_1 = O(1/\sigma^2)$$

□

Proposition 1. *Consider the denoising regularizer $R(\theta\,;\sigma)$. Suppose $\mathcal{F}$ is closed under input translation and that, for any $\theta \in \Theta$, there exists $f \in \mathcal{F}$ such that $f(x)$ maps to the prior $p_\theta(z)$ all $x \in \mathcal{X}$. Furthermore, assume that there exists $\theta \in \Theta$ such that $p_\theta(x, z) = p_\theta(z)p_\theta(x)$. Then $R(\theta\,;\sigma_1)$ is stronger $R(\theta\,;\sigma_2)$ when $\sigma_1 \geq \sigma_2$; i.e., $\min_\theta R(\theta\,;\sigma_1) = \min_\theta R(\theta\,;\sigma_2) = 0$ and $R(\theta\,;\sigma_1) \geq R(\theta\,;\sigma_2)$ for all $\theta \in \Theta$.*

Proof. Proof is provided in two parts.

Part 1. Recall that R is always non-negative. Since there exists $\theta \in \Theta$ such that $p_\theta(x, z) = p_\theta(z)p_\theta(x)$, and $f \in \mathcal{F}$ such that $f(x) = p_\theta(z)$, then $\min_\theta R(\theta\,;\sigma) = 0$ for any choice of σ.

Part 2. Let $\varepsilon_1 \sim \mathcal{N}(\mathbf{0}, \sigma_1 \mathbf{I})$, $\varepsilon_2 \sim \mathcal{N}(\mathbf{0}, \sigma_2 \mathbf{I})$, and $\varepsilon_\delta = \mathcal{N}(\mathbf{0}, (\sigma_1 - \sigma_2)\mathbf{I})$. Then

$$\begin{aligned}
R(\theta\,;\sigma_1) &= \min_{f \in \mathcal{F}} \mathbb{E}_{\varepsilon_1} \mathbb{E}_{\hat{p}(x)} \left[D(f(x + \varepsilon_1) \parallel p_\theta(z \mid x)) \right] && (52) \\
&= \min_{f \in \mathcal{F}} \mathbb{E}_{\varepsilon_\delta} \mathbb{E}_{\varepsilon_2} \mathbb{E}_{\hat{p}(x)} \left[D(f(x + \varepsilon_\delta + \varepsilon_2) \parallel p_\theta(z \mid x)) \right] && (53) \\
&\geq \mathbb{E}_{\varepsilon_\delta} \min_{f \in \mathcal{F}} \mathbb{E}_{\varepsilon_2} \mathbb{E}_{\hat{p}(x)} \left[D(f(x + \varepsilon_\delta + \varepsilon_2) \parallel p_\theta(z \mid x)) \right]. && (54)
\end{aligned}$$

Since $\mathcal{F}$ is closed under input translation,

$$\mathbb{E}_{\varepsilon_\delta} \min_{f \in \mathcal{F}} \mathbb{E}_{\varepsilon_2} \mathbb{E}_{\hat{p}(x)} \left[D(f(x + \varepsilon_\delta + \varepsilon_2) \parallel p_\theta(z \mid x)) \right] = R(\theta \,;\, \varepsilon_2). \tag{55}$$

It thus follows that $R(\theta \,;\, \sigma_1) \geq R(\theta \,;\, \sigma_2)$ for all $\theta \in \Theta$. □

Proposition 2. *Let $\mathcal{P}$ be an exponential family with corresponding mean parameter space $\mathcal{M}$ and sufficient statistic function $T(\cdot)$. With a slight abuse of notation, consider $g \in \mathcal{G} : \mathcal{Z} \to \mathcal{M}$. Define $q(x, z) = \hat{p}(x)q(z \mid x)$, where $q(z \mid x)$ is a fixed inference model. Supposing $\mathcal{G}$ has infinite capacity, then the optimal generative model in Eq. (5) returns $g^*(z) = \mu \in \mathcal{M}$, where*

$$\mu = \sum_{i=1}^{n} q(x^{(i)} \mid z) \cdot T(x^{(i)}) = \sum_{i=1}^{n} \left(\frac{q(z \mid x^{(i)})}{\sum_j q(z \mid x^{(j)})} \cdot T(x^{(i)}) \right). \tag{19}$$

Proof. For a given inference model $q(z \mid x)$, the optimal generator maximizes the objective

$$\begin{aligned}
\max_{g \in \mathcal{G}} \mathbb{E}_{\hat{p}(x)} \mathbb{E}_{q(z|x)} \left[\ln g(z)(x) \right] &= \max_{g \in \mathcal{G}} \mathbb{E}_{q(x,z)} \left[\ln g(z)(x) \right]. && (56) \\
&= \max_{g \in \mathcal{G}} \mathbb{E}_{q(x,z)} \left[\ln p_{g(z)}(x) \right] && (57) \\
&\leq \mathbb{E}_{q(z)} \max_{p \in \mathcal{P}} \mathbb{E}_{q(x|z)} \ln p(x) && (58) \\
&= \mathbb{E}_{q(z)} \max_{\mu \in \mathcal{M}} \mathbb{E}_{q(x|z)} \ln p_\mu(x), && (59)
\end{aligned}$$

where p_μ denotes the distribution $p \in \mathcal{P}$ with associate mean parameter μ. This inequality is tight since we can select $g^* \in \mathcal{G}$ such that

$$g^*(z) = \arg\max_{\mu \in \mathcal{M}} \mathbb{E}_{q(x|z)} \ln p_\mu(x). \tag{60}$$

Recall that the maximum likelihood and maximum entropy solutions are equivalent for an exponential family. From the moment-matching condition of maximum entropy, it follows that

$$\begin{aligned}
g^*(z) &= \arg\max_{\mu \in \mathcal{M}} \mathbb{E}_{q(x|z)} \ln p_\mu(x) && (61) \\
&= \mathbb{E}_{q(x|z)} \left[T(x) \right] && (62) \\
&= \sum_{i=1}^{n} q(x^{(i)} \mid z) \cdot T(x^{(i)}) && (63) \\
&= \sum_{i=1}^{n} \left(\frac{q(z \mid x^{(i)})}{\sum_j q(z \mid x^{(j)})} \cdot T(x^{(i)}) \right). && (64)
\end{aligned}$$

□

Proposition 3. *Consider the regularizer $R_k(\theta \,;\, \sigma, \mathcal{F}_H)$. Under similar assumptions as Proposition 1, then R_{k_1} is stronger than R_{k_2} when $k_1 \leq k_2$; i.e., $\min_\theta R_{k_1}(\theta \,;\, \sigma, \mathcal{F}_H) = \min_\theta R_{k_2}(\theta \,;\, \sigma, \mathcal{F}_H) = 0$ and $R_{k_1}(\theta \,;\, \sigma, \mathcal{F}_H) \leq R_{k_2}(\theta \,;\, \sigma, \mathcal{F}_H)$ for all $\theta \in \Theta$.*

Proof. Proof is provided in two parts.

Part 1. The relevant assumptions are that there exists $\theta \in \Theta$ such that $p_\theta(x, z) = p_\theta(z)p_\theta(x)$, and $f \in \mathcal{F}_H$ such that $f(x) = p_\theta(z)$. Note that R_k is always non-negative. It follows readily that $\min_\theta R_k(\theta \,;\, \sigma, \mathcal{F}_H) = 0$ for any choice of k.

Part 2. We define $\mathcal{L}_k$ as

$$\begin{aligned}
\mathcal{L}_k &= \mathbb{E}_{\hat{p}(x)} \ln p_\theta(x) - R_k(\theta \,;\, \sigma, \mathcal{F}_H) && (65) \\
&= \max_{f \in \mathcal{F}_H} \mathbb{E}_{\hat{p}(x)} \mathbb{E}_\varepsilon \mathbb{E}_{z_1 \ldots z_k \sim f(x+\varepsilon)} \left[\ln \frac{1}{k} \sum_{i=1}^{k} \frac{p_\theta(x, z_i)}{f(x + \varepsilon)(z_i)} \right]. && (66)
\end{aligned}$$

It suffices to show that $\mathcal{L}_k \geq \mathcal{L}_m$ when $k \geq m$. We adapt the proof from [8]'s Theorem 1 as follows. Let $|I| = m$ denote a uniformly distributed subset of distinct indices from $\{1, \ldots, k\}$. For any choice of $f \in \mathcal{F}_H$, the following inequality holds

$$\mathbb{E}_{\hat{p}(x)}\mathbb{E}_{\varepsilon}\mathbb{E}_{z_1 \ldots z_k \sim f(x+\varepsilon)} \left[\ln \frac{1}{k} \sum_{i=1}^{k} \frac{p_\theta(x, z_i)}{f(x+\varepsilon)(z_i)} \right] \tag{67}$$

$$= \mathbb{E}_{\hat{p}(x)}\mathbb{E}_{\varepsilon}\mathbb{E}_{z_1 \ldots z_k \sim f(x+\varepsilon)} \left[\ln \mathbb{E}_{I=\{i_1 \ldots i_m\}} \left[\frac{1}{m} \sum_{j=1}^{m} \frac{p_\theta(x, z_{i_j})}{f(x+\varepsilon)(z_{i_j})} \right] \right] \tag{68}$$

$$\geq \mathbb{E}_{\hat{p}(x)}\mathbb{E}_{\varepsilon}\mathbb{E}_{z_1 \ldots z_k \sim f(x+\varepsilon)}\mathbb{E}_{I=\{i_1 \ldots i_m\}} \left[\ln \frac{1}{m} \sum_{j=1}^{m} \frac{p_\theta(x, z_{i_j})}{f(x+\varepsilon)(z_{i_j})} \right] \tag{69}$$

$$= \mathbb{E}_{\hat{p}(x)}\mathbb{E}_{\varepsilon}\mathbb{E}_{z_1 \ldots z_m \sim f(x+\varepsilon)} \left[\ln \frac{1}{m} \sum_{i=1}^{m} \frac{p_\theta(x, z_i)}{f(x+\varepsilon)(z_i))} \right]. \tag{70}$$

It thus follows that $\mathcal{L}_k \geq \mathcal{L}_m$. □