MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

AutoVAE: Mismatched Variational Autoencoder with Irregular Posterior Prior Pairing

Koike-Akino, Toshiaki; Wang, Ye

TR2022-071 July 20, 2022

Abstract

The variational autoencoder (VAE) has been used in a myriad of applications, e.g., dimensionality reduction and generative modeling. VAE uses a specific model for stochastic sampling in latent space. The normal distribution is the most commonly used one because it allows a straightforward sampling, a reparameterization trick, and a differentiable expression of the Kullback–Leibler divergence. Although various other distributions such as Laplace were studied in literature, the effect of heterogeneous use of different distributions for posterior-prior pair is less known to date. In this paper, we investigate numerous possibilities of such a mismatched VAE, e.g., where the uniform distribution is used as a posterior belief at the encoder while the Cauchy distribution is used as a prior belief at the decoder. To design the mismatched VAE, the total number of potential combinations to explore grows rapidly with the number of latent nodes when allowing different distributions across latent nodes. We propose a novel framework called AutoVAE, which searches for better pairing set of posterior-prior beliefs in the context of automated machine learning for hyperparameter optimization. We demonstrate that the proposed irregular pairing offers a potential gain in the variational Renyi bound. In addition, we analyze a variety of likelihood beliefs and divergence order.

IEEE International Symposium on Information Theory (ISIT) 2022

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

AutoVAE: Mismatched Variational Autoencoder with Irregular Posterior-Prior Pairing

Toshiaki Koike-Akino, Ye Wang
Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA 02139, USA.
Email: {koike, yewang}@merl.com

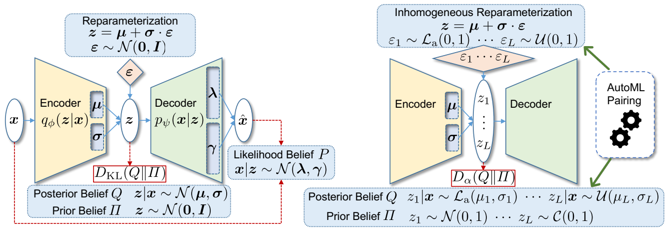
***Abstract*—The variational autoencoder (VAE) has been used in a myriad of applications, e.g., dimensionality reduction and generative modeling. VAE uses a specific model for stochastic sampling in latent space. The normal distribution is the most commonly used one because it allows a straightforward sampling, a reparameterization trick, and a differentiable expression of the Kullback–Leibler divergence. Although various other distributions such as Laplace were studied in literature, the effect of heterogeneous use of different distributions for posterior-prior pair is less known to date. In this paper, we investigate numerous possibilities of such a mismatched VAE, e.g., where the uniform distribution is used as a posterior belief at the encoder while the Cauchy distribution is used as a prior belief at the decoder. To design the mismatched VAE, the total number of potential combinations to explore grows rapidly with the number of latent nodes when allowing different distributions across latent nodes. We propose a novel framework called AutoVAE, which searches for better pairing set of posterior-prior beliefs in the context of automated machine learning for hyperparameter optimization. We demonstrate that the proposed irregular pairing offers a potential gain in the variational Rényi bound. In addition, we analyze a variety of likelihood beliefs and divergence order.**

***Index Terms*—Deep learning, variational Bayes, autoencoder**

I. Introduction

The variational autoencoder (VAE) [1, 2] is configured with a parametric encoder and decoder, as shown in Fig. 1(a), to learn a latent variable model underlying the data within a variational inference (VI) framework. There are many variants of VAE, e.g., as listed in Table I. For example, β-VAE [3] uses an emphasized Kullback–Leibler divergence (KLD) to regularize the latent distribution more strongly. The continuous Bernoulli and beta distributions are studied as alternative likelihood models in [4]. Laplace and Cauchy distributions are considered as an alternative prior belief for sparse latent in [5]. The normal posterior belief is adjusted by inverse autoregressive flow (IAF) [6], importance weighted autoencoder (IWAE) [7, 8], and Gibbs sampling [9]. The IWAE is further extended to the variational Rényi (VR) [10] based on the Rényi's α-divergence. The generalized VI (GVI) [11] then discusses any arbitrary loss, divergence, and posterior selections.

While the generalized VAE offers great degrees of freedom, it in turn makes it difficult to design those selections in addition to other architecture hyperparameters. In this paper, we propose a concept called 'AutoVAE' depicted in Fig. 1(b), which facilitates finding a proper choice of posterior, prior, likelihood, and divergence with an automated machine learning (AutoML) framework [12]. Although the normal distribution often works well for latent sampling, the effect of mismatched posterior-prior pairing is less known in literature. We show some interesting behaviors when exploring diverse settings — the VR bound can be improved by heterogeneous pairing, e.g., when the logistic distribution is used as a posterior belief at the VAE encoder despite that the VAE decoder assumes the normal distribution as a prior belief.


Reparameterization
$\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \cdot \boldsymbol{\varepsilon}$
$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
$\boldsymbol{\varepsilon}$
$\boldsymbol{x}$
Encoder
$q_\phi(\boldsymbol{z}|\boldsymbol{x})$
$\boldsymbol{\mu}$
$\boldsymbol{\sigma}$
$\boldsymbol{z}$
Decoder
$p_\psi(\boldsymbol{x}|\boldsymbol{z})$
$\boldsymbol{\lambda}$
$\boldsymbol{\gamma}$
$\hat{\boldsymbol{x}}$
$D_{\mathrm{KL}}(Q\|\mathit{\Pi})$
Likelihood Belief P
$\boldsymbol{x}|\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\lambda}, \boldsymbol{\gamma})$
Posterior Belief Q $\boldsymbol{z}|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$
Prior Belief $\mathit{\Pi}$ $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
Inhomogeneous Reparameterization
$\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \cdot \boldsymbol{\varepsilon}$
$\varepsilon_1 \sim \mathcal{L}_{\mathrm{a}}(0,1) \cdots \varepsilon_L \sim \mathcal{U}(0,1)$
$\varepsilon_1 \cdots \varepsilon_L$
Encoder
$\boldsymbol{\mu}$
$\boldsymbol{\sigma}$
z_1
z_L
Decoder
AutoML
Pairing
$D_\alpha(Q\|\mathit{\Pi})$
Posterior Belief Q $z_1|\boldsymbol{x} \sim \mathcal{L}_{\mathrm{a}}(\mu_1, \sigma_1) \cdots z_L|\boldsymbol{x} \sim \mathcal{U}(\mu_L, \sigma_L)$
Prior Belief $\mathit{\Pi}$ $z_1 \sim \mathcal{N}(0,1) \cdots z_L \sim \mathcal{C}(0,1)$

(a) Standard VAE with Normal Posterior-Prior-Likelihood (b) Mismatched VAE with Automated Irregular Posterior-Prior Pairing

Fig. 1. VAE pipeline: (a) Standard VAE; (b) Mismatched VAE.

TABLE I
Typical setting for variational inference methods

Method	Likelihood	Discrepancy	Posterior	Prior
Standard VAE [1, 2]	$\mathcal{B}$, $\mathcal{N}$	KLD	$\mathcal{N}$	$\mathcal{N}$
β-VAE [3]	$\mathcal{B}$, $\mathcal{N}$	$\beta\times$KLD	$\mathcal{N}$	$\mathcal{N}$
$\mathcal{CB}$-VAE [4]	$\mathcal{CB}$	KLD	$\mathcal{N}$	$\mathcal{N}$
Sparse-VAE [5]	$\mathcal{B}$, $\mathcal{N}$	KLD	$\mathcal{L}_{\mathrm{a}}$, $\mathcal{C}$	$\mathcal{L}_{\mathrm{a}}$, $\mathcal{C}$
IAF-VI [6]	$\mathcal{B}$, $\mathcal{N}$	KLD	IAF-$\mathcal{N}$	$\mathcal{N}$
IWAE [7]	$\mathcal{B}$, $\mathcal{N}$	KLD	IW-$\mathcal{N}$ [8]	$\mathcal{N}$
Rényi-VAE [10]	$\mathcal{B}$, $\mathcal{N}$	D_α	IW-$\mathcal{N}$	$\mathcal{N}$
Gibbs VI [9]	Any	KLD	Gibbs	$\mathcal{N}$
Generalized VI [11]	Any	Any	Any	Any
Mismatched VAE	P	D_α	$Q \neq \mathit{\Pi}$	$\mathit{\Pi}$
AutoVAE	$\mathcal{P}$	$\mathcal{D}$	$\mathcal{Q}$	$\boldsymbol{\Pi}$

II. Mismatched VAE

Table II lists our notations of various random distributions under consideration and its probability density function (PDF).

A. Variational Inference (VI)

Let $\boldsymbol{x} \in \mathbb{R}^N$ be an N-dimensional data input to the VAE encoder. The encoder generates an L-dimensional latent variable $\boldsymbol{z} \in \mathbb{R}^L$. The latent variable is then input to the VAE decoder to generate a reconstructed data $\hat{\boldsymbol{x}} \in \mathbb{R}^N$. The encoder and decoder are configured with parameterized deep neural networks (DNNs) mapping as $q_\phi : \boldsymbol{x} \to \boldsymbol{z}$ and $p_\psi : \boldsymbol{z} \to \hat{\boldsymbol{x}}$, respectively, with ϕ and ψ being DNN parameters (weights,

TABLE II
PROBABILITY DISTRIBUTION NOTATION AND PDF

Distribution	Notation	PDF $f(x)$
Normal	$\mathcal{N}(\mu,\sigma)$	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\big(-\frac{(x-\mu)^2}{2\sigma^2}\big)$.
Laplace	$\mathcal{L}_{\mathrm{a}}(\mu,\sigma)$	$\frac{1}{2\sigma}\exp\big(-\frac{\lvert x-\mu\rvert}{\sigma}\big)$
Cauchy	$\mathcal{C}(\mu,\sigma)$	$\frac{1}{\pi}\frac{\sigma}{\sigma^2+(x-\mu)^2}$
Logistic	$\mathcal{L}_{\mathrm{o}}(\mu,\sigma)$	$\frac{1}{\sigma}\Big(\exp\big(\frac{x-\mu}{2\sigma}\big)+\exp\big(\frac{\mu-x}{2\sigma}\big)\Big)^{-2}$
Uniform	$\mathcal{U}(\mu,\sigma)$	$\frac{1}{2\sigma},\quad \mu-\sigma\le x\le\mu+\sigma$
Gumbel	$\mathcal{G}(\mu,\sigma)$	$\frac{1}{\sigma}\exp\Big(-\frac{x-\mu}{\sigma}-\exp\big(-\frac{x-\mu}{\sigma}\big)\Big)$
Exponential	$\mathcal{E}(\sigma)$	$\frac{1}{\sigma}\exp\big(-\frac{x}{\sigma}\big),\quad x\ge 0$
Bernoulli	$\mathcal{B}(\lambda)$	$\lambda^x(1-\lambda)^{1-x},\qquad x\in\{0,1\}$
Cont. Bernoulli [4]	$\mathcal{CB}(\lambda)$	$C(\lambda)\lambda^x(1-\lambda)^{1-x},\quad 0\le x\le 1$
Beta	$\mathcal{B}_{\mathrm{e}}(\lambda,\gamma)$	$\frac{\mathit{\Gamma}(\lambda+\gamma)}{\mathit{\Gamma}(\lambda)\mathit{\Gamma}(\gamma)}x^{\lambda-1}(1-x)^{\gamma-1}$

biases, etc.). The DNN tries to minimize a reconstruction loss, which would be typically negative log-likelihood (NLL).

For a given choice of parameters ϕ and ψ, the VAE encoder and decoder models imply a conditional distribution (a.k.a., posterior) $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and a conditional distribution (a.k.a., likelihood) $p_\psi(\boldsymbol{x}|\boldsymbol{z})$, respectively. Letting $\pi(\boldsymbol{z})$ be a prior distribution for the latent variable $\boldsymbol{z}$, we wish to maximize the marginal distribution $\Pr(\boldsymbol{x})$, given by

$$\Pr(\boldsymbol{x}) = \int p_\psi(\boldsymbol{x}|\boldsymbol{z})\pi(\boldsymbol{z})\mathrm{d}\boldsymbol{z}, \tag{1}$$

which is generally intractable to compute exactly. While it could be possible to approximate the integration with sampling of $\boldsymbol{z}$, the crux of the VAE approach is to utilize a variational lower-bound of the posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ implied by the generative model $p_\psi(\boldsymbol{x}|\boldsymbol{z})$. With $q_\psi(\boldsymbol{z}|\boldsymbol{x})$ representing the variational approximation of the posterior, the *evidence lower-bound* (ELBO) is given by

$$\begin{aligned}\log\Pr(\boldsymbol{x}) &= \log \mathop{\mathbb{E}}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{p_\psi(\boldsymbol{x}|\boldsymbol{z})\pi(\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] &(2)\\ &\ge \mathop{\mathbb{E}}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log\frac{p_\psi(\boldsymbol{x}|\boldsymbol{z})\pi(\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] &(3)\\ &= \mathop{\mathbb{E}}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\big[\log p_\psi(\boldsymbol{x}|\boldsymbol{z})\big] - D_{\mathrm{KL}}\big(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|\pi(\boldsymbol{z})\big), &(4)\end{aligned}$$

where $D_{\mathrm{KL}}(Q\|\mathit{\Pi})$ denotes the KLD, measuring discrepancy between the posterior and prior distributions, defined as

$$D_{\mathrm{KL}}(Q\|\mathit{\Pi}) = \mathop{\mathbb{E}}_{z\sim Q}\left[\log\left(\frac{Q(z)}{\mathit{\Pi}(z)}\right)\right]. \tag{5}$$

The VAE encoder and decoder are jointly trained such that the ELBO is maximized under the variational Bayes framework.

B. Generalized VAE

In the ELBO (4), there are four important factors to specify: i) likelihood belief $P = p_\psi(\boldsymbol{x}|\boldsymbol{z})$; ii) posterior belief $Q = q_\phi(\boldsymbol{z}|\boldsymbol{x})$; iii) prior belief $\mathit{\Pi} = \pi(\boldsymbol{z})$; and iv) discrepancy measure $D = D_{\mathrm{KL}}(.\|.)$. As shown in Fig. 1(a), a standard VAE often uses the normal distribution (or Bernoulli distribution for nearly binary image reconstruction) for likelihood belief $P = \mathcal{N}(\lambda,\gamma)$, and a specific KLD $D_{\mathrm{KL}}(Q\|\mathit{\Pi})$ to regularize latent variables considering normal posterior $Q = \mathcal{N}(\mu,\sigma)$ and normal prior $\mathit{\Pi} = \mathcal{N}(0,1)$. The GVI [11] discusses any arbitrary loss (due to generalized likelihood), divergence (non-KLD), and posterior selections. It was proven that the standard VAE is asymptotically optimal for infinite dimension, while it is no longer justified when likelihood/posterior/prior beliefs are misspecified as opposed to the real data distribution.

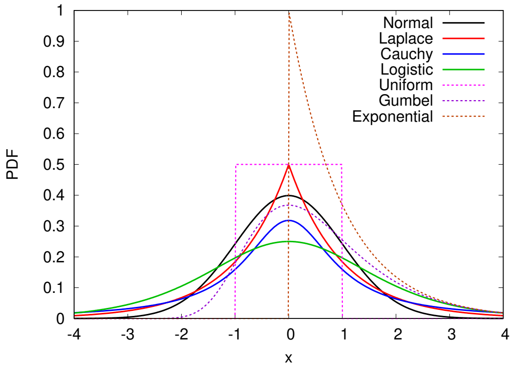


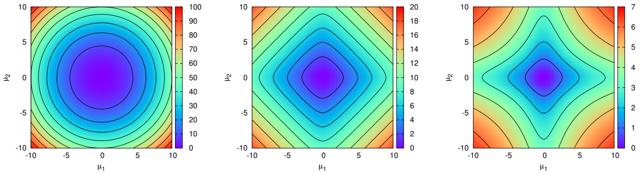
Fig. 2. PDF of standard location-scale family $\mathbb{LSF}(0,1)$.

Although an exponential-family prior belief will result into the same exponential-family posterior belief under some condition [13], the GVI motivates us to consider mismatched pairing for posterior-prior beliefs — e.g., logistic posterior and Laplace prior. Moreover, the generalized VAE can allow an irregular inhomogeneous pairing — e.g., 30% latent nodes use Laplace-normal pairs and 70% nodes uniform-Cauchy pairs as in Fig. 1(b). Although non-Gaussian prior has been considered in literature [5], irregular mismatched pairing has not been investigated thoroughly to the best of our knowledge. To design such irregular VAEs, this paper provides a convenient set of pairings that have closed-form differentiable expressions of KLD and straighforward reparameterization, without requiring high complexity like IAF [6] and Gibbs sampling [9].

C. Reparameterization Trick: Location-Scale Family

As mentioned, the normal distribution is the most widely used model for stochastic DNNs, allowing simple sampling, a reparameterization trick, and a closed-form expression of KLD. In this paper, we consider diverse alternatives using a location-scale family $\mathbb{LSF}(\mu,\sigma)$, which holds the same distribution within a transform with a location of $\mu\in\mathbb{R}$ and a scale of $\sigma\in\mathbb{R}_+$. For instance, given a random variable ε drawn as $\varepsilon\sim\mathbb{LSF}(0,1)$, its translated random variable $Z = \mu+\sigma\cdot\varepsilon$ follows the same family as $Z\sim\mathbb{LSF}(\mu,\sigma)$. This transform is known as *reparameterization trick* — one of key enabling methods of stochastic DNN training to back-propagate a gradient for μ and σ while keeping a desired distribution in forward sampling, as depicted in Fig. 1.

We consider seven $\mathbb{LSF}$ distributions (denoted in Table II): normal $\mathcal{N}$; Laplace $\mathcal{L}_{\mathrm{a}}$; Cauchy $\mathcal{C}$; logistic $\mathcal{L}_{\mathrm{o}}$; uniform $\mathcal{U}$; Gumbel $\mathcal{G}$; and exponential $\mathcal{E}$ (which is not $\mathbb{LSF}$ but a scale family). The PDF of its standard form $\mathbb{LSF}(0,1)$ is plotted in Fig. 2. These distributions are useful as a choice of potential



(a) Normal-Normal (b) Laplace-Laplace (c) Cauchy-Cauchy

Fig. 3. KLD landscape $D_{\mathrm{KL}}(Q\|\mathit{\Pi})$ for matched posterior-prior pairs.

TABLE III
KLD FOR VARIOUS POSTERIOR-PRIOR PAIRS

Posterior Q	Prior $\mathit{\Pi}$	KLD $D_{\mathrm{KL}}(Q\|\mathit{\Pi})$
$\mathcal{N}(\mu,\sigma)$	$\mathcal{N}(0,1)$	$\frac{1}{2}\big(\mu^2+\sigma^2-1-\log(\sigma^2)\big)$
$\mathcal{L}_{\mathrm{a}}(\mu,\sigma)$	$\mathcal{N}(0,1)$	$\frac{1}{2}\mu^2+\sigma^2-1-\frac{1}{2}\log\big(\frac{2\sigma^2}{\pi}\big)$
$\mathcal{L}_{\mathrm{o}}(\mu,\sigma)$	$\mathcal{N}(0,1)$	$\frac{1}{2}\mu^2+\frac{\pi^2}{6}\sigma^2-2-\frac{1}{2}\log\big(\frac{\sigma^2}{2\pi}\big)$
$\mathcal{U}(\mu,\sigma)$	$\mathcal{N}(0,1)$	$\frac{1}{2}\mu^2+\frac{1}{6}\sigma^2-\frac{1}{2}\log\big(\frac{2\sigma^2}{\pi}\big)$
$\mathcal{G}(\mu,\sigma)$	$\mathcal{N}(0,1)$	$\log(\frac{\sqrt{2\pi}}{\sigma})+\frac{\pi^2\sigma^2}{12}+\frac{(\mu+\sigma\gamma_0)^2}{2}-\gamma_0-1$
$\mathcal{E}(\sigma)$	$\mathcal{N}(0,1)$	$\sigma^2-1-\frac{1}{2}\log\big(\frac{\sigma^2}{2\pi}\big)$
$\mathcal{N}(\mu,\sigma)$	$\mathcal{L}_{\mathrm{a}}(0,1)$	$\mu\cdot\mathrm{erf}\frac{\mu}{\sqrt{2\sigma^2}}+\sqrt{\frac{2\sigma^2}{\pi}}\exp\big(-\frac{\mu^2}{2\sigma^2}\big)$ $-\frac{1}{2}-\frac{1}{2}\log\big(\frac{\pi\sigma^2}{2}\big)$
$\mathcal{L}_{\mathrm{a}}(\mu,\sigma)$	$\mathcal{L}_{\mathrm{a}}(0,1)$	$\lvert\mu\rvert+\sigma\exp\big(-\frac{\lvert\mu\rvert}{\sigma}\big)-1-\log(\sigma)$
$\mathcal{L}_{\mathrm{o}}(\mu,\sigma)$	$\mathcal{L}_{\mathrm{a}}(0,1)$	$2\sigma\log\big(2\cosh\big(\frac{\mu}{2\sigma}\big)\big)-2-\log\big(\frac{\sigma}{2}\big)$
$\mathcal{E}(\sigma)$	$\mathcal{L}_{\mathrm{a}}(0,1)$	$\sigma-\log(\sigma)-1+\log(2)$
$\mathcal{C}(\mu,\sigma)$	$\mathcal{C}(0,1)$	$\log(\mu^2+(1+\sigma)^2)-\log(4\sigma)$
$\mathcal{U}(\mu,\sigma)$	$\mathcal{C}(0,1)$	$\frac{1}{\sigma}\tan^{-1}(\sigma-\mu)+\frac{1}{\sigma}\tan^{-1}(\sigma+\mu)-2$ $-\log\big(\frac{2\sigma}{\pi}\big)+\frac{\sigma-\mu}{2\sigma}\log\big(1+(\sigma-\mu)^2\big)$ $+\frac{\sigma+\mu}{2\sigma}\log\big(1+(\sigma+\mu)^2\big)$
$\mathcal{N}(\mu,\sigma)$	$\mathcal{G}(0,1)$	$-\log(\sigma)+\mu+\exp(-\mu+\frac{\sigma^2}{2})-\frac{1+\log(2\pi)}{2}$
$\mathcal{U}(\mu,\sigma)$	$\mathcal{G}(0,1)$	$\mu+\frac{1}{\sigma}\exp(-\mu)\sinh(\sigma)-\log(2\sigma)$
$\mathcal{G}(\mu,\sigma)$	$\mathcal{G}(0,1)$	$\mu-\log(\sigma)+\mathit{\Gamma}(\sigma+1)\mathrm{e}^{-\mu}-1+\gamma_0(\sigma-1)$
$\mathcal{E}(\sigma)$	$\mathcal{G}(0,1)$	$\sigma+(1+\sigma)^{-1}-1-\log(\sigma)$

posterior beliefs Q for the VAE encoder because of the simple reparameterization trick. For example, sampling the logistic distribution can be done as $Z=\mu+\sigma\varepsilon\sim\mathcal{L}_{\mathrm{o}}(\mu,\sigma)$ for $\varepsilon=\log(V/W)\sim\mathcal{L}_{\mathrm{o}}(0,1)$ with $V,W\sim\mathcal{E}(1)$. The locations vector $\boldsymbol{\mu}\in\mathbb{R}^L$ and scales vector $\boldsymbol{\sigma}\in\mathbb{R}^L$ are produced from data $\boldsymbol{x}$ by the parametric VAE encoder, while the latent $\boldsymbol{z}$ is stochastically sampled by the transformed standard random variables $\boldsymbol{\varepsilon}\sim\mathbb{LSF}(0,1)^L$.

D. Discrepancy Measure: α-Divergence

For the choice of prior beliefs $\mathit{\Pi}=\pi(\boldsymbol{z})$ used in generative models (i.e., VAE decoder), we typically use the standard normal distribution $\mathcal{N}(0,1)$, or a matched standard prior $\mathit{\Pi}=\mathbb{LSF}(0,1)$ for a particular choice of posterior belief $Q=\mathbb{LSF}(\mu,\sigma)$. Fig. 3 shows the KLD landscape for matched normal, Laplace and Cauchy beliefs, where we can see Laplace and Cauchy have non-isotropic distribution which promotes sparse regularization [5]. While we can select any arbitrary distribution for the prior belief regardless of the posterior belief, it is desirable to have a closed-form simple expression of KLD to measure the discrepancy between posterior Q and prior $\mathit{\Pi}$. In this paper, we consider sixteen such pairs of posterior and prior beliefs, whose KLD is listed in Table III.

Fig. 4 shows the KLD $D_{\mathrm{KL}}(Q\|\mathit{\Pi})$ for various pairs of posterior $Q=\mathbb{LSF}(\mu,\sigma)$ and prior $\mathit{\Pi}=\mathbb{LSF}(0,1)$ as a function of location μ for a certain scale σ such that the KLD is minimal. While it is generally true that a matched posterior-prior pairing has smaller KLD, mismatched pairs such as logistic-Laplace has slightly smaller KLD than Laplace-Laplace at a moderate location value. More importantly, those mismatched cases have sufficiently small KLD values for a wide range of location μ. Because KLD cannot be exactly zero unless $\mu=0$ and $\sigma=1$ even for matched pairs, minimizing the KLD term is not necessarily important. It implies that there is a chance that irregular pairing may offer better ELBO eventually.

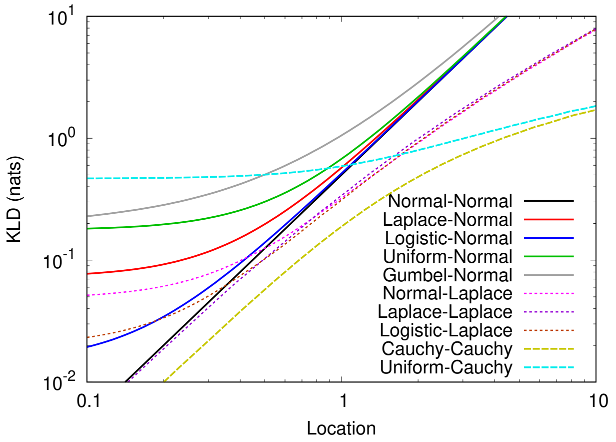


Fig. 4. KLD $D_{\mathrm{KL}}(Q\|\mathit{\Pi})$ as a function of location μ for various posteriors $Q=\mathbb{LSF}(\mu,\sigma)$ against priors $\mathit{\Pi}=\mathbb{LSF}(0,1)$.

Note that the KLD is the most commonly used discrepancy measure to assess the 'difference' between the posterior Q and prior $\mathit{\Pi}$ for VAE. However, besides KLD, the GVI [11] discussed various other discrepancy measures, including the Fisher, Jeffrey, α-, β-, and γ-divergences. In particular, the Rényi's α-divergence is attractive since it covers many variants such as IWAE [7] and the standard VAE as a special case [10] as shown in Table IV. For example, the α-divergence $D_\alpha(Q\|\mathit{\Pi})$ is reduced to the KLD $D_{\mathrm{KL}}(Q\|\mathit{\Pi})$ when $\alpha\to1$. The Rényi divergence of order $\alpha\geq0$ is expressed as

$$D_\alpha(Q\|\mathit{\Pi})=\frac{1}{\alpha-1}\log\mathop{\mathbb{E}}_{z\sim Q}\left[\left(\frac{Q(z)}{\mathit{\Pi}(z)}\right)^{\alpha-1}\right]. \tag{6}$$

More importantly, it was shown that the VR bound [10] based on the Rényi divergence has a tighter bound for ELBO than KLD case ($\alpha\to1$) in (4). Specifically, the VR bound is approximated with K-times latent samples $\boldsymbol{z}_k\sim p_\phi(\boldsymbol{z}|\boldsymbol{x})$ as:

$$\hat{\mathcal{L}}_{\alpha,K}=\frac{1}{1-\alpha}\log\frac{1}{K}\sum_{k=1}^{K}\left(\frac{p_\psi(\boldsymbol{x}|\boldsymbol{z}_k)\pi(\boldsymbol{z}_k)}{q_\phi(\boldsymbol{z}_k|\boldsymbol{x})}\right)^{1-\alpha}. \tag{7}$$

The IWAE [7] is a special case of this VR framework with $\alpha=0$, and it converges to the marginal probability in (2) when $K\to\infty$. As α-divergence is closely related to KLD [14], the aforementioned pairs in Table III have also closed-form expressions [15]. Accordingly, we use the Rényi's α-divergence as a generalized discrepancy measure in this paper.

TABLE IV
RÉNYI DIVERGENCE $D_\alpha(Q\|\mathit{\Pi})$ SPECIAL CASES [10]

Order α	Definition	Correspondence
$\alpha \to 0$	$-\log \int_{Q(z)>0} \mathit{\Pi}(z)\mathrm{d}z$	Overlap (i.e., IWAE [7])
$\alpha = 0.5$	$-2\log(1 - \mathsf{Hel}^2[Q\|\mathit{\Pi}])$	Square Hellinger distance
$\alpha \to 1$	$\int Q(z) \log \frac{Q(z)}{\mathit{\Pi}(z)}\mathrm{d}z$	KLD (i.e., standard VAE [1])
$\alpha = 2$	$-\log(1 - \chi^2[Q\|\mathit{\Pi}])$	χ^2-divergence
$\alpha \to \infty$	$\log \max \frac{Q(z)}{\mathit{\Pi}(z)}$	Worst-case regret

TABLE V
GENERALIZED NLL FOR VARIOUS LIKELIHOOD BELIEFS P

Likelihood P	Generalized NLL Loss ℓ
$\mathcal{B}(\lambda)$	$\mathsf{BCE}(x;\lambda) = -x\log(\lambda) - (1-x)\log(1-\lambda)$
$\mathcal{CB}(\lambda)$	$\mathsf{NLL}(x;\lambda) = \mathsf{BCE}(x;\lambda) - \log C(\lambda)$
$\mathcal{N}(\lambda, *)$	$\mathsf{MSE}(x;\lambda) = (x-\lambda)^2$ (omitting unspecified variance)
$\mathcal{L}_\mathrm{a}(\lambda, *)$	$\mathsf{MAE}(x;\lambda) = \lvert x-\lambda\rvert$ (omitting unspecified variance)
$\mathcal{N}(\lambda, \gamma)$	$\mathsf{NLL}(x;\lambda,\gamma) = \frac{1}{2\gamma^2}\mathsf{MSE}(x;\lambda) + \frac{1}{2}\log(2\pi\gamma^2)$
$\mathcal{L}_\mathrm{a}(\lambda, \gamma)$	$\mathsf{NLL}(x;\lambda,\gamma) = \frac{1}{\gamma}\mathsf{MAE}(x;\lambda) + \log(2\gamma)$
$\mathcal{B}_\mathrm{e}(\lambda, \gamma)$	$\mathsf{NLL}(x;\lambda,\gamma) = (1-\lambda)\log(x) + (1-\gamma)\log(1-x)$ $+ \log\mathit{\Gamma}(\lambda) + \log\mathit{\Gamma}(\gamma) - \log\mathit{\Gamma}(\lambda+\gamma)$

E. Reconstruction Measure: Likelihood Belief

For VAE, any differentiable loss measure can be used as a reconstruction loss in practice; e.g., mean-square error (MSE), mean-absolute error (MAE), and binary cross-entropy (BCE) besides NLL. Nevertheless, most loss functions are closely related to generalized NLL under a specific likelihood belief $P = p_\psi(\boldsymbol{x}|\boldsymbol{z})$ to represent the generative model of the data.

We consider various likelihood beliefs P for VAE as listed in Table V, where we present the generalized NLL loss. For example, BCE is often used for nearly binary images such as MNIST. The BCE is equivalent to the NLL under the likelihood belief based on Bernoulli distribution $\mathcal{B}(\boldsymbol{\lambda})$. In [4], a beta distribution $\mathcal{B}_\mathrm{e}$ was compared as a proper likelihood and a modified belief called *continuous Bernoulli* distribution $\mathcal{CB}$ was proposed to improve the VAE for not-strictly binary images. MSE corresponds to NLL under the normal distribution likelihood $\mathcal{N}(\hat{\boldsymbol{x}}, *)$ when omitting an unspecified variance. MAE corresponds to NLL under Laplace distribution likelihood $\mathcal{L}_\mathrm{a}(\hat{\boldsymbol{x}}, *)$ when omitting an unspecified scale.

The VAE decoder may require multiple variational outputs to generate $\hat{\boldsymbol{x}}$ given likelihood belief P. For example, the normal distribution likelihood $P = \mathcal{N}(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ provides the mean as the reconstructed data $\hat{\boldsymbol{x}} = \boldsymbol{\lambda}$ and its standard deviation of $\boldsymbol{\gamma}$ as a confidence. Likewise, we can use $\hat{\boldsymbol{x}} = \boldsymbol{\lambda}$ for the Laplace distribution in the sense of maximum likelihood. However, in general, given the decoder variational outputs ($\boldsymbol{\lambda}$, $\boldsymbol{\gamma}$, etc.), the data reconstruction should be done by its mode (peak of PDF). Although for Bernoulli likelihood $\mathcal{B}(\lambda)$, the mode is binary as $\hat{x} = 0$ or 1 depending on λ, we use a mean as the reconstruction $\hat{x} = \lambda$, following most papers [4].

F. AutoVAE: Automated Posterior-Prior Pairing

As we discussed above, the VAE needs to specify four factors: posterior Q; prior $\mathit{\Pi}$; likelihood P; and divergence D. To design those factors, we often need manual efforts in searching for the best combinations of different likelihood beliefs as in Table V, different posterior-prior pairs in Table III, and different divergence order $\alpha > 0$ in (7). The searching space will be rapidly exploded when we consider irregular inhomogeneous pairing at each latent node. We use an AutoML framework [12] to explore those factors for VAE design in an automated fashion — we refer to this concept as AutoVAE.

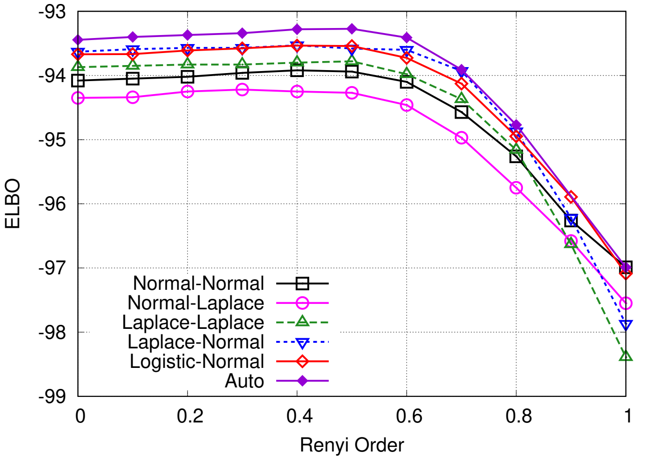


Fig. 5. ELBO $\hat{\mathcal{L}}_{\alpha,50}$ vs. Rényi order α for different posterior-prior pairs with likelihood belief of $P = \mathcal{B}$.

III. EXPERIMENTS

A. Dataset and Architecture

We use benchmark datasets of MNIST, KMNIST, FMNIST, Omniglot, CIFAR10, CIFAR100, SVHN, and STL10 for VAE experiments. We use a simple VAE architecture based on a multi-layer perceptron (MLP) for encoder and decoder, where MLP is composed of two fully-connected linear layers with 400 hidden nodes having rectified linear unit activation. The number of latent nodes is chosen to be $L = 20$, which was found to work well for most cases. We use the adaptive momentum gradient optimization with a learning rate of 1×10^{-3} over 100 epochs at a mini-batch size of 1000. Although AutoML can also design those hyperparameters, we leave them simple as the optimization of DNN architecture and optimization strategy is outside of our main scope.

B. Inception Scores

Once the VAE is trained, the decoder can be used as a generative model to reproduce fake data by feeding random samples drawn from the prior belief $\mathit{\Pi}$. The generated images are evaluated by Fréchet inception distance (FID) [16] and kernel inception distance (KID) [17] to assess its natural distribution from the original image dataset. To evaluate the inception scores, we generate 50,000 images by sampling random latent variables $\boldsymbol{z}$ from prior beliefs $\mathit{\Pi}$.

C. ELBO Performance

Table VI shows results of mismatched VAEs under various combinations of likelihood, posterior, and prior beliefs. For

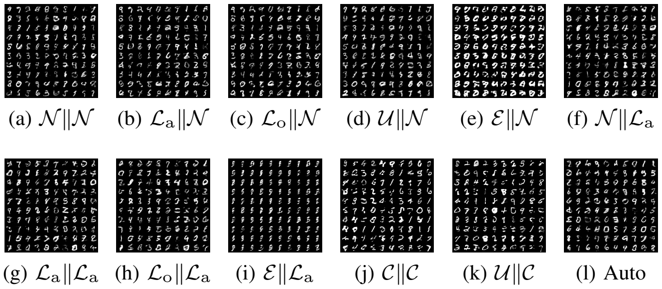
TABLE VI

ELBO, NLL, INCEPTION SCORES FOR VARIOUS POSTERIOR-PRIOR PAIRS ($Q\|\mathit{\Pi}$) WITH DIFFERENT LIKELIHOOD BELIEF P

	$\mathcal{N}\|\mathcal{N}$	$\mathcal{L}_\mathrm{a}\|\mathcal{N}$	$\mathcal{L}_\mathrm{o}\|\mathcal{N}$	$\mathcal{U}\|\mathcal{N}$	$\mathcal{G}\|\mathcal{N}$	$\mathcal{E}\|\mathcal{N}$	$\mathcal{N}\|\mathcal{L}_\mathrm{a}$	$\mathcal{L}_\mathrm{a}\|\mathcal{L}_\mathrm{a}$	$\mathcal{L}_\mathrm{o}\|\mathcal{L}_\mathrm{a}$	$\mathcal{E}\|\mathcal{L}_\mathrm{a}$	$\mathcal{C}\|\mathcal{C}$	$\mathcal{U}\|\mathcal{C}$	$\mathcal{G}\|\mathcal{G}$	Auto
(a) Likelihood Belief $P = \mathcal{N}(\lambda, *)$: Unspecified Normal Distribution, i.e., MSE Loss														
$\hat{\mathcal{L}}_{1,1}$	-19.74	-20.35	-19.23	-22.47	-20.60	-39.76	-19.74	-19.42	-19.39	-34.06	-26.56	-26.33	-19.49	**-19.01**
MSE	12.71	12.76	**12.59**	12.84	12.91	20.0	12.61	13.00	12.72	16.76	26.44	12.84	13.14	12.74
FID	119.0	119.4	**113.2**	142.6	139.2	126.0	119.4	126.9	120.7	223.5	348.4	147.2	134.7	126.1
KID	0.125	0.127	**0.118**	0.138	0.154	0.164	0.145	0.133	0.126	0.246	0.528	0.142	0.149	0.135
(b) Likelihood Belief $P = \mathcal{B}(\lambda)$: Bernoulli Distribution, i.e., BCE Loss														
$\hat{\mathcal{L}}_{1,1}$	-102.5	-103.9	-102.9	-105.9	-103.5	-163.6	-102.7	-103.9	-103.2	-148.1	-203.4	-108.6	-103.4	**-101.6**
BCE	77.15	77.01	76.62	76.66	76.20	124.4	76.81	78.26	77.35	121.6	202.5	76.67	76.90	**76.30**
FID	42.91	43.50	44.01	42.76	41.82	113.27	40.80	41.59	42.13	152.6	389.3	42.42	40.88	**40.19**
KID	0.0369	0.0370	0.0378	0.0359	0.0349	0.1236	0.0347	0.0348	0.0360	0.1908	0.6302	0.0338	0.0344	**0.0337**
(c) Likelihood Belief $P = \mathcal{L}_\mathrm{a}(\lambda, *)$: Unspecified Laplace Distribution, i.e., MAE Loss														
$\hat{\mathcal{L}}_{1,1}$	-65.34	-62.34	-61.83	-64.64	-66.26	-98.29	-62.18	-62.54	-62.27	-88.07	-98.73	-76.29	-65.47	**-61.08**
MAE	49.86	46.71	46.86	46.54	50.86	74.10	46.75	48.32	48.17	69.11	98.54	**44.80**	51.39	46.54
FID	46.02	46.91	48.85	46.41	52.19	159.8	50.48	48.00	48.55	174.2	219.9	102.7	54.58	**44.02**
KID	0.0343	0.0357	0.0375	0.0340	0.0428	159.8	0.0388	0.0342	0.0348	0.1767	0.2233	0.0845	0.0456	**0.0313**
(d) Likelihood Belief $P = \mathcal{N}(\lambda, \gamma^2)$: Normal Distribution														
$\hat{\mathcal{L}}_{1,1}$	1888.6	1899.2	1774.5	1819.8	-8000.1	658.6	2079.9	1521.5	2122.6	-30000	177.6	1969.7	2399.1	**2490.4**
NLL	-1954.8	-1976.0	-1839.4	-1886.3	552.1	-705.1	-2114.4	-1584.2	-2195.1	-748.3	-193.4	-2042.6	-2469.9	**-2575.0**
FID	170.2	167.9	161.5	162.3	294.3	267.7	182.0	167.9	177.2	321.1	423.4	283.4	**87.67**	98.19
KID	0.1982	0.1968	0.1840	0.1932	0.3624	0.3431	0.2195	0.2176	0.2057	0.7301	0.6555	0.3733	**0.0816**	0.0976
(e) Likelihood Belief $P = \mathcal{CB}(\lambda)$: Continuous Bernoulli Distribution														
$\hat{\mathcal{L}}_{1,1}$	1838.0	1835.4	1837.2	1833.2	1838.2	1656.3	**1840.8**	1837.4	1838.4	1656.3	1355.2	1834.8	1838.2	**1840.8**
NLL	-1882.1	-1880.9	-1881.4	-1881.2	-1883.3	-1678.3	-1885.4	-1883.5	-1883.1	-1696.0	-1356.9	**-1885.7**	-1882.4	-1883.9
FID	61.25	63.05	62.21	64.17	57.27	116.13	62.28	60.06	59.21	132.9	318.1	66.85	**52.85**	55.86
KID	0.0576	0.0589	0.0586	0.0609	0.0514	0.1223	0.0597	0.0565	0.0546	0.1467	0.7745	0.0611	**0.0471**	0.0501
(f) Likelihood Belief $P = \mathcal{B}_\mathrm{e}(\lambda, \gamma)$: Beta Distribution														
$\hat{\mathcal{L}}_{1,1}$	6970.2	6062.9	4136.2	4916.3	5362.9	—	5450.7	4301.3	6258.4	7214.3	—	5510.9	5866.2	**8044.7**
NLL	-7053.0	-6158.5	-4223.9	-5002.0	-5430.4	—	-5501.7	-4381.6	-6353.5	-7275.0	—	-5588.8	-5946.9	**-8122.7**
FID	136.44	134.61	135.23	141.30	103.69	—	127.74	134.68	133.45	204.28	—	119.23	**98.15**	98.64
KID	0.1540	0.1525	0.1533	0.1591	0.1119	—	0.1423	0.1514	0.1516	0.2134	—	**0.1036**	0.1050	0.1070
(g) Likelihood Belief $P = \mathcal{B}(\lambda)$: Bernoulli Distribution, i.e., BCE Loss; Rényi order $\alpha = 0$, i.e., IWAE														
$\hat{\mathcal{L}}_{0,50}$	-94.08	-93.63	-93.67	-98.54	-94.44	-132.6	-94.35	-93.87	-94.04	-119.6	-98.06	-100.34	-94.28	**-93.44**
BCE	77.04	76.56	76.70	77.73	**76.55**	103.9	77.50	77.01	77.45	97.02	78.12	79.16	77.10	76.81
FID	38.18	37.01	37.24	42.76	38.13	80.30	41.35	38.09	40.29	111.99	**35.71**	36.82	36.99	36.43
KID	0.0310	0.0299	0.0304	0.0359	0.0321	0.0803	0.0354	0.0316	0.0342	0.1180	**0.0256**	0.0268	0.0312	0.0293

MSE loss with unspecified normal likelihood $P = \mathcal{N}(\lambda, *)$ in (a), it was found that the ELBO $\hat{\mathcal{L}}_{1,1}$ of regular normal-normal pair can be improved by using logistic-normal, Laplace-Laplace, and logistic-Laplace pairs. MSE and inception scores are also improved, e.g., by the logistic-normal pair. When exploring different likelihood beliefs in (a) through (f), we can observe benefits of mismatched pairing. Moreover, an irregular inhomogeneous use of posterior-prior pairs at individual latent nodes (denoted as 'Auto') may further improve the performance via automatic exploration of mixed pairs in some cases. In addition, exploring the Rényi order α, the VR bound $\hat{\mathcal{L}}_{\alpha,K}$ in (7) can be further improved as shown in (g) compared to (b). Note that some combinations (when involving Cauchy or exponential distributions) had numerical instability causing overflow/underflow failure during the training as denoted as '—'. This may be due to the unbounded variance of the Cauchy distribution and the unbalanced support with no location adaptability of exponential distributions.

Fig. 5 shows the impact of the divergence order α. One can see that uncommon pairing such as logistic-normal or Laplace-normal can outperform common choice of normal-normal pair, when exploring the best divergence order α. It is also confirmed that irregular pairing explored in AutoVAE achieves the best performance. The best posterior-prior mixture was $\mathcal{L}_\mathrm{o}\|\mathcal{N}$, $\mathcal{L}_\mathrm{a}\|\mathcal{N}$, and $\mathcal{L}_\mathrm{a}\|\mathcal{L}_\mathrm{a}$ pairs, respectively, for 50%, 30% and 20% of $L = 20$ latent nodes at $\alpha = 0.5$. Fig. 6 shows snapshots of randomly generated images.



(a) $\mathcal{N}\|\mathcal{N}$ (b) $\mathcal{L}_\mathrm{a}\|\mathcal{N}$ (c) $\mathcal{L}_\mathrm{o}\|\mathcal{N}$ (d) $\mathcal{U}\|\mathcal{N}$ (e) $\mathcal{E}\|\mathcal{N}$ (f) $\mathcal{N}\|\mathcal{L}_\mathrm{a}$

(g) $\mathcal{L}_\mathrm{a}\|\mathcal{L}_\mathrm{a}$ (h) $\mathcal{L}_\mathrm{o}\|\mathcal{L}_\mathrm{a}$ (i) $\mathcal{E}\|\mathcal{L}_\mathrm{a}$ (j) $\mathcal{C}\|\mathcal{C}$ (k) $\mathcal{U}\|\mathcal{C}$ (l) Auto

Fig. 6. Image snapshot generated by VAE decoder given $\boldsymbol{z} \sim \mathit{\Pi}$ after trained with various posterior-prior pairs $Q\|\mathit{\Pi}$ and likelihood $P = \mathcal{B}$ with $\alpha = 0$.

IV. Conclusion

We investigated mismatched and irregular posterior-prior pairing for generalized VAE design. It was demonstrated that the mismatched VAE can outperform standard VAE. In addition, we explored the impact of different likelihood and divergence order. We also proposed the concept of AutoVAE to facilitate searching for proper combinations of those factors. We note that the mismatched pairing technique is also applicable to other stochastic DNNs besides VAEs.

References

[1] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[2] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International conference on machine learning*. PMLR, 2014, pp. 1278–1286.

[3] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017.

[4] G. Loaiza-Ganem and J. P. Cunningham, "The continuous Bernoulli: fixing a pervasive error in variational autoencoders," *arXiv preprint arXiv:1907.06845*, 2019.

[5] G. Barello, A. S. Charles, and J. W. Pillow, "Sparse-coding variational auto-encoders," *bioRxiv*, p. 399246, 2018.

[6] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improved variational inference with inverse autoregressive flow," *Advances in neural information processing systems*, vol. 29, pp. 4743–4751, 2016.

[7] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," *arXiv preprint arXiv:1509.00519*, 2015.

[8] C. Cremer, Q. Morris, and D. Duvenaud, "Reinterpreting importance-weighted autoencoders," *arXiv preprint arXiv:1704.02916*, 2017.

[9] P. Alquier, J. Ridgway, and N. Chopin, "On the properties of variational approximations of Gibbs posteriors," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 8374–8414, 2016.

[10] Y. Li and R. E. Turner, "Rényi divergence variational inference," *arXiv preprint arXiv:1602.02311*, 2016.

[11] J. Knoblauch, J. Jewson, and T. Damoulas, "Generalized variational inference: Three arguments for deriving new posteriors," *arXiv preprint arXiv:1904.02063*, 2019.

[12] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[13] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[14] T. Van Erven and P. Harremos, "Rényi divergence and Kullback–Leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[15] M. Gil, F. Alajaji, and T. Linder, "Rényi divergence measures for commonly used univariate continuous distributions," *Information Sciences*, vol. 249, pp. 124–131, 2013.

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[17] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," *arXiv preprint arXiv:1801.01401*, 2018.