Constraining Variational Inference with Geometric Jensen-Shannon Divergence

Jacob Deasy*, **Nikola Simidjievski, Pietro Liò**
Department of Computer Science and Technology
University of Cambridge
{jd645,ns779,pl219}@cam.ac.uk

Abstract

We examine the problem of *controlling divergences* for latent space regularisation in variational autoencoders. Specifically, when aiming to reconstruct example $x \in \mathbb{R}^m$ via latent space $z \in \mathbb{R}^n$ ($n \leq m$), while balancing this against the need for generalisable latent representations. We present a regularisation mechanism based on the *skew-geometric Jensen-Shannon divergence* $\left(\text{JS}^{\text{G}_\alpha}\right)$. We find a variation in $\text{JS}^{\text{G}_\alpha}$, motivated by limiting cases, which leads to an intuitive interpolation between forward and reverse KL in the space of both distributions and divergences. We motivate its potential benefits for VAEs through low-dimensional examples, before presenting quantitative and qualitative results. Our experiments demonstrate that skewing our variant of $\text{JS}^{\text{G}_\alpha}$, in the context of $\text{JS}^{\text{G}_\alpha}$-VAEs, leads to better reconstruction and generation when compared to several baseline VAEs. Our approach is entirely unsupervised and utilises only one hyperparameter which can be easily interpreted in latent space.

1 Introduction

The problem of controlling regularisation strength for generative models is often data-dependent and poorly understood [3, 7]. Post-hoc analysis of coefficients dictating regularisation strength is rarely carried out and even more rarely provides an intuitive explanation (e.g. β-VAE, [13]). Although evidence suggests that stronger regularisation in variational settings leads to desirable disentangled representations of latent factors and better generalisation [38], scaling factors remain opaque and unrelated to the task at hand.

To learn useful latent representations for reconstruction and generation of high-dimensional distributions, the variational inference problem can be addressed through the use of Variational Autoencoders (VAEs) [17, 34]. VAE learning requires optimisation of an objective balancing the quality of samples that are encoded and then decoded, with a regularisation term penalising latent space deviations from a fixed prior distribution. VAEs have favourable properties when compared with other families of generative models, such as Generative Adversarial Networks (GANs) [10] and autoregressive models [9, 20]. In particular, GANs are known to necessitate more stringent and problem-dependent training regimes, while autoregressive models are computationally expensive and inefficient to sample.

VAEs often assume latent variables to be parameterised by a multivariate Gaussian $p_\theta(z) = N(\mu, \sigma^2)$ with $z, \mu, \sigma \in \mathbb{R}^n$, which is approximated by $q_\phi(z|x)$ with $x \in \mathbb{R}^m$ and $n \leq m$. In variational Bayesian methods, using the Evidence Lower BOund (ELBO) [4], the model can be naturally constrained to prevent overfitting by minimising the Kullback-Leibler (KL) [19] divergence to an isotropic unit Gaussian ball $\text{KL}\left(p_\theta(z) \parallel \mathcal{N}(0, I)\right)$. One line of work has sought to better understand this divergence term to induce disentanglement, robustness, and generalisation [5, 6]. Meanwhile, the

*Corresponding author.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

broader framework of learning a VAE as a constrained optimisation problem [13], has allowed for increasing use of more exotic statistical divergences and distances for latent space regularisation [8, 12, 22, 37], such as the regularisation term in InfoVAE [38], the Maximum Mean Discrepancy (MMD) [11].

As regularisation terms increase in complexity, it is advantageous to maintain intuition as to how they operate in latent space and to avoid exponential hyperparameter search spaces on real-world problems. In order to properly capitalise on the advantages of each divergence, it is also desirable that the meaning of scaling factors remains clear when combining multiple divergence terms. For instance, as forward KL and reverse KL are known to have distinct beneficial properties—*zero-avoidance* allowing for exploration of new areas in the latent space [3] and *zero-forcing* more easily ignoring noise for sharper selection of strong modes [37] respectively—there are instances where favouring one over the other would be beneficial. Even better would be to balance use of both properties at the same time in a comprehensible manner.

In this regard, we propose the *skew-geometric Jensen-Shannon Variational Autoencoder* ($\text{JS}^{\text{G}_\alpha}$-VAE) as an unsupervised approach to learning strongly regularised latent spaces. More specifically, we make several contributions: we first discuss the skew-geometric Jensen-Shannon divergence (and its dual form) [30] in the context of the well known KL and Jensen-Shannon (JS) divergences and outline its limited use. We proceed to propose an adjustment of the skew parameter, and show how its effect on an intermediate distribution in $\text{JS}^{\text{G}_\alpha}$ furnishes us with a more intuitive divergence and permits interpolation between forward and reverse KL divergence. We then study the skew-geometric Jensen-Shannon in the wider context of latent space regularisation and use it to derive a loss function for $\text{JS}^{\text{G}_\alpha}$-VAE.

To test the utility of the proposed skew-geometric Jensen-Shannon adjustments, we investigate how $\text{JS}^{\text{G}_\alpha}$ operates on low-dimensional examples. We demonstrate that $\text{JS}^{\text{G}_\alpha}$ has beneficial properties for light-tailed posterior distributions and is a more useful (and tractable) intermediate divergence than standard JS. We further exhibit that $\text{JS}^{\text{G}_\alpha}$ for VAEs has a positive impact on test set reconstruction loss. Namely, we show that the dual form, $\text{JS}_*^{\text{G}_\alpha}$ consistently outperforms forward and reverse KL across several standard benchmark datasets and skew values.[2]

2 $\text{JS}^{\text{G}_\alpha}$ VAE derivation

Existing work suggests that there exists no tractable interpolation between forward and reverse KL for multivariate Gaussians. In this section, we will show that one can be found by adapting $\text{JS}^{\text{G}_\alpha}$. We also exhibit how this interpolation, well-motivated in the space of distributions, reduces to a simple quadratic interpolation in the space of divergences.

2.1 The $\text{JS}^{\text{G}_\alpha}$ divergences family

Problems with KL and JS minimisation. For distributions P and Q of a continuous random variable $X = [X_1, \ldots, X_n]^{\text{T}}$, the Kullback-Leibler (KL) divergence [19] is defined as

$$\text{KL}(P \parallel Q) = \int_X p(x) \log \left[\frac{p(x)}{q(x)} \right] dx, \tag{1}$$

where p and q are the probability densities of P and Q respectively, and $x \in \mathbb{R}^n$. In particular, Equation (1) is known as the forward KL divergence from P to Q, whereas reverse KL divergence refers to $\text{KL}(Q \parallel P)$.

Due to Gaussian distributions being the self-conjugate distributions of choice in variational learning, we are interested in using divergences to compare two multivariate normal distributions $\mathcal{N}_1(\mu_1, \Sigma_1)$ and $\mathcal{N}_2(\mu_2, \Sigma_2)$ with the same dimension n. In this case, the KL divergence is

$$\text{KL}\left(\mathcal{N}_1 \parallel \mathcal{N}_2\right) = \frac{1}{2} \left(\text{tr}\left(\Sigma_2^{-1}\Sigma_1\right) + \ln \left[\frac{|\Sigma_2|}{|\Sigma_1|} \right] + (\mu_2 - \mu_1)^{\text{T}} \Sigma_2^{-1} (\mu_2 - \mu_1) - n \right). \tag{2}$$

[2] Code is available at: https://github.com/jacobdeasy/geometric-js

This expression is well-known in variational inference and, for the case of reverse KL from a standard normal distribution $\mathcal{N}_2(0, I)$ to a diagonal multivariate normal distribution, reduces to the expression

$$\mathrm{KL}\left(\mathcal{N}_1\left(\mu_1, \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right)\right) \parallel \mathcal{N}_2(0, I)\right) = \frac{1}{2}\sum_{i=1}^{n}\left(\sigma_i^2 - \ln\left[\sigma_i^2\right] + \mu_i^2 - 1\right), \tag{3}$$

used as a regularisation term in variational models [13, 17, 27] and is known to enforce zero-avoiding parameters on $\mathcal{N}_1$ when minimised [3, 26]. On the other hand, the forward KL divergence reduces to

$$\mathrm{KL}\left(\mathcal{N}_2(0, I) \parallel \mathcal{N}_1\left(\mu_1, \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_n^1\right)\right)\right) = \frac{1}{2}\sum_{i=1}^{n}\left(\sigma_i^{-2} + \ln\left[\sigma_i^2\right] + \frac{\mu_i^2}{\sigma_i^2} - 1\right), \tag{4}$$

and is known for its zero-forcing property [3, 26]. However, there exist well-known drawbacks of the KL divergence, such as no upper bound leading to unstable optimization and poor approximation [12], as well as its asymmetric property $\mathrm{KL}(P \parallel Q) \neq \mathrm{KL}(Q \parallel P)$. Underdispersed approximations relative to the exact posterior also produce difficulties with light-tailed posteriors when the variational distribution has heavier tails [8].

One attempt at remedying these issues is the well-known symmetrisation, the Jensen-Shannon (JS) divergence [23]

$$\mathrm{JS}(p(z) \parallel q(x)) = \frac{1}{2}\mathrm{KL}\left(p \,\middle\|\, \frac{p+q}{2}\right) + \frac{1}{2}\mathrm{KL}\left(q \,\middle\|\, \frac{p+q}{2}\right). \tag{5}$$

Although the JS divergence is bounded (in $[0, 1]$ when using base 2), and offers some intuition through symmetry, it includes the problematic mixture distribution $\frac{p+q}{2}$. This term means that no closed-form expression exists for the JS divergence between two multivariate normal distributions using Equation (5).

Divergence families. To circumvent these problems, prior work has sought more general families of distribution divergence [29]. For example, when $\lambda = \frac{1}{2}$, JS is a special case of the more general family of λ divergences, defined by

$$\lambda(p(x) \parallel q(x)) = \lambda\mathrm{KL}\left(p \parallel (1-\lambda)p + \lambda q\right) + (1-\lambda)\mathrm{KL}\left(q \parallel (1-\lambda)p + \lambda q\right), \tag{6}$$

for $\lambda \in [0, 1]$, which interpolates between forward and reverse KL, and provides control over the degree of *divergence skew* (how closely related the intermediate distribution is to p or q).

Although λ divergences do not prevent the intractable comparison to a mixture distribution, their broader goal is to measure weighted divergence to an intermediate distribution in the space of possible distributions over X. In the case of the JS divergence, this is the (arithmetic) mean divergence to the arithmetic mean distribution. Recently, [30] and [32] have proposed a further generalisation of the JS divergence using abstract means (*quasi-arithmetic means* [28], also known as *Kolmogorov-Nagumo means*). By choosing the *weighted geometric mean* $\mathrm{G}_\alpha(x, y) = x^{1-\alpha}y^\alpha$ for $\alpha \in [0, 1]$, and using the property that the weighted product of exponential family distributions (which includes the multivariate normal) stays in the exponential family [31], a new divergence family has arisen

$$\mathrm{JS}^{\mathrm{G}_\alpha}(p(x) \parallel q(x)) = (1-\alpha)\mathrm{KL}\left(p \parallel G_\alpha(p, q)\right) + \alpha\mathrm{KL}\left(q \parallel G_\alpha(p, q)\right). \tag{7}$$

$\mathrm{JS}^{\mathrm{G}_\alpha}$, the *skew-geometric Jensen-Shannon divergence*, between two multivariate Gaussians $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ then admits the closed form

$$\mathrm{JS}^{\mathrm{G}_\alpha}\left(\mathcal{N}_1 \parallel \mathcal{N}_2\right) = (1-\alpha)\mathrm{KL}\left(\mathcal{N}_1 \parallel \mathcal{N}_\alpha\right) + \alpha\mathrm{KL}\left(\mathcal{N}_2 \parallel \mathcal{N}_\alpha\right) \tag{8}$$

$$= \frac{1}{2}\Bigg(\mathrm{tr}\left(\Sigma_\alpha^{-1}((1-\alpha)\Sigma_1 + \alpha\Sigma_2)\right) + \log\left[\frac{|\Sigma_\alpha|}{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha}\right] + (1-\alpha)(\mu_\alpha - \mu_1)^{\mathrm{T}}\Sigma_\alpha^{-1}(\mu_\alpha - \mu_1) + \alpha(\mu_\alpha - \mu_2)^{\mathrm{T}}\Sigma_\alpha^{-1}(\mu_\alpha - \mu_2) - n\Bigg), \tag{9}$$

with the equivalent dual divergence being

$$\mathrm{JS}_*^{\mathrm{G}_\alpha}\left(\mathcal{N}_1 \parallel \mathcal{N}_2\right) = (1-\alpha)\mathrm{KL}\left(\mathcal{N}_\alpha \parallel \mathcal{N}_1\right) + \alpha\mathrm{KL}\left(\mathcal{N}_\alpha \parallel \mathcal{N}_2\right) \tag{10}$$

$$= \frac{1}{2}\left((1-\alpha)\mu_1^{\mathrm{T}}\Sigma_1^{-1}\mu_1 + \alpha\mu_2^{\mathrm{T}}\Sigma_2^{-1}\mu_2 - \mu_\alpha^{\mathrm{T}}\Sigma_\alpha^{-1}\mu_\alpha + \log\left[\frac{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha}{|\Sigma_\alpha|}\right]\right), \tag{11}$$

where $\mathcal{N}_\alpha$ has parameters

$$\Sigma_\alpha = \left((1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1}\right)^{-1}, \tag{12}$$

(the matrix harmonic barycenter) and

$$\mu_\alpha = \Sigma_\alpha \left((1-\alpha)\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2\right). \tag{13}$$

Throughout this paper we explore how to incorporate these expressions into variational learning.

2.2 $\text{JS}^{\text{G}_\alpha}$ and $\text{JS}_*^{\text{G}_\alpha}$ in variational neural networks

Interpolation between forward and reverse KL. Before applying $\text{JS}^{\text{G}_\alpha}$, we note that although the mean distribution $\mathcal{N}_\alpha$ can be intuitively understood, the limiting skew cases still seem to offer no insight, as

$$\lim_{\alpha\to 0}\left[\text{JS}^{\text{G}_\alpha}\right] = 0 \qquad \lim_{\alpha\to 1}\left[\text{JS}^{\text{G}_\alpha}\right] = 0 \tag{14}$$

$$\lim_{\alpha\to 0}\left[\text{JS}_*^{\text{G}_\alpha}\right] = 0 \qquad \lim_{\alpha\to 1}\left[\text{JS}_*^{\text{G}_\alpha}\right] = 0. \tag{15}$$

Therefore, we instead choose to consider the more useful intermediate mean distribution

$$\mathcal{N}_{\alpha'} = \mathcal{N}\left(\mu_{(1-\alpha)}, \Sigma_{(1-\alpha)}\right). \tag{16}$$

This is equivalent to simply reversing the geometric mean (using $G_\alpha(y,x)$ rather than $G_\alpha(x,y)$) and trivially still permits a valid divergence as a weighted sum of valid divergences.

Proposition 1. *The alternative divergence*

$$\text{JS}^{\text{G}_{\alpha'}}\left(\mathcal{N}_1 \parallel \mathcal{N}_2\right) = (1-\alpha)\text{KL}\left(\mathcal{N}_1 \parallel \mathcal{N}_{\alpha'}\right) + \alpha\text{KL}\left(\mathcal{N}_2 \parallel \mathcal{N}_{\alpha'}\right), \tag{17}$$

and its dual $\text{JS}_*^{\text{G}_{\alpha'}}$*, interpolate between forward and reverse* KL*, satisfying*

$$\lim_{\alpha\to 0}\left[\text{JS}^{\text{G}_{\alpha'}}\right] = \text{KL}\left(\mathcal{N}_1 \parallel \mathcal{N}_2\right) \qquad \lim_{\alpha\to 1}\left[\text{JS}^{\text{G}_{\alpha'}}\right] = \text{KL}\left(\mathcal{N}_2 \parallel \mathcal{N}_1\right) \tag{18}$$

$$\lim_{\alpha\to 0}\left[\text{JS}_*^{\text{G}_{\alpha'}}\right] = \text{KL}\left(\mathcal{N}_2 \parallel \mathcal{N}_1\right) \qquad \lim_{\alpha\to 1}\left[\text{JS}_*^{\text{G}_{\alpha'}}\right] = \text{KL}\left(\mathcal{N}_1 \parallel \mathcal{N}_2\right). \tag{19}$$

The proof of this is given in Appendix A.1. Henceforth in the paper, unless *explicitly* stated, $\text{JS}^{\text{G}_\alpha}$ refers to $\text{JS}^{\text{G}_{\alpha'}}$ (without the prime ($'$)).

Variational autoencoders. We can now introduce a new VAE loss function based on this finding by using the formulation of VAE optimisation as a constrained optimisation problem given in [13]. For generative models, a suitable objective to maximise is the marginal (log-)likelihood of the observed data $x \in \mathbb{R}^m$ as an expectation over the whole distribution of latent factors $z \in \mathbb{R}^n$

$$\max_\theta \left[\mathbb{E}_{p_\theta(z)}\left[p_\theta(x|z)\right]\right]. \tag{20}$$

More generalisable latent representations can be achieved by imposing an isotropic unit Gaussian constraint on the prior $p(z) = \mathcal{N}(0, I)$, arriving at the constrained optimisation problem

$$\max_{\phi,\theta} \mathbb{E}_{p_\mathcal{D}(x)}\left[\log \mathbb{E}_{q_\phi(z|x)}\left[p_\theta(x|z)\right]\right] \qquad \text{subject to } \; D(q_\phi(z|x) \parallel p(z)) < \varepsilon, \tag{21}$$

where ε dictates the strength of the constraint and D is a divergence. We can then re-write Equation (21) as a Lagrangian under the KKT conditions [15, 18], obtaining

$$\mathcal{F}(\theta, \phi, \lambda; x, z) = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \lambda\left(D(q_\phi(z|x) \parallel p(z)) - \varepsilon\right). \tag{22}$$

By setting $D(\alpha) = \text{JS}^{\text{G}_\alpha}$ or $D(\alpha) = \text{JS}_*^{\text{G}_\alpha}$, we immediately note that our family of divergences includes the β-VAE by setting $\alpha = 1$ and varying λ. In simple terms, a broader family of divergences using both α and β, would dictate *where* and with how much strength to skew an intermediate distribution.

Before experimentation, in order to use $\text{JS}^{\text{G}_\alpha}$ and $\text{JS}_*^{\text{G}_\alpha}$ as divergence measures in variational learning, we first simplify Equations (9) and (11).

Proposition 2. *For a diagonal multivariate normal distribution $\mathcal{N}_1(\mu, diag\left(\sigma_1^2, \ldots, \sigma_n^1\right))$ and a standard normal distribution $\mathcal{N}_2(0, I)$, the skew-geometric Jensen-Shannon divergence $\mathrm{JS}^{\mathrm{G}_\alpha}$—an intermediate of forward and reverse* KL *regularisation—and its dual $\mathrm{JS}_*^{\mathrm{G}_\alpha}$ reduce to*

$$\mathrm{JS}^{\mathrm{G}_\alpha}(\mathcal{N}_1 \parallel \mathcal{N}_2) = \frac{1}{2}\sum_{i=1}^{n}\left(\frac{(1-\alpha)\sigma_i^2 + \alpha}{\sigma_{\alpha,i}^2} + \log\left[\frac{\sigma_{\alpha,i}^2}{\sigma_i^{2(1-\alpha)}}\right] + \frac{(1-\alpha)(\mu_{\alpha,i}-\mu_i)^2}{\sigma_{\alpha,i}^2} + \frac{\alpha\mu_{\alpha,i}^2}{\sigma_{\alpha,i}^2} - 1\right), \tag{23}$$

and

$$\mathrm{JS}_*^{G_\alpha} = \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\mu_i^2}{\sigma_i^2} - \frac{\mu_{\alpha,i}^2}{\sigma_\alpha^2} + \log\left[\frac{\sigma_i^{2(1-\alpha)}}{\sigma_{\alpha,i}^2}\right]\right), \tag{24}$$

respectively, where

$$\sigma_{\alpha,i}^2 = \frac{\sigma_i^2}{(1-\alpha) + \alpha\sigma_i^2}, \tag{25}$$

and

$$\mu_{\alpha,i} = \frac{\sigma_{\alpha,i}^2(1-\alpha)\mu_i}{\sigma_i^2}. \tag{26}$$

The proof of this is given in Appendix A.2.

3 Experiments

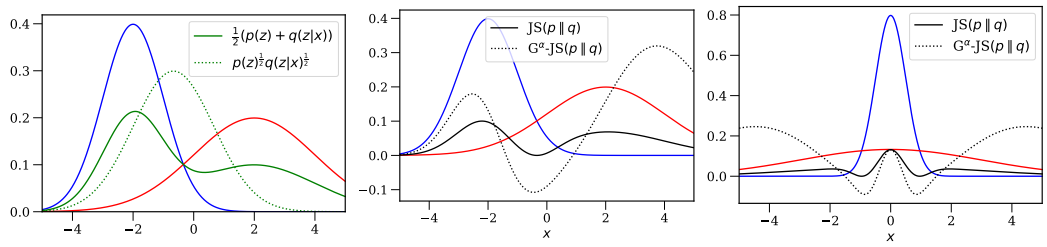
Thus far we have discussed the $\mathrm{JS}^{\mathrm{G}_\alpha}$ divergence and its relationship to KL and in particular VAEs. In this section, we begin by offering a better understanding of where $\mathrm{JS}^{\mathrm{G}_\alpha}$ and its variants differ in distributional space. We then provide a quantitative and qualitative exploration, justifying the immediate benefit of skewing α away from 0 or 1, before finishing with an exploration of the effects this has on VAE reconstruction as well as on the generative capabilities. Note that, in the analyses that follow, we set $\lambda = 1$ for all variants of $\mathrm{JS}^{\mathrm{G}_\alpha}$-VAEs[3].

3.1 Characteristic behaviour of $\mathrm{JS}^{\mathrm{G}_\alpha}$

To elucidate how $\mathrm{JS}^{\mathrm{G}_\alpha}$ will behave in the higher dimensional setting of variational inference, we highlight its properties in the case of one and two dimensions. In Figure 1, univariate Gaussians illustrate how the integrand for $\mathrm{JS}^{\mathrm{G}_\alpha}$ differs favourably from the intractable JS. As the intermediate distribution $\mathcal{N}_\alpha$ in Figure 1a is a Gaussian, $\mathrm{JS}^{\mathrm{G}_\alpha}$ not only permits a closed-form integral, but also offers a more natural interpolation between $p(z)$ and $q(z|x)$, which raises questions about whether intuitive regularisation strength (relative to a known intermediate Gaussian) may be possible in variational settings. Moreover, Figure 1c demonstrates symmetry for $\alpha = 0.5$, and both Figure 1b and Figure 1c depict the increased integrand in areas of low probability density—addressing the issues touched upon earlier, where KL struggles with light-tailed posteriors.

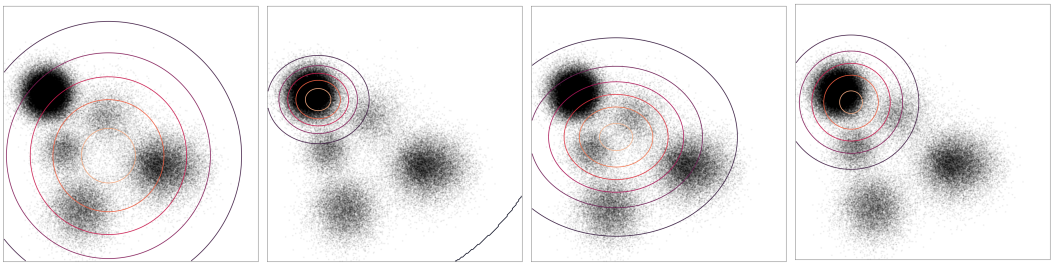
In Figure 2, we use two dimensions to depict the effect of changing divergence measures on optimisation. As the integral of JS divergence is not tractable (and to make comparison fair), we directly optimise a bivariate Gaussian via samples from the data for all divergences. We see that the example mixture of Gaussians leads to the zero-avoiding property of KL divergence in Figure 2a and zero-forcing (i.e. mode dropping) for reverse KL in Figure 2b. While JS divergence provides an intermediate solution in Figure 2c, there is still considerable unnecessary spreading and direct optimisation of the integral will not scale. Finally, $\mathrm{JS}^{\mathrm{G}_\alpha}$ with α naively set to the symmetric case $\alpha = 0.5$ leads to a more reasonable intermediate distribution which both tends towards the dominant mode and offers localised exploration.

[3]Details on the influence of λ on the reconstructive performance of VAEs, with respect to $\mathrm{JS}^{\mathrm{G}_\alpha}$ and $\mathrm{JS}_*^{\mathrm{G}_\alpha}$, are given in Appendix E



(a) Mean comparison. (b) $\mathcal{N}(-2, 1) \parallel \mathcal{N}(2, 2)$ (c) $\mathcal{N}(0, 0.5) \parallel \mathcal{N}(0, 3)$

Figure 1: Comparison of mean distributions (green) for two univariate Gaussians (red and blue), as well as comparison of arithmetic Jensen-Shannon integrand against skew-geometric Jensen-Shannon integrand with $\alpha = 0.5$ for univariate Gaussians.



(a) $\mathrm{KL}(p(z)) \parallel (q(z|x)$ (b) $\mathrm{KL}(q(z|x) \parallel p(z))$ (c) JS. (d) $\mathrm{JS}^{\mathrm{G}_\alpha}$ ($\alpha = 0.5$).

Figure 2: Level sets for optimised bivariate Gaussians fit to data drawn from a mixture of Gaussians. $\mathrm{JS}^{\mathrm{G}_\alpha}$ with α naively set to the symmetric case $\alpha = 0.5$ leads to a more reasonable intermediate distribution which both tends towards the dominant mode and offers localised exploration

3.2 Variational autoencoder benchmarks

We present quantitative evaluation results following standard experimental protocols from the literature [5, 13, 38]. In this regard, VAEs are known to have a strong capacity to reproduce images when used in conjunction with convolutional encoders and decoders. For fair comparison, we follow Higgins et al. [13] in selecting a common neural architecture across experiments[4]. Although the margin for error ε in Equation (21) will vary with dataset and architecture, the point here is to standardise comparison and isolate the effect of the new divergence measure, rather than searching within architecture and hyperparameter spaces for the best performing model by some metric.

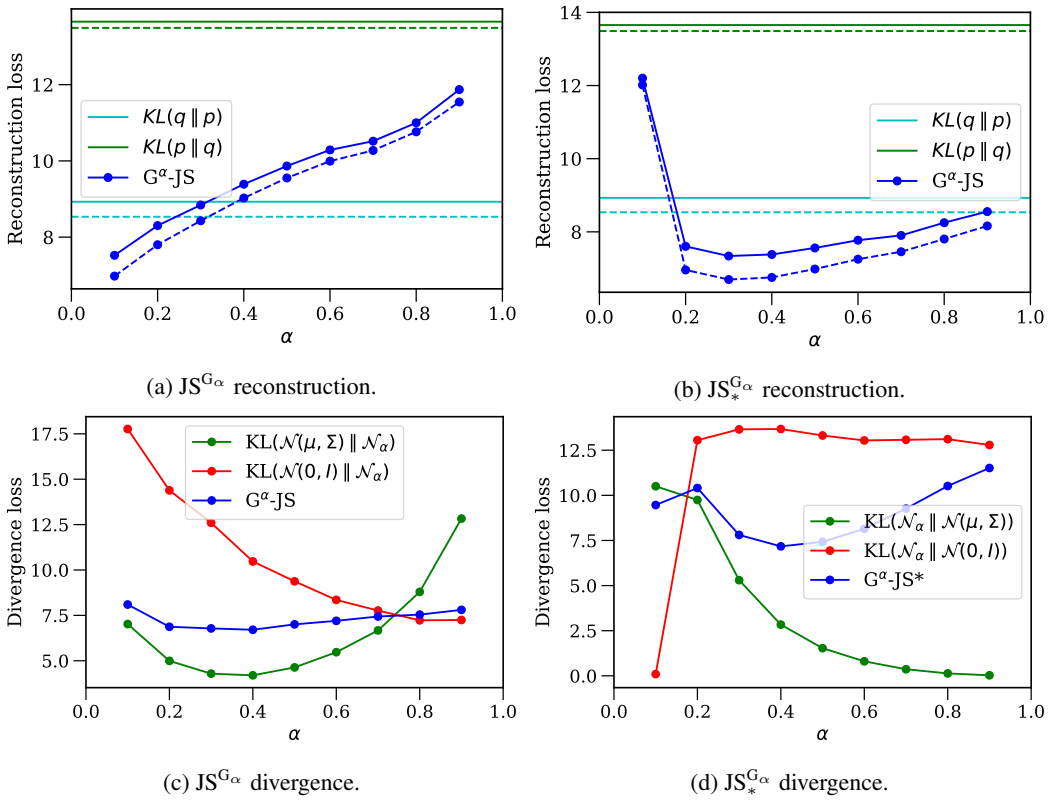
Throughout our experiments we make use of four standard benchmark datasets: **MNIST**, 28×28 black and white images of handwritten digits [21]; **Fashion-MNIST**, 28×28 black and white images of clothing [36]; **Chairs**, 64×64 black and white images of 3D chairs [1]; **dSprites** 64×64 black and white images of 2D shapes procedurally generated from 6 ground truth independent latent factors [25].

Influence of skew coefficient. In Figure 3, we demonstrate several immediately useful properties of skewing our divergence away from $\alpha = 0$ or $\alpha = 1$. Firstly, intermediate skew values of $\mathrm{JS}^{\mathrm{G}_\alpha}$ do not compromise reconstruction loss and remain considerably below $\mathrm{KL}(p(z) \parallel q(z|x))$, which we find to induce the expected mode collapse across datasets. Secondly, $\mathrm{JS}^{\mathrm{G}_\alpha}$ regularisation effectively generalises to unseen data, as can be seen by the small discrepancy between train and test set evaluation. Finally, there are ranges of α values which produce superior reconstructions when compared to either direction of KL for identical architectures.

Furthermore, Figure 3 indicates that $\mathrm{JS}^{\mathrm{G}_\alpha}_*$ outperforms $\mathrm{KL}(q(z|x) \parallel p(z))$ for nearly all values of α. We verify that the trend, $\mathrm{JS}^{\mathrm{G}_\alpha}$ *outperforms traditional divergences for* $\alpha < 0.3$ *and* $\mathrm{JS}^{\mathrm{G}_\alpha}_*$ *performs even better for nearly all* α, generalises across datasets in Table 1 and Supplementary Figures 7–9. In

[4]The specific model details are given in Appendix C

Figure 3c and 3d, we also include the corresponding divergence loss contributions to verify that $\text{JS}^{\text{G}_\alpha}$ does not simply minimise regularisation strength in order to improve reconstruction.



(a) $\text{JS}^{\text{G}_\alpha}$ reconstruction.

(b) $\text{JS}_*^{\text{G}_\alpha}$ reconstruction.

(c) $\text{JS}^{\text{G}_\alpha}$ divergence.

(d) $\text{JS}_*^{\text{G}_\alpha}$ divergence.

Figure 3: Reconstruction (top) and divergence (bottom) loss comparison for $\text{JS}^{\text{G}_\alpha}$ (left) and $\text{JS}_*^{\text{G}_\alpha}$ (right) against $\text{KL}(q(z|x) \parallel p(z))$ (VAE) and $\text{KL}(p(z) \parallel q(z|x))$ on the MNIST dataset. Throughout this work, dashed or full lines represent evaluation (sampling the mean with no variance) on the training or test sets, respectively. The comaprisons performed on the remaining three datasets are given in Appendix B.

In Table 1, we compare the naive symmetric case $\text{JS}^{\text{G}_{0.5}}$ against the skew value with the lowest reconstruction loss (selected from $\{0.1, \ldots, 0.9\}$) for $\text{JS}^{\text{G}_\alpha}$ and $\text{JS}_*^{\text{G}_\alpha}$, as well as baseline regularisation terms: $\text{KL}(q(z|x) \parallel p(z))$, $\text{KL}(p(z) \parallel q(z|x))$, β-VAE (with $\beta = 4$)[5] and MMD (with $\lambda = 500$). $\text{JS}_*^{\text{G}_\alpha}$ is clearly stronger than all baselines across datasets. We reinforce this point in Figure 4 where KL divergence fails to capture sharper reconstructions (such as delineating trouser legs or the heel of high-heels in the case of Fashion-MNIST) and MMD produces blurred reconstructions (we also tested $\lambda = 1000$ from [38] to no avail). We additionally extend qualitative results in Supplementary Figures 13–15. We sample each latent dimension at 10 equi-spaced points, while keeping the other 9 dimensions fixed in order to highlight the trends learnt by each dimension. As $\alpha \to 1$, the expected mode collapse occurs when approaching reverse KL across datasets, impeding reconstruction loss across more than a few modes. However, for α values close to 0, reverse KL images suffer from blur due to the aforementioned over-dispersion property.

reduce this

Figure 4 too small + quite confusing. R4: The paper leaves too much for the reader to interpret regarding Figure 4. Better caption and maybe labels

Generative capacity. In Figure 5, we demonstrate the generative capabilities when skewing $\text{JS}^{\text{G}_\alpha}$ across different α values. More specifically, we present the model evidence (ME) estimates for $\text{JS}^{\text{G}_\alpha}$ in comparison to forward KL, reverse KL, and MMD. ME estimates are generated by Monte Carlo estimation of the marginal distribution $p_\theta(x)$ with mean and 95% confidence intervals bootstrapped

[5]Details on the performance of β-VAEs for varying β is given in Appendix F

Divergence	MNIST	Fashion-MNIST	dSprites	Chairs
$\mathrm{KL}(q(z\|x) \parallel p(z))$	8.46	11.98	13.55	12.27
$\mathrm{KL}(p(z) \parallel q(z\|x))$	11.61	14.42	14.18	19.88
β-VAE ($\beta = 4$)	11.75	13.32	10.51	20.79
MMD ($\lambda = 500$)	13.19	11.10	11.87	18.85
$\mathrm{JS}^{\mathrm{G}_{0.5}}$	9.87	11.29	9.89	13.57
$\mathrm{JS}^{\mathrm{G}_\alpha}$	7.52 ($\alpha = 0.1$)	10.04 ($\alpha = 0.2$)	5.54 ($\alpha = 0.1$)	11.95 ($\alpha = 0.2$)
$\mathrm{JS}_*^{\mathrm{G}_\alpha}$	**7.34** ($\alpha = 0.3$)	**9.58** ($\alpha = 0.4$)	**4.97** ($\alpha = 0.5$)	**11.64** ($\alpha = 0.4$)

Table 1: Final model reconstruction error including optimal α for $\mathrm{JS}^{\mathrm{G}_\alpha}$ and $\mathrm{JS}_*^{\mathrm{G}_\alpha}$. The reconstruction errors for different α values for $\mathrm{JS}^{\mathrm{G}_\alpha}$ and $\mathrm{JS}_*^{\mathrm{G}_\alpha}$ are given in Appendix B

(a) $\mathrm{JS}_*^{\mathrm{G}_{0.4}}$. (b) $\mathrm{KL}(q(z|x) \parallel p(z))$. (c) MMD ($\lambda = 500$).

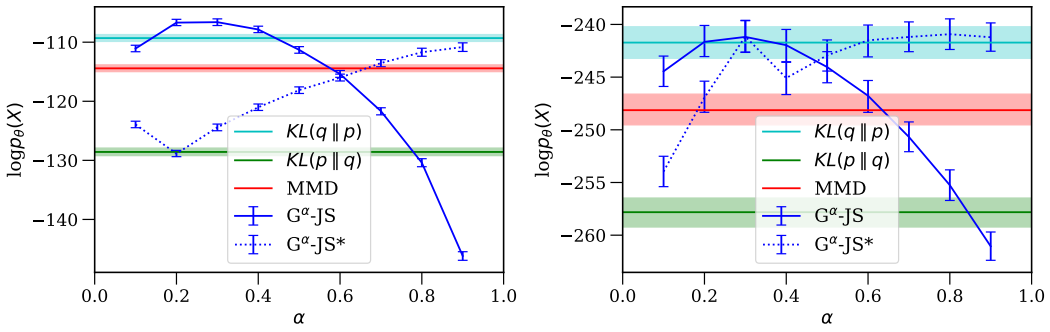
Figure 4: Latent space traversal for Fashion-MNIST. Each row represents a latent dimension and each column represents an equidistant point in the traversal. Analyses for dSprites and Chairs are given in AppendixG.

from 1000 resamples of estimated batch evidence across 100 *test* set batches. We emphasise here that we are not looking for state-of-the-art results, but *relative* improvement which isolates the impact of the proposed regularisation and extends our analysis of $\mathrm{JS}^{\mathrm{G}_\alpha}$. We see that in the case of MNIST (Figure 5a) the increased reconstructive power of $\mathrm{JS}_*^{\mathrm{G}_\alpha}$ does come at a cost to generative performance, however this trend is not consistent in the noisier Fashion-MNIST dataset (Figure 5b). Nevertheless, note that the reconstruction error of $\mathrm{JS}_*^{\mathrm{G}_\alpha}$ for $\alpha > 0.8$ and $\alpha > 0.6$, in the case of MNIST and Fashion-MNIST, respectively, is still lower than the benchmarks. We also find $0.15 < \alpha < 0.4$ for $\mathrm{JS}^{\mathrm{G}_\alpha}$ is competitive with or better than all alternatives on both datasets.

Taken all together, we make several pragmatic suggestions for selecting α values when using our variant of $\mathrm{JS}^{\mathrm{G}_\alpha}$ or its dual form. Firstly, when using $\mathrm{JS}^{\mathrm{G}_\alpha}$, lower α values are to be preferred, this goes some way to explaining the poor performance of the initial attempts to use $\mathrm{JS}^{\mathrm{G}_{0.5}}$ in the literature (see Section 4). Whereas for the dual divergence, although lower α values ($\alpha <= 0.5$) lead to the lowest reconstruction error, higher α values ($\alpha > 0.6$) exhibit better generative capabilities while having lower reconstruction error than the benchmarks. Therefore, the symmetric case is a reasonably strong choice. Moreover, the plots of reconstruction loss against α clearly demonstrate a strong correlation between train and test set performance. This can be applied in practice, by selecting an optimal value of α using the training performance, circumventing the need for a separate validation set.

4 Related work

$\mathrm{JS}^{\mathrm{G}_\alpha}$-VAEs build upon traditional VAEs [17, 34], with a regularisation constraint inspired by recent work on closed-form expressions for statistical divergences [30, 32]. $\mathrm{JS}^{\mathrm{G}_\alpha}$-VAEs, offer simpler and



(a) MNIST

(b) Fashion-MNIST

Figure 5: Log model evidence for $\text{JS}^{\text{G}_\alpha}$ and $\text{JS}_*^{\text{G}_\alpha}$ across different α values compared against $\text{KL}(q(z|x) \parallel p(z))$, $\text{KL}(p(z) \parallel q(z|x))$ and MMD on the (a) MNIST and (b) Fashion-MNIST datasets.

more intuitive regularisation by skewing the intermediate distribution, allowing interpolation between forward and reverse KL divergence, and therefore combatting the issue of posterior collapse [24]. In this regard, our work is related to approaches that address this issue through KL annealing during training [5, 14]. In a more general sense, this work is also related to other approaches that utilise various statistical divergences and distances for latent space regularisation as an alternative to the conventional KL divergence [8, 12, 22, 37, 38].

Since its recent introduction, [2] used $\text{JS}^{\text{G}_{0.5}}$ as a plug-and-play replacement for JS divergence with little success, while [35] used $\text{JS}^{\text{G}_{0.5}}$ to decompose and estimate a multimodal ELBO loss. In contrast to these papers, we do not overlook the potential of $\text{JS}^{\text{G}_\alpha}$. We reverse the intermediate distribution parameterisation, allowing a principled interpolation of forward and reverse KL, we simplify the subsequent closed-form loss to that needed for VAEs, and we demonstrate improved empirical performance against several baselines (application, rather than the theory of [31]). Our more natural parameterisation and pragmatic advice on how to properly use the skew parameter α ultimately lead to better image reconstruction. We are not aware of any prior work exploring the dual form $\text{JS}_*^{\text{G}_\alpha}$.

5 Conclusion

Prior work assumed that no tractable interpolation existed between forward and reverse KL for multivariate Gaussians. We have overcome this with our variant of $\text{JS}^{\text{G}_\alpha}$, before translating it to the variational learning setting with $\text{JS}^{\text{G}_\alpha}$-VAE. The benefits of our variant of $\text{JS}^{\text{G}_\alpha}$ include symmetry (at $\alpha = 0.5$) and having closed-form expression. Alongside this, we have demonstrated that the advantages of its role in VAEs include quantitatively and qualitatively better reconstructions than several baselines. Although we accept that use of "vanilla" VAEs may not out-compete some of the leading flow and GAN based architectures, we believe our regularisation mechanism addresses the trade-off between zero-avoidance and zero-forcing in latent space, which goes some way to bridge this gap while being intuitive in both divergence and distribution space. Our experiments demonstrate that the flexibility accorded to VAEs by skewing $\text{JS}^{\text{G}_\alpha}$ is worth considering across a broad range of applications.

Broader Impact

For the statistics community, our introduction of the alternative $\mathrm{JS}^{\mathrm{G}_{\alpha'}}$ and $\mathrm{JS}_*^{\mathrm{G}_{\alpha'}}$, rather than the "original" $\mathrm{JS}^{\mathrm{G}_{\alpha}}$ and $\mathrm{JS}_*^{\mathrm{G}_{\alpha}}$, immediately presents a benefit as a more intuitive interpolation through divergence and distribution space. As we have shown the benefits of such an interpolation on the task of image reconstruction, the first impact of our model lies in better image compression and generation from latent samples. However, in a more general setting, VAEs present multiple impactful opportunities.

Applications include compression (of any data type), generation of new samples in fields with data paucity, as well as extraction of underlying relationships. As our exploration of the $\mathrm{JS}^{\mathrm{G}_{\alpha}}$ family of VAEs has improved performance, after translation to data types with other structures, our VAE could be used for all of these applications. Our experiments also indicate strong regions for the skew parameter α which could be used as a standard regularisation mechanism across variational learning.

In settings with sensitive data, all of these applications bear some risks. As VAEs provide a form of *lossy* compression, in healthcare and social settings there is the risk of misrepresenting personal information in latent space. In areas of data paucity, without additional constraints, VAEs may generate samples which are unrealistic and severely bias any downstream training. Finally, when using VAEs in science, to extract underlying associations, it remains important to analyse the true meaning of any independent components extracted, rather than taking these rules at face value.

Acknowledgments and Disclosure of Funding

We thank Cristian Bodnar, Cătălina Cangea, Ben Day, Felix Opolka, Emma Rocheteau, Ramon Viñas Torne and Duo Wang from the Department of Computer Science and Technology, University of Cambridge, for their helpful comments. We would like to also thank the reviewers for their constructive feedback and efforts towards improving our paper. We acknowledge the support of The Mark Foundation for Cancer Research and Cancer Research UK Cambridge Centre [C9685/A25177] for N.S. The authors declare no competing interests.

References

[1] Mathieu Aubry, Daniel Maturana, Alexei Efros, Bryan Russell, and Josef Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014.

[2] Vikash Balasubramanian, Ivan Kobyzev, Hareesh Bahuleyan, Ilya Shapiro, and Olga Vechtomova. Polarized-vae: Proximity based disentangled representation learning for text generation. *arXiv preprint arXiv:2004.10809*, 2020.

[3] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[4] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[5] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

[6] Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018.

[7] Bin Dai and David Wipf. Diagnosing and enhancing vae models. *arXiv preprint arXiv:1903.05789*, 2019.

[8] Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems*, pages 2732–2741, 2017.

[9] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pages 881–889, 2015.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[12] James Hensman, Max Zwießele, and Neil Lawrence. Tilted variational bayes. In *Artificial Intelligence and Statistics*, pages 356–364, 2014.

[13] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *Iclr*, 2(5):6, 2017.

[14] Chin-Wei Huang, Shawn Tan, Alexandre Lacoste, and Aaron C Courville. Improving explorability in variational inference with annealed variational objectives. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9701–9711. Curran Associates, Inc., 2018.

[15] William Karush. Minima of functions of several variables with inequalities as side constraints. *M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago*, 1939.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR 2014*, 2014. URL abs/1312.6114.

[18] Harold W Kuhn and Albert W Tucker. Nonlinear programming. In *Traces and emergence of nonlinear programming*, pages 247–258. Springer, 2014.

[19] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[20] Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 29–37, 2011.

[21] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[22] Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1073–1081. Curran Associates, Inc., 2016.

[23] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[24] James Lucas, George Tucker, Roger Baker Grosse, and Mohammad Norouzi. Understanding posterior collapse in generative latent variable models. In *DGS@ICLR*, 2019.

[25] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[26] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[27] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[28] Constantin Niculescu and Lars-Erik Persson. *Convex functions and their applications*. Springer, 2006.

[29] Frank Nielsen. A family of statistical symmetric divergences based on jensen's inequality. *arXiv preprint arXiv:1009.4004*, 2010.

[30] Frank Nielsen. On the jensen–shannon symmetrization of distances relying on abstract means. *Entropy*, 21(5):485, 2019.

[31] Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards. *arXiv preprint arXiv:0911.4863*, 2009.

[32] Tomohiro Nishiyama. Generalized bregman and jensen divergences which include some f-divergences. *arXiv preprint arXiv:1808.06148*, 2018.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[34] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286. PMLR, 2014.

[35] Thomas Sutter, Imant Daunhawer, and Julia E Vogt. Multimodal generative learning utilizing jensen-shannon divergence. In *Workshop on Visually Grounded Interaction and Language at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

[36] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[37] Mingtian Zhang, Thomas Bird, Raza Habib, Tianlin Xu, and David Barber. Variational f-divergence minimization. *arXiv preprint arXiv:1907.11891*, 2019.

[38] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA*, pages 5885–5892. AAAI Press, Palo Alto, CA, USA, 2019.

A Proofs

A.1 Proof of proposition 1

Proof. We first present the more general case of distributions p and q permitting a geometric mean distribution (e.g. p and q members of the exponential family), as we believe this more general case to be of note.

$$\begin{aligned}
\mathrm{JS}^{\mathrm{G}_{\alpha'}} &= (1-\alpha)\mathrm{KL}\left(p \parallel G_{\alpha'}(p,q)\right) + \alpha\mathrm{KL}\left(q \parallel G_{\alpha'}(p,q)\right) && (27)\\
&= (1-\alpha)\mathrm{KL}\left(p \parallel p^{\alpha}q^{1-\alpha}\right) + \alpha\mathrm{KL}\left(q \parallel p^{\alpha}q^{1-\alpha}\right) && (28)\\
&= (1-\alpha)\int_x p\log\left[\frac{p}{p^{\alpha}q^{1-\alpha}}\right]dx + \alpha\int_x q\log\left[\frac{q}{p^{\alpha}q^{1-\alpha}}\right]dx && (29)\\
&= (1-\alpha)^2\int_x p\log\left[\frac{p}{q}\right]dx + \alpha^2\int_x q\log\left[\frac{q}{p}\right]dx && (30)\\
&= (1-\alpha)^2\mathrm{KL}(p \parallel q) + \alpha^2\mathrm{KL}(q \parallel p) && (31)
\end{aligned}$$

Therefore, the respective cases disappear in the limits $\alpha \to 0$ and $\alpha \to 1$ and for $\mathrm{JS}^{\mathrm{G}_{\alpha'}}$ we have, in fact, recovered an equivalence between linear scaling in distribution space and quadratic scaling in the space of divergences.

The dual case $\mathrm{JS}_*^{\mathrm{G}_{\alpha'}}$ does not simplify in the same way because the geometric mean term lies outside of the logarithm. However, instead we have

$$\begin{aligned}
\mathrm{JS}_*^{\mathrm{G}_{\alpha'}} &= (1-\alpha)\mathrm{KL}\left(G_{\alpha'}(p,q) \parallel p\right) + \alpha\mathrm{KL}\left(G_{\alpha'}(p,q) \parallel q\right) && (32)\\
&= (1-\alpha)\mathrm{KL}\left(p^{\alpha}q^{1-\alpha} \parallel p\right) + \alpha\mathrm{KL}\left(p^{\alpha}q^{1-\alpha} \parallel q\right) && (33)\\
&= (1-\alpha)\int_x p^{\alpha}q^{1-\alpha}\log\left[\frac{p^{\alpha}q^{1-\alpha}}{p}\right]dx + \alpha\int_x p^{\alpha}q^{1-\alpha}\log\left[\frac{p^{\alpha}q^{1-\alpha}}{q}\right]dx && (34)\\
&= (1-\alpha)^2\int_x p^{\alpha}q^{1-\alpha}\log\left[\frac{q}{p}\right]dx + \alpha^2\int_x p^{\alpha}q^{1-\alpha}\log\left[\frac{p}{q}\right]dx. && (35)
\end{aligned}$$

The final step is to recognise the two limits

$$\lim_{\alpha\to 0}\left[p^{\alpha}q^{1-\alpha}\right] = q \qquad\qquad \lim_{\alpha\to 1}\left[p^{\alpha}q^{1-\alpha}\right] = p, \tag{36}$$

mean that we recover

$$\lim_{\alpha\to 0}\left[\mathrm{JS}_*^{\mathrm{G}_{\alpha'}}\right] = \mathrm{KL}\left(\mathcal{N}_2 \parallel \mathcal{N}_1\right) \qquad\qquad \lim_{\alpha\to 1}\left[\mathrm{JS}_*^{\mathrm{G}_{\alpha'}}\right] = \mathrm{KL}\left(\mathcal{N}_1 \parallel \mathcal{N}_2\right). \tag{37}$$

□

Overall, although the limiting cases are reversed between $\mathrm{JS}^{\mathrm{G}_{\alpha'}}$ and $\mathrm{JS}_*^{\mathrm{G}_{\alpha'}}$, we note that the approach to either limiting case is distinct and comes with its own benefits through the weighting (non-logarithmic) term used in the integrand.

A.2 Proof of proposition 2

We choose to prove proposition 1 via reduction of the form in Equation (9) , although we note it is also reasonable to simply follow through the weighted sum in Equation (8).

Proof. After defining $\Sigma_{ii} = \sigma_i^2$, $(\Sigma_\alpha)_{ii} = \sigma_{\alpha,i}^2$ and $(\mu_\alpha)_i = \mu_{\alpha,i}$, it is apparent $\Sigma_2 = I$ gives

$$\sigma_{\alpha,i}^2 = \frac{1}{((1-\alpha)\sigma_i^2 + \alpha)}, \tag{38}$$

and $\mu_2 = 0$ (the zero vector) gives

$$\mu_{\alpha,i} = \sigma_{\alpha,i}^2\left((1-\alpha)\frac{\mu_i}{\sigma_i^2}\right) \tag{39}$$

We can then reduce Equation (9) using diagonal matrix properties

$$\mathrm{JS}^{\mathrm{G}_\alpha}\left(\mathcal{N}_1 \parallel \mathcal{N}_2\right) = \frac{1}{2}\left(\sum_{i=1}^{n} \frac{1}{\sigma_{\alpha,i}^2}\left((1-\alpha)\sigma_i^2 + \alpha\right) + \log\left[\frac{\prod_{i=1}^{n}\sigma_{\alpha,i}^2}{\prod_{i=1}^{n}\left(\sigma_i^2\right)^{1-\alpha}}\right]\right. \tag{40}$$

$$\left. + \frac{(1-\alpha)(\mu_{\alpha,i}-\mu_i)^2}{\sigma_{\alpha,i}^2} + \frac{\alpha\mu_{\alpha,i}^2}{\sigma_{\alpha,i}^2} - n\right), \tag{41}$$

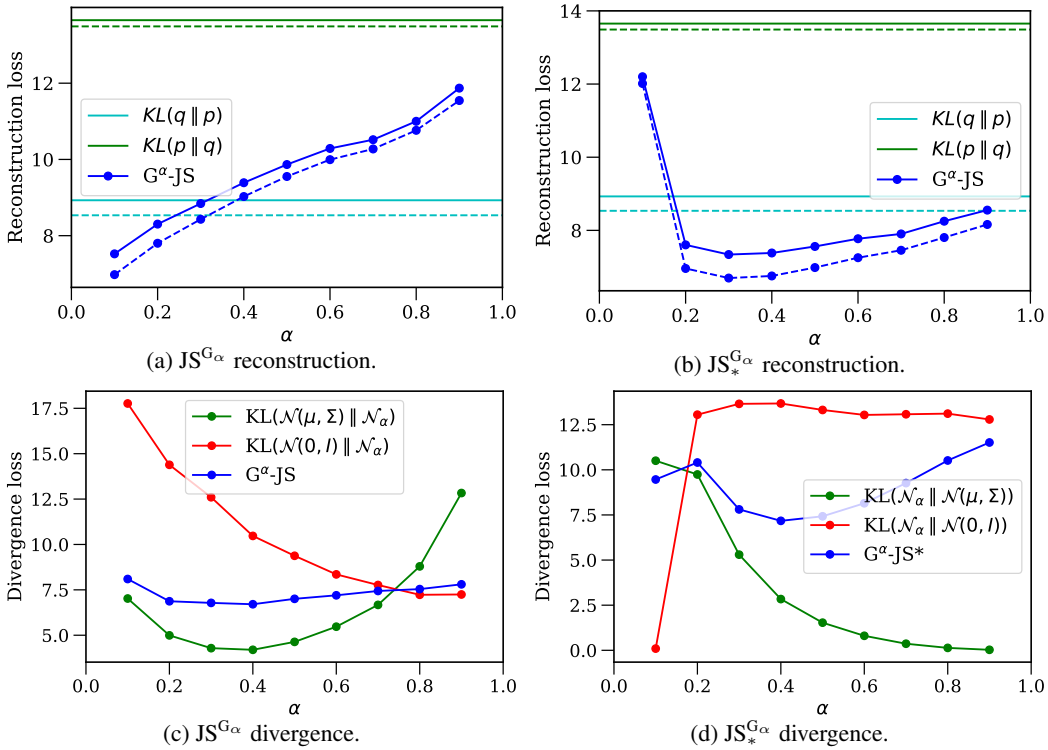
and application of log laws recovers Equation (23).

The proof of the dual form in Equation (25) is carried out similarly. □

B Additional training and evaluation information

Divergence	**MNIST**	**Fashion-MNIST**	**dSprites**	**Chairs**
$\mathrm{KL}(q(z\|x) \parallel p(z))$	8.46	11.98	13.55	12.27
$\mathrm{KL}(p(z) \parallel q(z\|x))$	11.61	14.42	14.18	19.88
β-VAE ($\beta = 4$)	11.75	13.32	10.51	20.79
β-VAE ($\beta = 0.25$)	8.09	*9.07*	10.39	14.09
MMD ($\lambda = 500$)	13.19	11.10	11.87	18.85
$\mathrm{JS}^{\mathrm{G}_{0.1}}$	**7.52**	10.04	**6.63**	12.62
$\mathrm{JS}^{\mathrm{G}_{0.2}}$	8.30	**10.04**	7.50	**11.95**
$\mathrm{JS}^{\mathrm{G}_{0.3}}$	8.84	10.50	8.56	12.40
$\mathrm{JS}^{\mathrm{G}_{0.4}}$	9.39	10.93	9.16	12.96
$\mathrm{JS}^{\mathrm{G}_{0.5}}$	9.87	11.29	9.89	13.57
$\mathrm{JS}^{\mathrm{G}_{0.6}}$	10.28	11.72	10.38	14.15
$\mathrm{JS}^{\mathrm{G}_{0.7}}$	10.51	12.09	10.80	14.68
$\mathrm{JS}^{\mathrm{G}_{0.8}}$	11.00	12.44	11.40	15.48
$\mathrm{JS}^{\mathrm{G}_{0.9}}$	11.87	13.21	12.05	16.27
$\mathrm{JS}_*^{\mathrm{G}_{0.1}}$	12.20	13.52	5.54	15.53
$\mathrm{JS}_*^{\mathrm{G}_{0.2}}$	7.60	10.90	5.18	13.06
$\mathrm{JS}_*^{\mathrm{G}_{0.3}}$	**7.34**	10.51	5.06	12.09
$\mathrm{JS}_*^{\mathrm{G}_{0.4}}$	7.38	**9.58**	5.17	**11.64**
$\mathrm{JS}_*^{\mathrm{G}_{0.5}}$	7.56	9.80	**4.97**	11.75
$\mathrm{JS}_*^{\mathrm{G}_{0.6}}$	7.77	10.01	5.30	12.07
$\mathrm{JS}_*^{\mathrm{G}_{0.7}}$	7.90	10.34	5.23	12.53
$\mathrm{JS}_*^{\mathrm{G}_{0.8}}$	8.25	10.84	5.42	13.11
$\mathrm{JS}_*^{\mathrm{G}_{0.9}}$	8.55	11.40	5.74	13.52

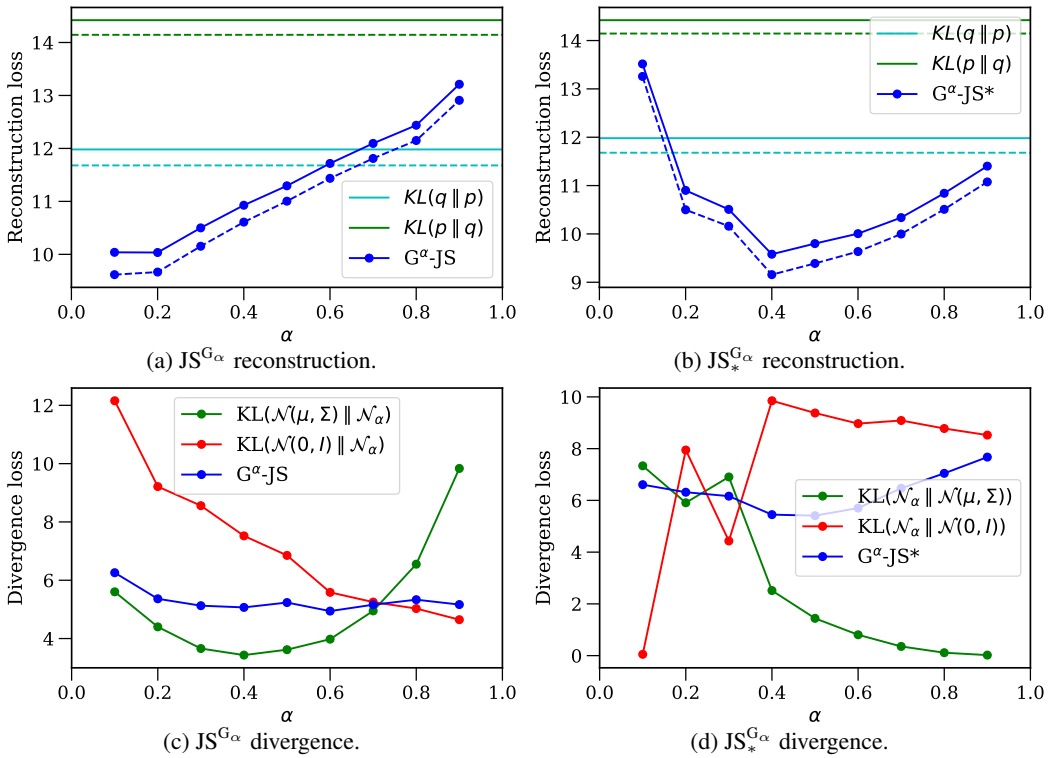
Table 2: Final model reconstruction error for different α values for $\mathrm{JS}^{\mathrm{G}_\alpha}$ and $\mathrm{JS}_*^{\mathrm{G}_\alpha}$.



(a) $\mathrm{JS}^{\mathrm{G}_\alpha}$ reconstruction. (b) $\mathrm{JS}_*^{\mathrm{G}_\alpha}$ reconstruction.

(c) $\mathrm{JS}^{\mathrm{G}_\alpha}$ divergence. (d) $\mathrm{JS}_*^{\mathrm{G}_\alpha}$ divergence.

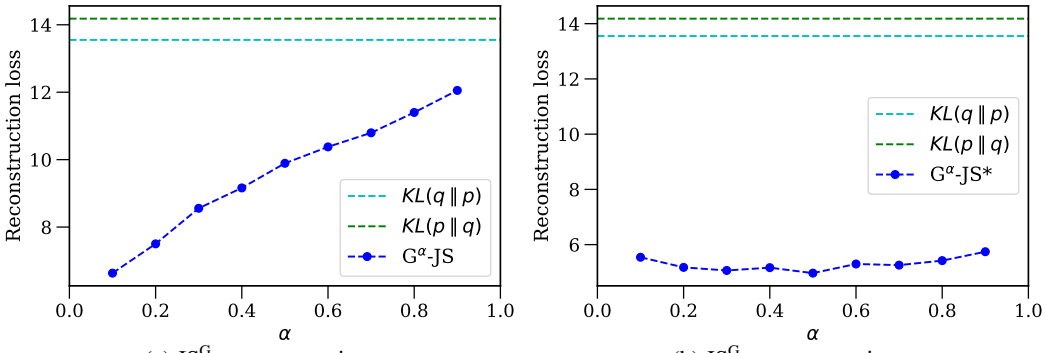
Figure 6: Breakdown of final model loss components on the MNIST dataset.



(a) $\mathrm{JS}^{\mathrm{G}_\alpha}$ reconstruction. (b) $\mathrm{JS}_*^{\mathrm{G}_\alpha}$ reconstruction.

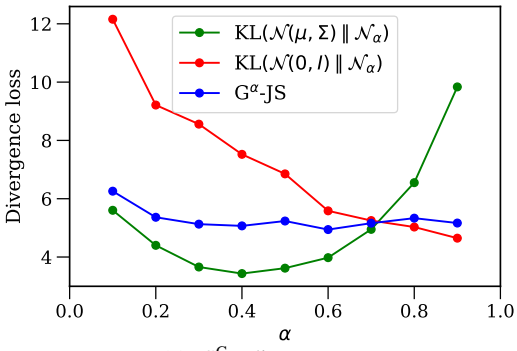
(c) $\mathrm{JS}^{\mathrm{G}_\alpha}$ divergence. (d) $\mathrm{JS}_*^{\mathrm{G}_\alpha}$ divergence.

Figure 7: Breakdown of final model loss on the Fashion-MNIST dataset.

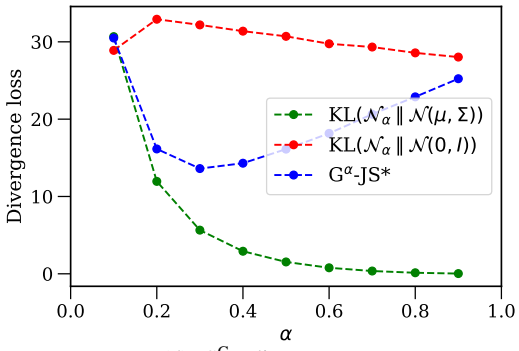


(a) $\mathrm{JS}^{\mathrm{G}_\alpha}$ reconstruction.

(b) $\mathrm{JS}_*^{\mathrm{G}_\alpha}$ reconstruction.

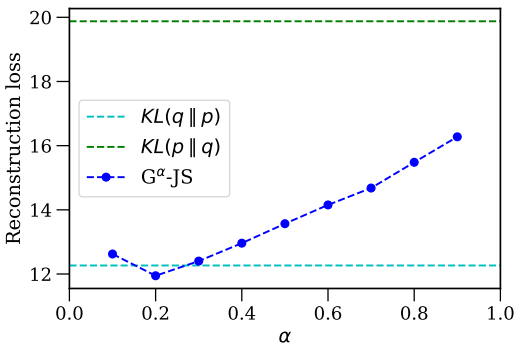


(c) $\mathrm{JS}^{\mathrm{G}_\alpha}$ divergence.

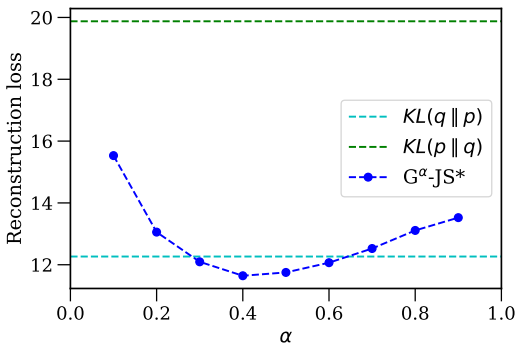


(d) $\mathrm{JS}_*^{\mathrm{G}_\alpha}$ divergence.

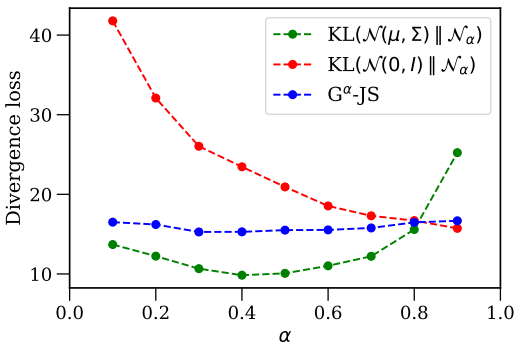
Figure 8: Breakdown of final model loss components on the dSprites dataset.



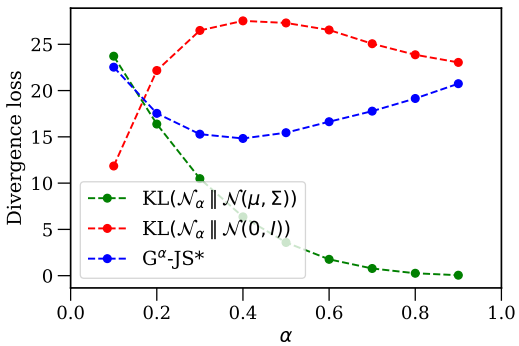
(a) $\mathrm{JS}^{\mathrm{G}_\alpha}$ reconstruction.



(b) $\mathrm{JS}_*^{\mathrm{G}_\alpha}$ reconstruction.



(c) $\mathrm{JS}^{\mathrm{G}_\alpha}$ divergence.



(d) $\mathrm{JS}_*^{\mathrm{G}_\alpha}$ divergence.

Figure 9: Breakdown of final model loss components on the Chairs dataset.

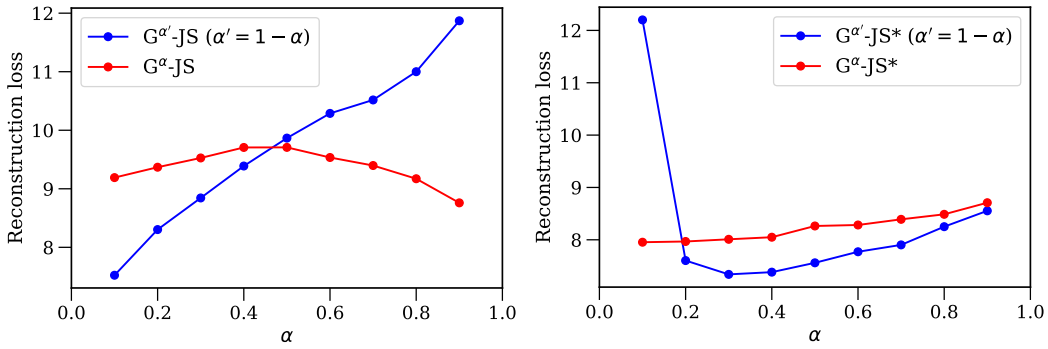
C Model details

We use the architectures specified in Table 3 throughout experiments. We pad 28x28x1 images to 32x32x1 with zeros as we found resizing images negatively affected performance. We use a learning rate of 1e-4 throughout and use batch size 64 and 256 for the two MNIST variants and the other datasets respectively. Where not specified (e.g. momentum coefficients in Adam [16]), we use the default values from PyTorch [33]. The only architectural change we make between datasets is an additional convolutional (and transpose convolutional) layer for encoding (and decoding) when inputs are 64x64x1 instead of 32x32x1. We train dSprites for 30 epochs and all other datasets for 100 epochs.

Dataset	**Stage**	**Architecture**
MNIST	Input	28x28x1 zero padded to 32x32x1.
	Encoder	Repeat Conv 32x4x4 for 3 layers (stride 2, padding 1). FC 256, FC 256. ReLU activation.
	Latents	10.
	Decoder	FC 256, FC 256, Repeat Deconv 32x4x4 for 3 layers (stride 2, padding 1). ReLU activation, Sigmoid. MSE.
Fashion-MNIST	Input	28x28x1 zero padded to 32x32x1.
	Encoder	Repeat Conv 32x4x4 for 3 layers (stride 2, padding 1). FC 256, FC 256. ReLU activation.
	Latents	10.
	Decoder	FC 256, FC 256, Repeat Deconv 32x4x4 for 3 layers (stride 2, padding 1). ReLU activation, Sigmoid. Bernoulli.
dSprites	Input	64x64x1.
	Encoder	Repeat Conv 32x4x4 for 4 layers (stride 2, padding 1). FC 256, FC 256. ReLU activation.
	Latents	10.
	Decoder	FC 256, FC 256, Repeat Deconv 32x4x4 for 4 layers (stride 2, padding 1). ReLU activation, Sigmoid. Bernoulli.
Chairs	Input	64x64x1.
	Encoder	Repeat Conv 32x4x4 for 4 layers (stride 2, padding 1). FC 256, FC 256. ReLU activation.
	Latents	32.
	Decoder	FC 256, FC 256, Repeat Deconv 32x4x4 for 4 layers (stride 2, padding 1). ReLU activation, Sigmoid. Bernoulli.

Table 3: Detail of model architectures.

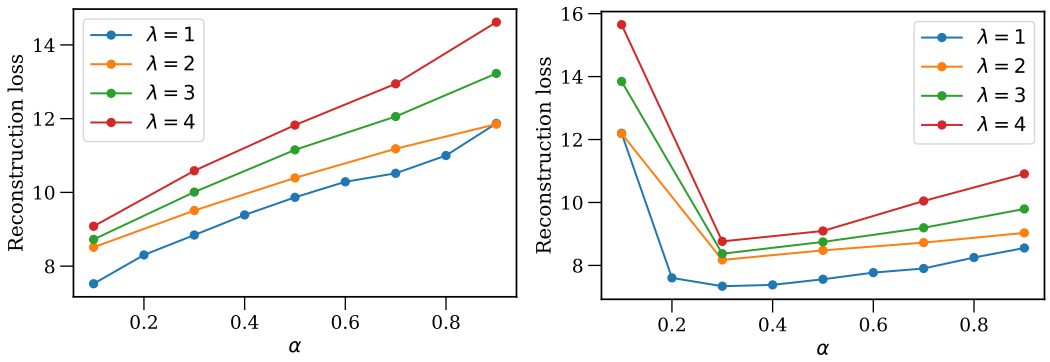
D $\mathrm{JS}^{\mathrm{G}_{\alpha'}}$ vs. $\mathrm{JS}^{\mathrm{G}_{\alpha}}$



(a) $\mathrm{JS}^{\mathrm{G}_{\alpha}}$. (b) $\mathrm{JS}_{*}^{\mathrm{G}_{\alpha}}$.

Figure 10: Comparison of the original $\mathrm{JS}^{\mathrm{G}_{\alpha}}$ and our variant, $\mathrm{JS}^{\mathrm{G}_{\alpha'}}$, on the MNIST dataset.

E Influence of the λ parameter on the performance of $\mathrm{JS}^{\mathrm{G}_\alpha}$-VAEs and $\mathrm{JS}_*^{\mathrm{G}_\alpha}$-VAEs



(a) $\mathrm{JS}^{\mathrm{G}_\alpha}$. (b) $\mathrm{JS}_*^{\mathrm{G}_\alpha}$.

Figure 11: Comparison of the reconstruction loss of $\mathrm{JS}^{\mathrm{G}_\alpha}$-VAEs and $\mathrm{JS}_*^{\mathrm{G}_\alpha}$-VAEs for different values of λ, on the MNIST dataset.

F Performance of β-VAEs for varying β

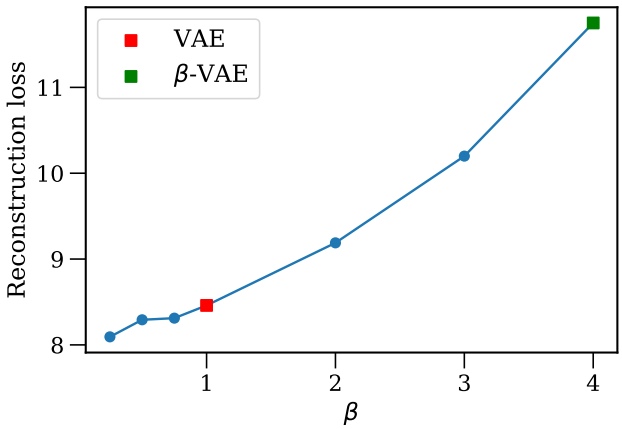
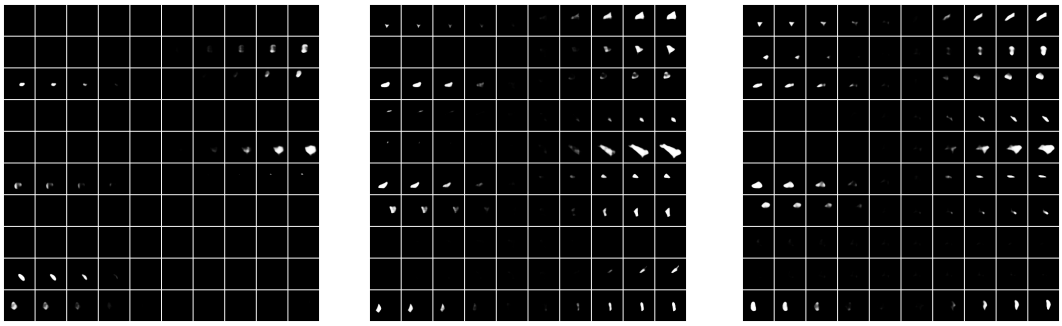


Figure 12: Comparison of the reconstruction loss of β-VAEs for different values of β, on the MNIST dataset.

G Latent samples

(a) $\mathrm{JS}_*^{\mathrm{G}_{0.1}}$. (b) $\mathrm{JS}_*^{\mathrm{G}_{0.4}}$. (c) $\mathrm{JS}^{\mathrm{G}_{0.9}}$.

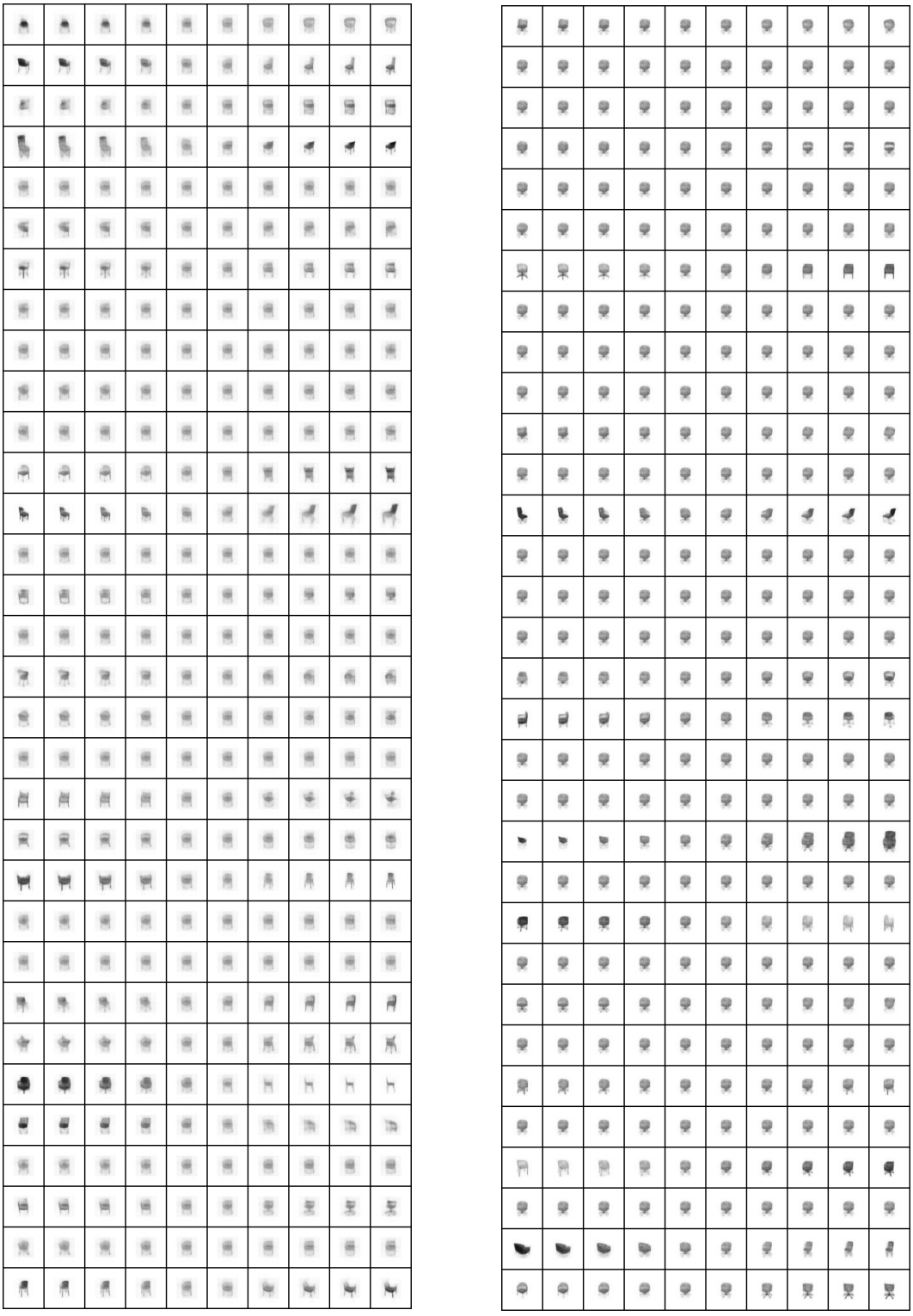
Figure 13: Latent space traversal of Fashion-MNIST for different skew values of $\mathrm{JS}_*^{\mathrm{G}_\alpha}$.



(a) $\mathrm{JS}_*^{\mathrm{G}_{0.1}}$ (b) $\mathrm{JS}_*^{\mathrm{G}_{0.5}}$ (c) $\mathrm{JS}_*^{\mathrm{G}_{0.9}}$

(d) KL.

Figure 14: Latent space traversal dSprites for different skew values and KL divergence.



(a) $\mathrm{JS}_*^{\mathrm{G}_{0.4}}$. (b) KL.

Figure 15: Latent space traversal for the Chairs dataset (32 latent dimensions).