α-VAEs : Optimising variational inference by learning data-dependent divergence skew

Anonymous Authors[1]

Abstract

The *skew-geometric Jensen-Shannon divergence* $\left(\mathrm{JS}^{\mathrm{G}_\alpha}\right)$ allows for an intuitive interpolation between forward and reverse Kullback-Leibler (KL) divergence based on the skew parameter α. While the benefits of the skew in $\mathrm{JS}^{\mathrm{G}_\alpha}$ are clear—balancing forward/reverse KL in a comprehensible manner—the choice of optimal skew remains opaque and requires an expensive grid search. In this paper we introduce α-VAEs, which extend the $\mathrm{JS}^{\mathrm{G}_\alpha}$ variational autoencoder by allowing for learnable, and therefore data-dependent, skew. We motivate the use of a parameterised skew in the dual divergence by analysing trends dependent on data complexity in synthetic examples. We also prove and discuss the dependency of the divergence minimum on the input data and encoder parameters, before empirically demonstrating that this dependency does not reduce to either direction of KL divergence for benchmark datasets. Finally, we demonstrate that optimised skew values consistently converge across a range of initial values and provide improved denoising and reconstruction properties. These render α-VAEs an efficient and practical modelling choice across a range of tasks, datasets, and domains.

1. Introduction

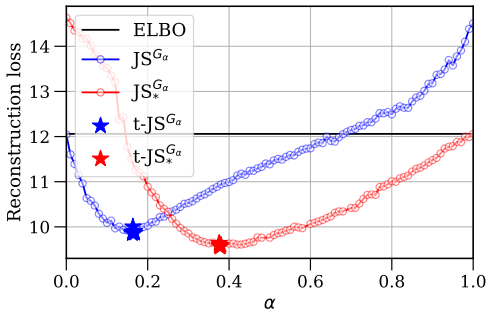
As variational inference (VI) progresses, state of the art Variational AutoEncoders (VAEs) increase in complexity (Vahdat & Kautz, 2020; Child, 2020) while continuing to minimise the Evidence Lower BOund (ELBO) (Blei et al., 2017). Compared to generative adversarial networks (GANs) (Goodfellow et al., 2014), VAEs necessitate less stringent and problem-dependent training regimes, and compared to autoregressive models (Larochelle & Murray, 2011; Germain et al., 2015) (that can be interpreted as instances of very deep VAEs (Child, 2020)) are less computationally expensive and more efficient to sample. VAE learning requires optimisation of an objective which balances the quality of decoded reconstructions from encoded representations, with a regularising divergence term penalising latent space deviations from a fixed prior distribution.

The divergence term should be strongly coupled to the predominant assumption of latent variables parameterised by a multivariate Gaussian, $p_\theta(z) = N(\mu, \sigma^2)$ with $z, \mu, \sigma \in \mathbb{R}^n$, approximated by $q_\phi(z|x)$ with $x \in \mathbb{R}^m$ and $n \leq m$. For instance, in the original VAE (Kingma & Welling, 2014), KL divergence (Kullback & Leibler, 1951) naturally constrained the variational distribution to an isotropic Gaussian unit ball $\mathrm{KL}\left(q_\phi(z|x) \parallel \mathcal{N}(0, I)\right)$, despite unfavourable properties (Bishop, 2006), such as unboundedness and asymmetry. Moreover, KL does not capitalise on the full flexibility of the wider family of exponential distributions, a recent direction which has tightened the ELBO (Brekelmans et al., 2020; Masrani et al., 2019) and rendered VAE divergence regularisation more interpretable in distribution space via *skew-geometric Jensen-Shannon* $\left(\mathrm{JS}^{\mathrm{G}_\alpha}\right)$ divergence (Nielsen, 2019; Deasy et al., 2020). We henceforth refer to VAEs with skewed geometric divergences as α-VAEs, subsuming β-VAEs.

Divergence skew in VAEs balances the contrasting properties of forward and reverse KL (such as zero-avoidance/forcing) and circumvents opaque divergence terms by interpolating between them (Bishop, 2006). However, an expensive grid search over skew values fixed through training is necessary to optimise for tasks such as image reconstruction. This is particularly problematic as the link between optimal skew and dataset properties is not clear and is not easily resolved. Moreover, when skew is treated as static, the divergence constraint does not change during training and therefore does not reflect improvements in the encoder and its embeddings. Instead, an improved optimisation would update the divergence skew relative to these factors without using prior knowledge or compromising performance.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by INNF+ 2021. Do not distribute.



(a) α-VAEs on the MNIST dataset.

Divergence	KL expression	Train skew
KL (ELBO)	$\mathrm{KL}(q \parallel p)$	✗
α-VAEs		
JS^{G_α}	$(1-\alpha)\mathrm{KL}\left(p \parallel p^\alpha q^{1-\alpha}\right) + \alpha\mathrm{KL}\left(q \parallel p^\alpha q^{1-\alpha}\right)$	✗
$\mathrm{JS}_*^{G_\alpha}$	$(1-\alpha)\mathrm{KL}\left(p^\alpha q^{1-\alpha} \parallel p\right) + \alpha\mathrm{KL}\left(p^\alpha q^{1-\alpha} \parallel q\right)$	✗
t-JS^{G_α}	$(1-\alpha)\mathrm{KL}\left(p \parallel p^\alpha q^{1-\alpha}\right) + \alpha\mathrm{KL}\left(q \parallel p^\alpha q^{1-\alpha}\right)$	✓
t-$\mathrm{JS}_*^{G_\alpha}$	$(1-\alpha)\mathrm{KL}\left(p^\alpha q^{1-\alpha} \parallel p\right) + \alpha\mathrm{KL}\left(p^\alpha q^{1-\alpha} \parallel q\right)$	✓

(b) Breakdown of α-VAE divergences.

Figure 1. The influence of divergence skew, α, on test set image reconstruction (measured in mean squared error). (Deasy et al., 2020) used an expensive grid search to fix divergence skew through training. Our separate optimisation of α in JS^{G_α} (prefix t-), as a competing objective to reconstruction loss, leads to strong performance and substantially improves over reverse KL from any initial skew. We plot optimal α, denoted by a star, for 5 training seeds which (visually) converge to the same point.

Contributions. To overcome these issues, we extend (Deasy et al., 2020) and introduce α-VAEs, by allowing for data- and encoder-dependent skew. Our findings indicate trends in final skew values which are dependent on proxies for data complexity in synthetic examples. We demonstrate that optimal divergence skew tends toward a balance of forward and reverse KL as dimensionality increases and, for increasing distribution modes, training skew marginally favours forward KL. In the higher dimensional setting of standard image benchmark datasets, we then establish that final skew values converge across seeds and are consistent for a range of initial values. We further exhibit that learning skew in α-VAEs has a positive impact on test set reconstruction loss (summarised in Figure 1), reconstructing denoised images from noisy images, and explain both advantages from the rate-distortion perspective (Alemi et al., 2018). Overall, we show that α-VAEs with learnable α consistently outperform forward and reverse KL independent of dataset, encoder parameters, and initial skew.

2. α-VAE optimisation

2.1. Learning divergence parameterisations is not constrained optimisation

In order to flexibly interchange divergence terms regularising the latent space of VAEs, it is common to formulate VAE training as a constrained optimisation problem (Higgins et al., 2017). A suitable objective to maximise is the marginal (log-)likelihood of the observed data $x \in \mathbb{R}^m$ as an expectation over the distribution of latent factors $z \in \mathbb{R}^n$

$$\max_\theta \left[\mathbb{E}_{p_\theta(z)}\left[p_\theta(x|z)\right]\right]. \tag{1}$$

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

$$\begin{aligned} &\max_{\phi,\theta} \mathbb{E}_{p_\mathcal{D}(x)}\left[\log \mathbb{E}_{q_\phi(z|x)}\left[p_\theta(x|z)\right]\right] \\ &\text{subject to } \; D(q_\phi(z|x) \parallel p(z)) < \varepsilon, \end{aligned} \tag{2}$$

where ε dictates the strength of the constraint and D is a divergence. Equation (2) is then rewritten as a Lagrangian under the KKT conditions (Karush, 1939; Kuhn & Tucker, 2014), obtaining

$$\begin{aligned} \mathcal{F}(\theta, \phi, \lambda; x, z) = &\ \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] \\ &- \lambda\left(D(q_\phi(z|x) \parallel p(z)) - \varepsilon\right). \end{aligned} \tag{3}$$

However, here we are interested in learning properties of the divergence itself, rendering Equation (2) and (3) invalid. In this setting, the constrained optimisation problem is no longer well-posed, as the constraint becomes part of the optimisation. Instead, we choose to relax these optimisation assumptions while maintaining competing objectives, the log-likelihood and the parameterised divergence

$$\begin{aligned} \mathcal{L}(\theta, \phi, \lambda; x, z) = &\ \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] \\ &- \lambda D_{\psi(x,z)}(q_\phi(z|x) \parallel p(z), \end{aligned} \tag{4}$$

where $\psi(x, z)$ parameterises D.

Such a formulation only relates back to Equation (3) in a valid manner when changes in ψ map to a family of divergences and redundant cases are avoided. For instance, it would be particularly useful if, for a given ψ during optimisation, training could still be understood as a reconstruction loss term and a closed-form divergence regularisation term. It is, therefore, of import to consider: which divergences allow for such a parameterisation, whether they will have non-trivial minima, and whether the subsequent properties of this optimisation generalise or are useful. To this end,

we consider learning α, the *skew*, in the $\mathrm{JS}^{\mathrm{G}_\alpha}$ family of divergences (Appendix A) and proceed by highlighting the connection to forward and reverse KL.

2.2. Skew optimisation

From an information geometric viewpoint, the skew parameter's influence on the intermediate distribution can be seen as the weighting of two distributions along a path in the statistical manifold. In (Masrani et al., 2019), the authors consider the geometric path between the variational distribution $q_\phi(z|x)$ and the model distribution $p(x, z)$. The expensive grid search therein required integration along this path and was later avoided in (Brekelmans et al., 2020) by using a moment-matching approach to learn the optimal point on the path. In contrast to these works, here we consider skew between $q_\phi(z|x)$ and the prior $p(z)$ as a form of regularisation.

Additionally, the property that $\mathrm{JS}^{\mathrm{G}_\alpha}$ interpolates between forward and reverse KL (Deasy et al., 2020) offers a clear mechanism for trading off forward and reverse KL properties. In VAE optimisation, as we have generalised the overall objective by extending the constrained optimisation problem to competing objectives, it is possible to consider direct parameterisation of ψ from Equation (4). As long as ψ maps to $[0, 1]$, the output is a valid divergence (a member of the $\mathrm{JS}^{\mathrm{G}_\alpha}$ family) with extremes at forward and reverse KL, so we can consider optimising ψ via gradient descent.

Before attempting the optimisation, we can derive useful properties of $\mathrm{JS}^{\mathrm{G}_\alpha}(P \parallel Q)$ which clarify how the optimisation should be carried out. In particular, it is important to understand the behaviour of $\mathrm{JS}^{\mathrm{G}_\alpha}$ with respect to α, as this will dictate whether convergence is possible.

Proposition 1. *For Gaussian distributions P and Q with probability density functions $p(z)$ and $q_\phi(z|x)$ respectively, the derivative of $\mathrm{JS}^{\mathrm{G}_\alpha}(P \parallel Q)$ with respect to α is*

$$\frac{d\mathrm{JS}^{\mathrm{G}_\alpha}}{d\alpha} = 2(\alpha - 1)\mathrm{KL}(p(z) \parallel q_\phi(z|x)) + 2\alpha\mathrm{KL}(q_\phi(z|x) \parallel p(z)), \tag{5}$$

with a stationary point at

$$\alpha^* = \frac{\mathrm{KL}(p(z) \parallel q_\phi(z|x))}{\mathrm{KL}(p(z) \parallel q_\phi(z|x)) + \mathrm{KL}(q_\phi(z|x) \parallel p)}, \tag{6}$$

which is a global minimum in the α dimension of the optimisation. Proof in Appendix D.1.

A useful sanity check from Equation (6), is that α^* is clearly bounded between 0 and 1 due to the non-negativity of KL divergence. Secondly, as Equation (6) relies on parameters ϕ, this expression is clearly data dependent, meaning optimal skew shifts during training and should not be fixed as in (Deasy et al., 2020). Finally, considered as a standalone $1D$ optimisation of α, the global minimum means that training divergence skew is convex and should be simple to minimise in practice, with closed-form convergence.

Proposition 2. *The upper bound on the convergence rate of gradient descent to α^* in (6) is*

$$2\left(\mathrm{KL}(p(z) \parallel q_\phi(z|x)) + \mathrm{KL}(q_\phi(z|x) \parallel p(z))\right) D_1^2 e^{-4T}, \tag{7}$$

where $D_1 = ||\alpha_1 - \alpha^*||_2^2$, α_1 is the initial skew value, and T is the number of optimisation steps. Proof in Appendix D.2.

In the next section, we find practical optimisation of skew in $\mathrm{JS}_*^{\mathrm{G}_\alpha}$ to be well behaved, suggesting that training skew in the dual divergence is also well-posed and does not reduce to an invalid divergence or one of the KL directions—despite an equivalent proof of convexity via differentiation not being obvious, due to a mixture term being outside of the log.

3. Experiments

As the convexity shown in Section 2.2 suggests simple optimisation of α, we directly train α as another parameter of the model, but use a separate optimiser to the model parameters so that our 1D analysis holds.

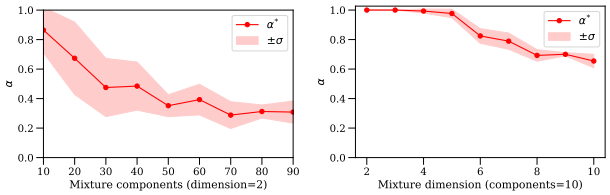
3.1. Characterising skew optimisation for $\mathrm{JS}_*^{\mathrm{G}_\alpha}$

To better understand how dual α-VAEs, α_*-VAEs, will behave in the more complex setting of modern variational inference benchmarks, we first highlight their properties on synthetic examples. In Figure 2, we depict different optimal skew values for a fit of a multivariate Gaussian to an underlying additive mixture of multivariate Gaussians with trained $\mathrm{JS}_*^{\mathrm{G}_\alpha}$ skew. As the divergence integrals are not tractable, we directly optimise the multivariate Gaussian parameters via samples from the data for all divergences.

In both plots of Figure 2, we depict the emergent low-dimensional trends in optimised skew. For Figure 2a, we fit a 2D Gaussian ball to a 2D additive mixture of Gaussian balls, with an increasing number of components in the mixture. Whereas, in Figure 2b, we increase the dimension of the fit and keep the ratio of mixtures to dimensions fixed at 5. As the number of mixture components increases in Figure 2a, optimal skew for $\mathrm{JS}_*^{\mathrm{G}_\alpha}$ decreases, favouring reverse KL as mass becomes more concentrated. Similarly, in Figure 2b, the optimal skew for $\mathrm{JS}_*^{\mathrm{G}_\alpha}$ also decreases, underlining the need for data-dependent skewed divergences and suggesting learnt skew in dual α_*-VAEs will be consistent and avoid trivial cases for more complex data.

VAE	MNIST	Fashion-MNIST	dSprites	Chairs	CelebA
β-VAE (KL($q \parallel p$), $\beta = 4$, (Deasy et al., 2020))	11.75	13.32	10.51	20.79	269.52
InfoVAE (MMD, $\lambda = 500$, (Deasy et al., 2020))	13.19	11.10	11.87	18.85	271.71
Vanilla VAE (KL($q \parallel p$))	10.67 ± 0.27	12.36 ± 0.22	7.78 ± 0.24	20.33 ± 0.34	262.53 ± 2.07
α-VAE (fixed $\alpha = 0.5$)	11.24 ± 0.05	11.07 ± 0.67	12.07 ± 0.06	19.11 ± 0.14	270.33 ± 0.78
α_*-VAE (fixed $\alpha_* = 0.5$)	8.82 ± 0.04	9.80 ± 0.06	5.72 ± 0.07	16.40 ± 0.15	264.27 ± 0.45
α-VAE	8.89 ± 0.07	9.90 ± 0.05	5.03 ± 0.31	16.48 ± 0.08	$\mathbf{259.50 \pm 0.32}$
α_*-VAE	$\mathbf{8.52 \pm 0.07}$	$\mathbf{9.59 \pm 0.03}$	$\mathbf{3.88 \pm 0.27}$	$\mathbf{15.98 \pm 0.17}$	$\mathbf{259.52 \pm 0.36}$

Table 1. Final model reconstruction error across regularisation divergences and datasets. For trainable skew α-VAEs (bottom two rows), final α values are given in Table 2. α_* indicates dual α-VAEs.



(a) Learnt α vs. mixture components.

(b) Learnt α vs. mixture dimension.

Figure 2. Emergent trends in optimised skew when fitting a Gaussian ball to an additive mixture of Gaussians in ND for increasing data complexity.

3.2. Benchmark image dataset performance

Divergence skew convergence. As the expression for the minimum in Equation (6) is encoder dependent, we begin our benchmark dataset assessment by depicting the alpha landscape. In Figure 3, we see that optimal α decreases during training, before stabilising in the final epochs. This supports the argument that naively fixing skew is not optimal, leads to inferior solutions, and even a hyperparameter search for fixed skew followed by training is not sufficient.

The divergences converge to a consistent minimum. This confirms that our $1D$ assessment of convexity and separate optimisation of α successfully leads to stable skew values (see Figure 4 in the Appendix). In addition, we exemplify how skew evolves across 5 different training seeds. These plots delineate data and encoder dependency, with consistency across seeds which initialise model parameters and training procedures (e.g. dropout and batching), suggesting learnt skew predominantly derives from, and varies between, datasets (see Table 2).

Robustness to noise. As learning skew allows for problem-dependent measurement of the distance to the prior, which better accommodates more dispersed encoded distributions, we tested how learning skew regularises VAEs in various noise settings. In Figure 5, we present denoising experiments where we add Gaussian noise, $\mathcal{N}(0, \sigma^2)$, to the normalised input images and clip to $[0, 1]$. Despite degraded intuition surrounding where the skew balance should lie in the noisy setting, Figure 5 clearly demonstrates that our trainable skew in $\mathrm{JS}^{\mathrm{G}_\alpha}$, and its dual form, consistently provides lower test set reconstruction loss across noise levels. As a sanity check, we can also verify expected trends in the other divergences, with the less dispersion-friendly reverse KL performing poorly at higher noise levels as forward KL becomes more appropriate, as well as the expected robust behaviour of MMD at high noise levels (Zhao et al., 2019).

Improved reconstruction as a rate-distortion trade-off. We test our model's reconstruction-loss performance in the standard clean-image setting. Table 1 demonstrates superior reconstruction loss across multiple datasets, outperforming both the naive choice of fixed $\alpha = 0.5$ and KL divergence—the latter by a substantial margin. We further detail the relationship with fixed α in Figure 1a, Figure 6, the supplementary figures in Appendix E, observing low reconstruction loss from an arbitrary starting skew. To explain this performance gain, we plot distortion (MSE) against rate (Figure 7 in Appendix E), measured using the reverse KL divergence for all divergences. The shift down and to the right for α-VAEs mean that they trade off rate for distortion, even more so when skew is learnt, giving higher quality image reconstruction when starting with unknown divergence scaling λ.

4. Conclusion

We recast VAE optimisation as a multi-objective task with competing reconstruction and divergence regularisation terms. This allowed us to use gradient descent to directly learn the divergence skew used to control the variational distribution $q_\phi(z|x)$ relative to the prior distribution $p(z)$. The resulting method, α-VAE, was shown to be well-posed as a 1D optimisation with a dependency on both the encoder and the data, has improved reconstruction and denoising properties over standard regularisation techniques, and avoids an expensive grid search over skew values. Moreover, as α-VAEs also generalise over the β-VAE family, they are a practical, efficient, and unbiased choice of VAE across a range of tasks and domains.

References

Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. Fixing a broken elbo. In *International Conference on Machine Learning*, pp. 159–168. PMLR, 2018.

Aubry, M., Maturana, D., Efros, A., Russell, B., and Sivic, J. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014.

Bishop, C. M. *Pattern recognition and machine learning*. Springer, 2006.

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

Bozkurt, A., Esmaeili, B., Tristan, J., Brooks, D. H., Dy, J. G., and van de Meent, J. Rate-regularization and generalization in variational autoencoders. In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3880–3888. PMLR, 2021.

Brekelmans, R., Masrani, V., Wood, F., Steeg, G. V., and Galstyan, A. All in the exponential family: Bregman duality in thermodynamic variational inference. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1111–1122. PMLR, 2020.

Bubeck, S. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Child, R. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020.

Deasy, J., Simidjievski, N., and Lió, P. Constraining variational inference with geometric jensen-shannon divergence. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems*, pp. 2732–2741, 2017.

Germain, M., Gregor, K., Murray, I., and Larochelle, H. Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pp. 881–889, 2015.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Hensman, J., Zwießele, M., and Lawrence, N. Tilted variational bayes. In *Artificial Intelligence and Statistics*, pp. 356–364, 2014.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. *Iclr*, 2(5):6, 2017.

Huang, C.-W., Tan, S., Lacoste, A., and Courville, A. C. Improving explorability in variational inference with annealed variational objectives. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 9701–9711. Curran Associates, Inc., 2018.

Huang, S., Makhzani, A., Cao, Y., and Grosse, R. B. Evaluating lossy compression rates of deep generative models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4444–4454. PMLR, 2020.

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

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *ICLR 2014*, 2014. URL `abs/1312.6114`.

Kuhn, H. W. and Tucker, A. W. Nonlinear programming. In *Traces and emergence of nonlinear programming*, pp. 247–258. Springer, 2014.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951.

Larochelle, H. and Murray, I. The neural autoregressive distribution estimator. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 29–37, 2011.

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

Li, Y. and Turner, R. E. Rényi divergence variational inference. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 1073–1081. Curran Associates, Inc., 2016.

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

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Lucas, J., Tucker, G., Grosse, R. B., and Norouzi, M. Understanding posterior collapse in generative latent variable models. In *DGS@ICLR*, 2019.

Masrani, V., Le, T. A., and Wood, F. The thermodynamic variational objective. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 11521–11530, 2019.

Matthey, L., Higgins, I., Hassabis, D., and Lerchner, A. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Neal, R. M. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Niculescu, C. and Persson, L.-E. *Convex functions and their applications*. Springer, 2006.

Nielsen, F. On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means. *Entropy*, 21(5):485, 2019. ISSN 1099-4300. doi: 10.3390/e21050485. Extended version available at https://arxiv.org/abs/1904.04017.

Nielsen, F. On a variational definition for the jensen-shannon symmetrization of distances based on the information radius. *Entropy*, 23(4), 2021. ISSN 1099-4300.

Nielsen, F. and Garcia, V. Statistical exponential families: A digest with flash cards. *arXiv preprint arXiv:0911.4863*, 2009.

Nishiyama, T. Generalized bregman and jensen divergences which include some f-divergences. *arXiv preprint arXiv:1808.06148*, 2018.

Rezende, D. J., Mohamed, S., and Wierstra, D. 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*, pp. 1278–1286. PMLR, 2014.

Rudin, W. et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

Vahdat, A. and Kautz, J. Nvae: A deep hierarchical variational autoencoder. *arXiv preprint arXiv:2007.03898*, 2020.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Zhang, M., Bird, T., Habib, R., Xu, T., and Barber, D. Variational f-divergence minimization. *arXiv preprint arXiv:1907.11891*, 2019.

Zhao, S., Song, J., and Ermon, S. Infovae: Balancing learning and inference in variational autoencoders. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA*, pp. 5885–5892. AAAI Press, Palo Alto, CA, USA, 2019.

A. The $\mathrm{JS}^{\mathrm{G}_\alpha}$ divergence family

For distributions P and Q of a continuous random variable $Z = [Z_1, \ldots, Z_n]^{\mathrm{T}}$, the forward KL divergence (Kullback & Leibler, 1951) is defined as

$$\mathrm{KL}(P \parallel Q) = \int_Z p(z) \log \left[\frac{p(z)}{q(z)}\right] dz, \tag{8}$$

where p and q are the probability densities of P and Q respectively, $z \in \mathbb{R}^n$, and reverse KL divergence refers to $\mathrm{KL}(Q \parallel P)$.

Reverse KL from a standard normal distribution $\mathcal{N}_2(0, I)$ to a diagonal multivariate normal distribution $\mathcal{N}_1(\mu, \Sigma)$, $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$, is used throughout variational models (Higgins et al., 2017; Kingma & Welling, 2014; Neal, 2012) and is known to enforce zero-avoiding hyperparameters on $\mathcal{N}_1$ when minimised (Bishop, 2006; Murphy, 2012). On the other hand, the forward KL divergence is known for its zero-forcing property (Bishop, 2006; Murphy, 2012). However, there exist well-known drawbacks of the KL divergence, such as no upper bound leading to unstable optimisation and poor approximation (Hensman et al., 2014), as well as its asymmetric property $\mathrm{KL}(P \parallel Q) \neq \mathrm{KL}(Q \parallel P)$. Under-dispersed approximations relative to the exact posterior also produce difficulties with light-tailed posteriors when the variational distribution has heavier tails (Dieng et al., 2017).

One attempt at remedying these issues is the well-known symmetrisation, the Jensen-Shannon (JS) divergence (Lin, 1991)

$$\mathrm{JS}(p(z) \parallel q(z)) = \frac{1}{2}\mathrm{KL}\left(p \,\middle\|\, \frac{p+q}{2}\right) + \frac{1}{2}\mathrm{KL}\left(q \,\middle\|\, \frac{p+q}{2}\right). \tag{9}$$

Although the JS divergence is bounded 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 (9).

Recently, (Nielsen, 2019) and (Nishiyama, 2018) have proposed a further generalisation of the JS divergence using abstract means (*quasi-arithmetic means* (Niculescu & Persson, 2006), also known as *Kolmogorov-Nagumo means*). By choosing the *weighted geometric mean* $\mathrm{G}_\alpha(x, y) = x^\alpha y^{1-\alpha}$ 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 (Nielsen & Garcia, 2009), we have the divergence

$$\mathrm{JS}^{\mathrm{G}_\alpha}(p(z) \parallel q(z)) = (1-\alpha)\mathrm{KL}\left(p \parallel G_\alpha(p, q)\right) \tag{10}$$
$$+ \alpha\mathrm{KL}\left(q \parallel G_\alpha(p, q)\right). \tag{11}$$

$\mathrm{JS}^{\mathrm{G}_\alpha}$, the *skew-geometric Jensen-Shannon divergence*, between two multivariate Gaussians $\mathcal{N}_1(\mu_1, \Sigma_1)$ and $\mathcal{N}_2(\mu_2, \Sigma_2)$ then admits a closed form

$$\mathrm{JS}^{\mathrm{G}_\alpha}\left(\mathcal{N}_1 \parallel \mathcal{N}_2\right) = (1-\alpha)\mathrm{KL}\left(\mathcal{N}_1 \parallel \mathcal{N}_\alpha\right) \tag{12}$$
$$+ \alpha\mathrm{KL}\left(\mathcal{N}_2 \parallel \mathcal{N}_\alpha\right) \tag{13}$$

with the equivalent dual divergence being

$$\mathrm{JS}_*^{\mathrm{G}_\alpha}\left(\mathcal{N}_1 \parallel \mathcal{N}_2\right) = (1-\alpha)\mathrm{KL}\left(\mathcal{N}_\alpha \parallel \mathcal{N}_1\right) \tag{14}$$
$$+ \alpha\mathrm{KL}\left(\mathcal{N}_\alpha \parallel \mathcal{N}_2\right) \tag{15}$$

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

$$\Sigma_\alpha = \left((1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1}\right)^{-1} \tag{16}$$
$$\mu_\alpha = \Sigma_\alpha\left((1-\alpha)\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2\right). \tag{17}$$

In simple terms, these divergences measure a weighted *arithmetic mean of divergences* from/to the prior/variational distribution to/from a weighted *geometric mean of distributions*. We proceed by examining why this skew parameter α should be learnt (replacing ψ in Equation (4)), what this means in divergence space, and where the optimal α lie.

B. Related work

Since its introduction (Nielsen, 2021), $\mathrm{JS}^{\mathrm{G}_\alpha}$ (with $\alpha = 0.5$) has been used to decompose and estimate multi-modal ELBO loss as regularisation in multi-modal VAEs. (Deasy et al., 2020) further investigated the potential of $\mathrm{JS}^{\mathrm{G}_\alpha}$ at different skew values, in turn, demonstrating improved reconstruction performance of VAEs. In contrast, in this paper we propose data-dependent learnable skew, which retains the benefits of using an optimal skew but avoids an expensive grid-search. This provides both efficient and practical way of training VAEs, leading to improved denoising and reconstruction performance as well as better lossy-compression rates focussed on in (Huang et al., 2020). In a similar context, our work is related to (Brekelmans et al., 2020), that introduces an optimal integration schedule via dynamic parameter selection when approximating the Thermodynamic Variational Objective (Masrani et al., 2019).

α-VAEs extend the standard VAEs (Kingma & Welling, 2014; Rezende et al., 2014) paradigm with regularisation constraint inspired by recent work on closed-form expressions for statistical divergences (Nielsen, 2019; Nishiyama, 2018). In particular, α-VAEs offer a stable and intuitive regularisation mechanism. This allows optimal interpolation between forward and reverse KL divergence, therefore combating the issue of posterior collapse (Lucas et al., 2019). In this regard, our work is related to approaches that address this issue through KL annealing during training (Bozkurt et al., 2021; Huang et al., 2018; Burgess et al., 2018). 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 (Hensman et al., 2014; Tolstikhin et al., 2018; Dieng et al., 2017; Zhang et al., 2019; Zhao et al., 2019; Li & Turner, 2016).

C. Benchmark datasets for skew exploration

Throughout our experiments we evaluate the reconstruction loss (mean squared error) on four standard benchmark image datasets : MNIST, 28×28 black and white images of handwritten digits (LeCun et al., 2010); Fashion-MNIST, 28×28 black and white images of clothing (Xiao et al., 2017); Chairs, 64×64 black and white images of 3D chairs (Aubry et al., 2014); dSprites 64×64 black and white images of 2D shapes procedurally generated from 6 ground truth independent latent factors (Matthey et al., 2017); CelebA resampled to $64 \times 64 \times 3$ colour images of celebrity faces (Liu et al., 2015). For fair comparison, we follow Higgins et al. (2017) by selecting a common neural architecture across experiments (full details in Appendix **??**). For consistent analysis, rather than searching within architecture and hyperparameter spaces for the best performing model by some metric, we standardise comparison and characterise the benefit of learning divergence skew.

D. Proofs

D.1. Proposition 1.

Proof. First, simplifying the divergence

$$\begin{aligned}
\mathrm{JS}^{G_\alpha} &= (1-\alpha)\mathrm{KL}\left(p \parallel p^\alpha q^{1-\alpha}\right) + \alpha \mathrm{KL}\left(q \parallel p^\alpha q^{1-\alpha}\right) \\
&= (1-\alpha)\int_x p \log\left[\frac{p}{p^\alpha q^{1-\alpha}}\right] dx \\
&+ \alpha \int_x q \log\left[\frac{q}{p^\alpha q^{1-\alpha}}\right] dx \\
&= (1-\alpha)^2 \int_x p \log\left[\frac{p}{q}\right] dx + \alpha^2 \int_x q \log\left[\frac{q}{p}\right] dx \\
&= (1-\alpha)^2 \mathrm{KL}(p \parallel q) + \alpha^2 \mathrm{KL}(q \parallel p). \qquad (18)
\end{aligned}$$

Then, differentiating (18) with respect to α

$$\frac{d\mathrm{JS}^{G_\alpha}}{d\alpha} = 2((\alpha-1)\mathrm{KL}(p \parallel q) + \alpha \mathrm{KL}(q \parallel p)) = 0, \qquad (19)$$

and rearranging gives Equation (6), before differentiating again

$$\frac{d^2\mathrm{JS}^{G_\alpha}}{d\alpha^2} = 2(\mathrm{KL}(p \parallel q) + \mathrm{KL}(q \parallel p)) \geq 0, \qquad (20)$$

demonstrates the global minimum property as KL divergence is always positive. □

D.2. Proposition 2.

Proof. We first define the convexity strength and smoothness of $f(\alpha) = \mathrm{JS}^{G_\alpha}$, using its simplified form in Equation (18).

A function $f(\alpha) : \mathbb{R}^n \mapsto \mathbb{R}^n$ is λ-strongly convex if

$$\nabla^2 f(\alpha) \succeq \lambda I, \qquad (21)$$

where $\succeq$ is a generalised inequality and, in our 1D case, this expression reduces to the second derivative with respect to α, giving

$$\lambda = 2(\mathrm{KL}(p \parallel q) + \mathrm{KL}(q \parallel p)). \qquad (22)$$

$f(\alpha)$ is also β-smooth if $\frac{df}{d\alpha}$ is β-Lipschitz

$$||\nabla f(x) - \nabla f(y)|| \leq \beta ||x - y||, \qquad (23)$$

and, as our f is doubly differentiable, the Mean Value Theorem (Rudin et al., 1976) with $g = \frac{df}{d\alpha}$

$$\frac{g(x) - g(y)}{x - y} \leq \frac{dg(z)}{d\alpha} = \frac{d^2 f(z)}{d\alpha^2} \quad \forall x < z < y, \qquad (24)$$

gives the necessary bound and the same constant

$$\beta = 2(\mathrm{KL}(p \parallel q) + \mathrm{KL}(q \parallel p)). \qquad (25)$$

We can now substitute into the result from convex optimisation with gradient descent (Bubeck, 2014) that, for a λ-strongly convex and β-smooth function, the optimal step size γ is

$$\begin{aligned}
\gamma &= \frac{2}{\lambda + \beta} \qquad (26) \\
&= \frac{1}{2(\mathrm{KL}(p \parallel q) + \mathrm{KL}(q \parallel p))}, \qquad (27)
\end{aligned}$$

and the upper bound on the convergence rate is

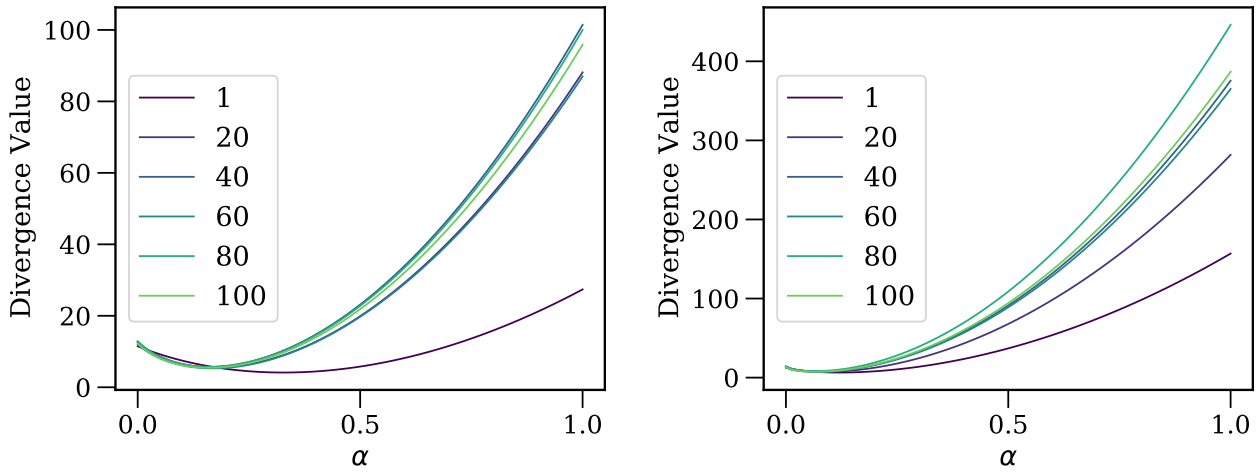
$$\beta ||\alpha_1 - \alpha^*||_2^2 e^{-\frac{4T}{\kappa}}, \qquad (28)$$

where $\kappa = \frac{\beta}{\lambda} = 1$ is the condition number, completing the proof. □

E. Further results

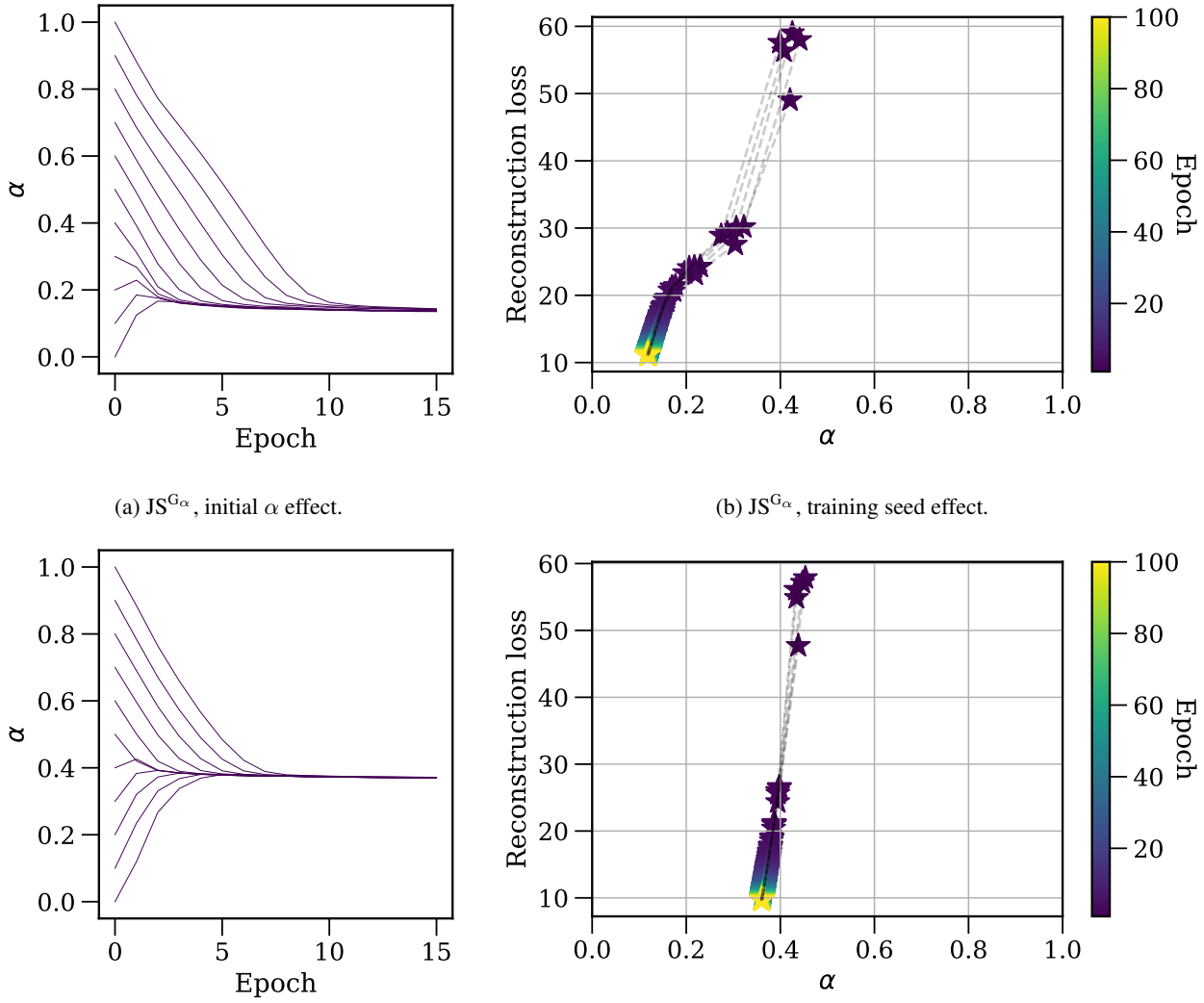
Divergence	MNIST	FashionMNIST	dSprites	Chairs	CelebA
t-JS$^{G_\alpha}$	0.118 ± 0.001	0.165 ± 0.002	0.06 ± 0.007	0.121 ± 0.001	0.0377 ± 0.000
t-JS$_*^{G_\alpha}$	0.360 ± 0.000	0.377 ± 0.002	0.39 ± 0.001	0.365 ± 0.001	0.310 ± 0.000

Table 2. Mean learnt α for JS$^{G_\alpha}$ and JS$_*^{G_\alpha}$ with standard deviation across 5 different training seeds.



(a) JS$^{G_\alpha}$ (b) JS$_*^{G_\alpha}$

Figure 3. α landscape at different training epochs in training for FashionMNIST.



(a) $\mathrm{JS}^{\mathrm{G}_\alpha}$, initial α effect.

(b) $\mathrm{JS}^{\mathrm{G}_\alpha}$, training seed effect.

(c) $\mathrm{JS}_*^{\mathrm{G}_\alpha}$, initial α effect.

(d) $\mathrm{JS}_*^{\mathrm{G}_\alpha}$, training seed effect.

Figure 4. The robust nature of our method's convergence. α convergence across a range of starting values and seeds using the MNIST dataset.

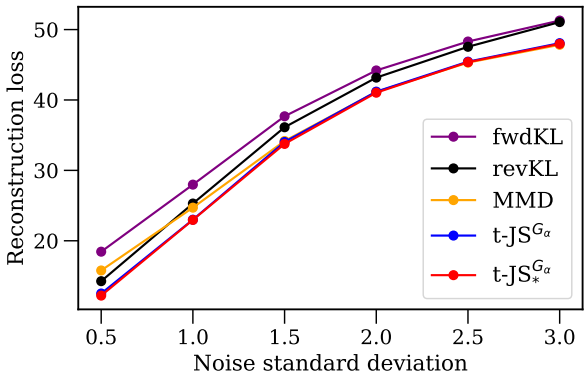


Figure 5. Reconstruction loss for noisy input images across different noise levels and regularisation divergences on MNIST.

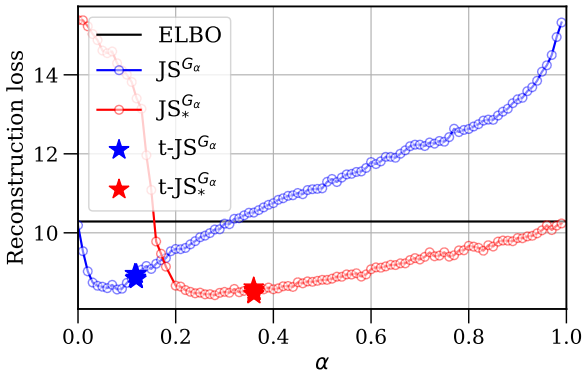
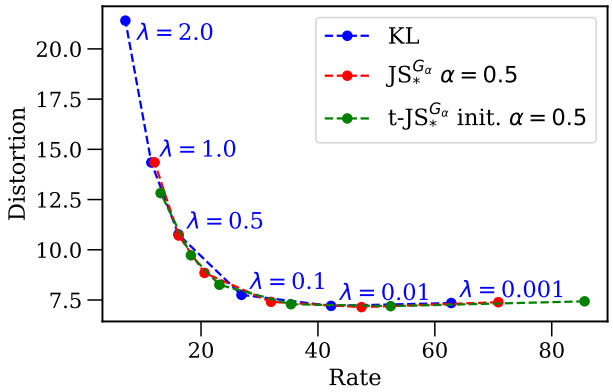
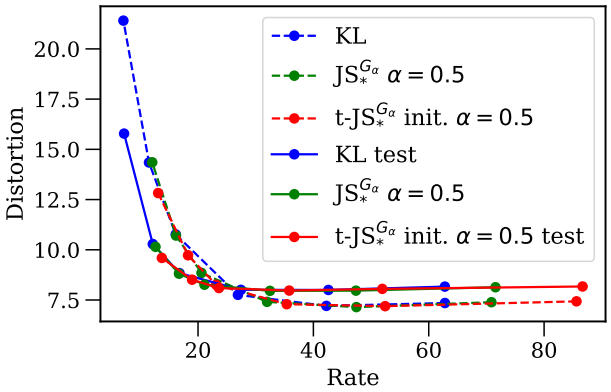


Figure 6. α-VAEs on the Fashion-MNIST dataset.



(a) Train set RD curves for varying λ.



(b) RD curve generalisation gap.

Figure 7. Rate distortion curves for KL, JS$_*^{G_\alpha}$, and JS$_*^{G_\alpha}$ with learnt α on MNIST. Dashed or full lines connect values from the training or test set respectively. λ values vary consistently for each divergence from left to right but are only annotated for KL.