Disentangling Disentanglement

Emile Mathieu* **Tom Rainforth*** **N. Siddharth*** **Yee Whye Teh**
University of Oxford
{emile.mathieu, rainforth, y.w.teh}@stats.ox.ac.uk nsid@robots.ox.ac.uk

Abstract

We develop a generalised notion of disentanglement in variational auto-encoders (VAEs) by casting it as a *decomposition* of the latent representation, characterised by i) enforcing an appropriate level of overlap in the latent encodings of the data, and ii) regularisation of the average encoding to a desired structure, represented through the prior. We motivate this by showing that a) the β-VAE disentangles purely through regularisation of the overlap in latent encodings, and b) disentanglement, as independence between latents, can be cast as a regularisation of the aggregate posterior to a prior with specific characteristics. We validate this characterisation by showing that simple manipulations of these factors, such as using rotationally variant priors, can help improve disentanglement, and discuss how this characterisation provides a more general framework to incorporate notions of decomposition beyond just independence between the latents.

1 Introduction

An oft-stated motivation for learning disentangled representations of data with deep generative models is a desire to achieve interpretability [4, 8]—particularly the *decomposability* [see §3.2.1 in 19] of latent representations to admit intuitive explanations. Most work on disentanglement has constrained the form of this decomposition to capturing purely *independent* factors of variation [1, 3, 6–8, 10–12, 16, 28, 29], typically evaluating this using purpose-built, artificial, data [7, 10, 12, 16], whose generative factors are themselves independent by construction. However, the high-level motivation for achieving decomposability places no a priori constraints on the form of the decompositions—just that they are captured effectively.

The conventional view of disentanglement, as recovering independence, has subsequently motivated the development of formal evaluation metrics for independence [10, 16], which in turn has driven the development of objectives that target these metrics, often by employing regularisers explicitly encouraging independence in the representations [10, 11, 16].

We argue that this methodological approach is not generalisable, and potentially even harmful, to learning decomposable representations for more complicated problems, wherein such simplistic representations will be unable to accurately mimic the generation of high dimensional data from low dimensional latent spaces. To see this, consider a typical measure of disentanglement-as-independence [e.g. 10], computed as the extent to which a latent dimension $d \in D$ predicts a generative factor $k \in K$ with each latent capturing at most one generative factor. This implicitly assumes $D \geq K$, as otherwise the latents are not able to explain all of the generative factors. However, for real data, the association is more likely $D \ll K$, with one learning a low-dimensional abstraction of a complex process involving many factors. Such complexities necessitate richly structured dependencies between latent dimensions—as reflected in the motivation for a handful of approaches [5, 11, 15, 25] that explore this through graphical models, although employing mutually-inconsistent, and not generalisable, interpretations of disentanglement.

Here, we develop a generalisation of disentanglement—*decomposing* latent representations—that can help avoid such pitfalls. Note that the typical assumption of independence *implicitly* makes a

*Equal Contribution

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

choice of decomposition—that the latent features are independent of one another. We make this *explicit*, and exploit it to provide improvement to disentanglement simply through judicious choices of structure in the prior, while also introducing a framework flexible enough to capture alternate, more complex, notions of decomposition such as sparsity [26], hierarchical structuring, or independent *subspaces*.

2 Decomposition: A Generalisation of Disentanglement

We characterise the decomposition of latent spaces in VAEs to be the fulfilment of two factors:

a. An "appropriate" level of overlap in the latent space—ensuring that the range of latent values capable of encoding a particular datapoint is neither too small, nor too large. This is, in general, dictated by the level of stochasticity in the encoder: the higher the encoder variance, the higher the number of datapoints which can plausibly give rise to a particular encoding.
b. The marginal posterior $q_\phi(\boldsymbol{z}) \triangleq \mathbb{E}_{p_\mathcal{D}(\mathbf{x})}[q_\phi(\boldsymbol{z} \mid \boldsymbol{x})]$ (for encoder $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ and true data distribution $p_\mathcal{D}(\mathbf{x})$) matching the prior $p_\theta(\boldsymbol{z})$, where the latter expresses the desired dependency structure between latents.

The overlap factor (a) is perhaps best understood by considering the extremes—too little, and the latent encodings effectively become a lookup table; too much, and the data and latents don't convey information about each other. In both cases, the meaningfulness of the latent encodings is lost. Thus, without the *appropriate* level of overlap—dictated both by noise in the true generative process and dataset size—it is not possible to enforce meaningful structure on the latent space.

The regularisation factor (b) enforces a congruence between the (aggregate) latent embeddings of data and the dependency structures expressed in the prior. We posit that such structure is best expressed in the prior, as opposed to explicit independence regularisation of the marginal posterior [7, 16], to enable the *generative* model to express the captured decomposition; and to avoid potentially violating the self-consistency between encoder, decoder, and true data generating distribution. Furthermore, the prior provides a rich and flexible means of expressing desired structure, by defining a generative process that encapsulates dependencies between variables, analogously to a graphical model.

Critically, *neither factor is sufficient in isolation*. An inappropriate level of overlap in the latent space (a) will impede interpretability, irrespective of how well the regularisation (b) goes, as the latent space need not be meaningful. On the other hand, without the pressure to regularise (b) to the prior, the latent space is under no constraint to exhibit the desired structure.

Deconstructing the β-VAE: To show how existing approaches fit into our proposed framework, we now consider, as a case study, the β-VAE [12]—an adaptation of the VAE objective (ELBO) to learn better-disentangled representations. We introduce new theoretical results that show its empirical successes are purely down to controlling the level of overlap, i.e. factor (a). In particular, we have the following result, the proof of which is given in Appendix A, along with additional results.

Theorem 1. *The β-VAE target*

$$\mathcal{L}_\beta(\boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \beta \,\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z})) \tag{1}$$

can be interpreted in terms of the standard ELBO, $\mathcal{L}(\boldsymbol{x})\,(\pi_{\theta,\beta}, q_\phi)$, for an adjusted target $\pi_{\theta,\beta}(\boldsymbol{x}, \boldsymbol{z}) \triangleq p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) f_\beta(\boldsymbol{z})$ with annealed prior $f_\beta(\boldsymbol{z}) \triangleq p_\theta(\boldsymbol{z})^\beta / F_\beta$ as

$$\mathcal{L}_\beta(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x})\,(\pi_{\theta,\beta}, q_\phi) + (\beta - 1) H_{q_\phi} + \log F_\beta \tag{2}$$

where $F_\beta \triangleq \int_{\boldsymbol{z}} p_\theta(\boldsymbol{z})^\beta d\boldsymbol{z}$ is constant given β, and H_{q_ϕ} is the entropy of $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$.

Clearly, the second term in (2), enforcing a *maxent* regulariser on the posterior $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$, allows the value of β to affect the overlap of encodings in the latent space; for Gaussian priors this effect is exactly equivalent to regularising the encoder to have higher variance. The annealed prior's effect though, is more subtle. While one could interpret its effect as simply inducing a fixed scaling on the parameters (c.f. Appendix A.1), which could be ignored and 'fixed' during learning, it actually has the effect of *exactly* counteracting the latent-space scaling due to the entropy regularisation—ensuring that the scaling of the marginal posterior matches that of the prior.

Taken together, these insights demonstrate that the β-VAE's disentanglement is purely down to controlling the level of induced overlap: it places no additional direct pressure on the latents to be

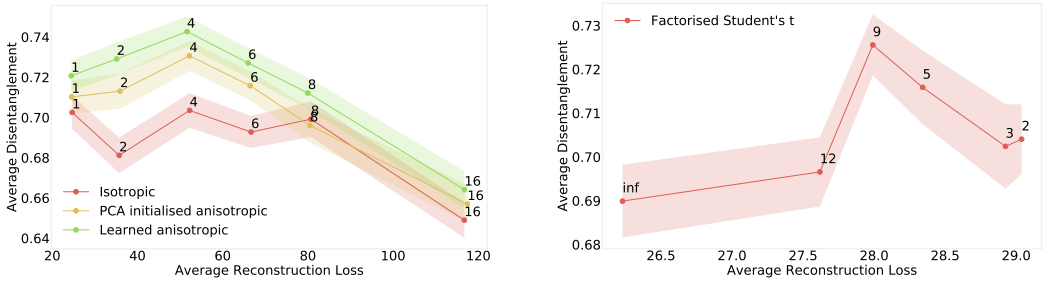

Figure 1: Reconstruction loss vs disentanglement metric [16] for β-VAE (i.e. (3) with $\alpha = 0$) trained on the *2D Shapes* dataset [21]. Shaded areas represent 95% confidence intervals for disentanglement metric estimate, calculated using 100 separately trained networks. See Appendix B for details. [Left] Using an anisotropic Gaussian with diagonal covariance either fixed to the principal component values or learned during training. Point labels represent different values of β. [Right] Using $p_\theta^\nu(\boldsymbol{z}) = \prod_i \text{STUDENT-T}(z_i; \nu)$ for different degrees of freedom ν with $\beta = 1$. Note that $p_\theta^\nu(\boldsymbol{z}) \to \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \mathbf{I})$ as $\nu \to \infty$, and reducing ν only incurs a minor increase in reconstruction loss.

independent, it only helps avoid the pitfalls of inappropriate overlap. Amongst other things, this explains why larger values of β are not universally beneficial for disentanglement, as the level of overlap can be increased too far. It also dispels the conjecture [6, 12] that the β-VAE encourages the latent variables to take on meaningful representations when using the standard choice of an isotropic Gaussian prior: for this prior, each term in (2) is invariant to rotation of the latent space. Our results show that the β-VAE encourages the latent states to match true generative factors no more than it encourages them to match rotations of the true generative factors, with the latter capable of exhibiting strong correlations between the latents. This view is further supported by our empirical results (see Figure 1), calculated by averaging over a large number of independently trained networks, where we did not observe any gains in disentanglement (using the metric from Kim and Mnih [16]) from increasing $\beta > 1$ with an isotropic Gaussian prior trained on the *2D Shapes* dataset [21].

A new objective: Given the characterisation set out above, we now develop an objective that incorporates the effect of both factors (a) and (b). From our analysis of the β-VAE, we see that its objective (1) allows expressing overlap, i.e. factor (a). To additionally capture the regularisation (b) between the marginal posterior and the prior, we add a divergence term $\mathbb{D}(q_\phi(z), p(\boldsymbol{z}))$, yielding

$$\mathcal{L}_{\alpha,\beta}(\boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] - \beta \ \text{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, p(\boldsymbol{z})) - \alpha \, \mathbb{D}(q_\phi(\boldsymbol{z}), p(\boldsymbol{z})) \tag{3}$$

where we can now control the extent to which factors (a) and (b) are enforced, through appropriate setting of β and α respectively.

Note that such an additional term has been previously considered by Kumar et al. [18], with $\mathbb{D}(q_\phi(\boldsymbol{z}), p(\boldsymbol{z})) = \text{KL}(q_\phi(\boldsymbol{z}) \,\|\, p(\boldsymbol{z}))$, although for the sake of tractability they rely instead on moment matching using covariances. There have also been a number of approaches that decompose the standard VAE objective in different ways [e.g. 2, 11, 13] to expose $\text{KL}(p(\boldsymbol{z}) \,\|\, q_\phi(\boldsymbol{z}))$ as a component, but, as we discuss in Appendix C, this is difficult to compute correctly in practice, with common previous approaches leading to highly biased estimates whose practical behaviour is very different than the divergence they are estimating. Wasserstein Auto-Encoders [27] formulate an objective that includes a general divergence term between the prior and marginal posterior, which are instantiated using either maximum mean discrepancy (MMD) or a variational formulation of the Jensen-Shannon divergence (a.k.a GAN loss). However, we find that empirically, choosing the MMD's kernel and numerically stabilising its U-statistics estimator to be tricky, and designing and learning a GAN to be cumbersome and unstable. Consequently, the problems of choosing an appropriate $\mathbb{D}(q_\phi(\boldsymbol{z}), p(\boldsymbol{z}))$ and generating reliable estimates for this choice are tightly coupled, with a general purpose solution remaining an important open problem in the field; further discussion is given in Appendix C.

3 Experiments

Prior for axis-aligned disentanglement First, we show how subtle changes to the prior distribution can yield improvement in terms of a common notion of disentanglement [see §4 in 16]. The

most common choice of prior, an isotropic Gaussian, $p_\theta(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \mathbf{I})$, has previously been justified by the correct assertion that the latents are independent under the prior [12]. However, an isotropic Gaussian is also *rotationally invariant* and so does not constrain the axes of the latents space to capture any meaning. Figure 1 demonstrates that substantial improvements in disentanglement can be achieved by simply using either a non-isotropic Gaussian or using a product of Student-t's prior, both of which break the rotational invariance.

Clustered prior We next consider an alternative decomposition one might wish to impose, namely a clustering of the latent space. For this, we use the "pinwheels" dataset from [15] and use a mixture of four equally-weighted Gaussians as our prior. We then conduct an ablation study to observe the effect of varying α and β in $\mathcal{L}_{\alpha,\beta}(\mathbf{x})$ (as per (3)) on the learned representations, taking the divergence to be $\mathrm{KL}\left(p(\boldsymbol{z})||q_\phi(\boldsymbol{z})\right)$ (see Appendix B for details). As shown in Figure 2, our framework allows one to impose this alternate decomposition, allowing control of both the level of overlap and the form of the marginal posterior.

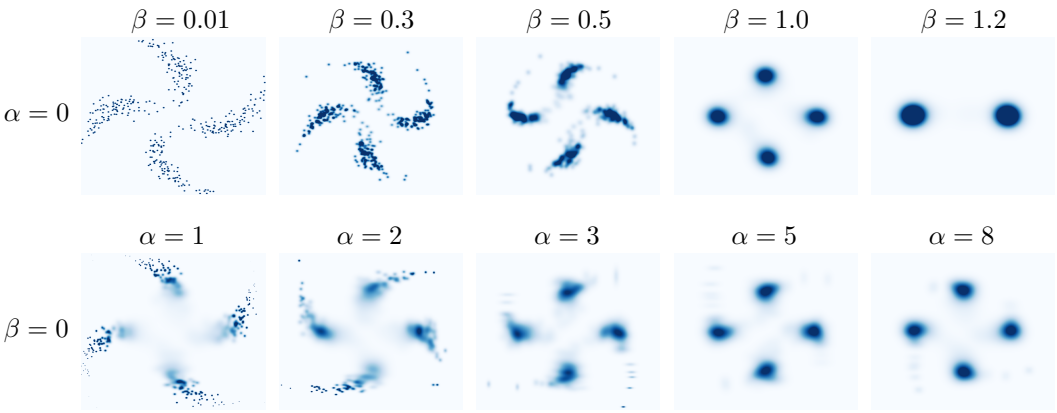

Figure 2: Density of aggregate posterior $q_\phi(\boldsymbol{z})$ for different values of α and β. [Top] $\alpha = 0$, $\beta \in \{0.01, 0.3, 0.5, 1.0, 1.2\}$. [Bottom] $\beta = 0$, $\alpha \in \{1, 2, 3, 5, 8\}$. We see that increasing β increases the level of overlap in $q_\phi(\boldsymbol{z})$, as a consequence of increasing the encoder variance for individual datapoints. When β is too large, the encoding of a datapoint loses meaning. Also, as a single datapoint encodes to a Gaussian distribution, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is unable to match $p_\theta(\boldsymbol{z})$ exactly. Because $q_\phi(\boldsymbol{z}|\boldsymbol{x}) \to q_\phi(\boldsymbol{z})$ when $\beta \to \infty$, this in turn means that overly large values of β actually cause a mismatch between $q_\phi(\boldsymbol{z})$ and $p_\theta(\boldsymbol{z})$ (see top right). Increasing α, instead always improves the match between $q_\phi(\boldsymbol{z})$ and $p_\theta(\boldsymbol{z})$. Here, the finiteness of the dataset and the choice of divergence results in an increase in the overlap with increasing α, but only up to the level required for a non-negligible overlap between the nearby datapoints, such that large values of α do not cause the encodings to lose significance.

Acknowledgements

EM, TR, YWT were supported in part by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013) / ERC grant agreement no. 617071. EM was also supported by Microsoft Research through its PhD Scholarship Programme. NS was funded by EPSRC/MURI grant EP/N019474/1.

References

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

[2] Anonymous. Explicit information placement on latent variables using auxiliary generative modelling task. In *Submitted to International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1l-SjA5t7. under review.

[3] Abdul Fatir Ansari and Harold Soh. Hyperprior Induced Unsupervised Disentanglement of Latent Representations. *arXiv.org*, September 2018.

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.50. URL http://dx.doi.org/10.1109/TPAMI.2013.50.

[5] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, 2018.

[6] Christopher P. Burgess, Irina Higgins, Arka Pal, Loïc Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *CoRR*, abs/1804.03599, 2018. URL http://arxiv.org/abs/1804.03599.

[7] Tian Qi Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

[8] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational Lossy Autoencoder. *arXiv.org*, November 2016.

[9] Justin Domke and Daniel Sheldon. Importance weighting and varational inference. *arXiv preprint arXiv:1808.09034*, 2018.

[10] Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=By-7dz-AZ.

[11] Babak Esmaeili, Hao Wu, Sarthak Jain, N Siddharth, Brooks Paige, and Jan-Willem van de Meent. Hierarchical Disentangled Representations. *arXiv.org*, April 2018.

[12] 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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[13] Matthew D Hoffman and Matthew J Johnson. ELBO surgery: yet another way to carve up the variational evidence lower bound. In *Workshop on Advances in Approximate Bayesian Inference, NIPS*, pages 1–4, 2016.

[14] Matthew D Hoffman, Carlos Riquelme, and Matthew J Johnson. The β-VAE's Implicit Prior. In *Workshop on Bayesian Deep Learning, NIPS*, pages 1–5, 2017.

[15] Matthew J Johnson, David Duvenaud, Alexander B Wiltschko, Sandeep R Datta, and Ryan P Adams. Composing graphical models with neural networks for structured representations and fast inference. *arXiv.org*, March 2016.

[16] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *CoRR*, abs/1802.05983, 2018. URL http://arxiv.org/abs/1802.05983.

[17] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv.org*, December 2014.

[18] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations. *arXiv.org*, November 2017.

[19] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[20] Chris J Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. Filtering variational objectives. In *Advances in Neural Information Processing Systems*, pages 6573–6583, 2017.

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

[22] Tom Rainforth, Robert Cornish, Hongseok Yang, Andrew Warrington, and Frank Wood. On nesting monte carlo estimators. In *International Conference on Machine Learning*, pages 4264–4273, 2018.

[23] Tom Rainforth, Adam R. Kosiorek, Tuan Anh Le, Chris J. Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. *International Conference on Machine Learning (ICML)*, 2018.

[24] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryQu7f-RZ.

[25] N. Siddharth, T Brooks Paige, Jan-Willem Van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems*, pages 5925–5935, 2017.

[26] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

[27] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein Auto-Encoders. *arXiv.org*, November 2017.

[28] Jiacheng Xu and Greg Durrett. Spherical Latent Spaces for Stable Variational Autoencoders. *arXiv.org*, August 2018.

[29] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *CoRR*, abs/1706.02262, 2017. URL http://arxiv.org/abs/1706.02262.

A Proofs and Additional Results for the Disentangling the β-VAE

Hoffman et al. [14] showed that the β-VAE target (1) can be interpreted as a standard evidence lower bound (ELBO) with the alternative prior $r(\boldsymbol{z}) \propto \bar{q}(\boldsymbol{z})^{(1-\beta)}p(\boldsymbol{z})^{\beta}$, where $\bar{q}(\boldsymbol{z}) = \frac{1}{n}\sum_{x_i} q(\boldsymbol{z}|\boldsymbol{x}_n)$, along with a term down-weighting mutual information and another based on the prior's normalising constant.

We derive the following alternate expression for the β-VAE.

Theorem 1. *The β-VAE target*

$$\mathcal{L}_\beta(\boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \beta\,\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z})) \tag{1}$$

can be interpreted in terms of the standard ELBO, $\mathcal{L}(\boldsymbol{x})\,(\pi_{\theta,\beta}, q_\phi)$, for an adjusted target $\pi_{\theta,\beta}(\boldsymbol{x},\boldsymbol{z}) \triangleq p_\theta(\boldsymbol{x} \mid \boldsymbol{z})f_\beta(\boldsymbol{z})$ with annealed prior $f_\beta(\boldsymbol{z}) \triangleq p_\theta(\boldsymbol{z})^\beta / F_\beta$ as

$$\mathcal{L}_\beta(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x})\,(\pi_{\theta,\beta}, q_\phi) + (\beta-1)H_{q_\phi} + \log F_\beta \tag{2}$$

where $F_\beta \triangleq \int_{\boldsymbol{z}} p_\theta(\boldsymbol{z})^\beta d\boldsymbol{z}$ is constant given β, and H_{q_ϕ} is the entropy of $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$.

Proof. Starting with (1), we have

$$\begin{aligned}
\mathcal{L}_\beta(\boldsymbol{x}) &= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] + \beta H_{q_\phi} + \beta\,\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{z})] \\
&= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] + (\beta-1)H_{q_\phi} + H_{q_\phi} + \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\Big[\log p_\theta(\boldsymbol{z})^\beta - \log F_\beta\Big] + \log F_\beta \\
&= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] + (\beta-1)H_{q_\phi} - \mathrm{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel f_\beta(\boldsymbol{z})) + \log F_\beta \\
&= \mathcal{L}(\boldsymbol{x})\,(\pi_{\theta,\beta}, q_\phi) + (\beta-1)H_{q_\phi} + \log F_\beta
\end{aligned}$$

as required. □

A.1 Special Case – Gaussians

We analyse the effect of the adjusted target in (2) by studying the often-used Gaussian prior, $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, \Sigma)$, where it is straightforward to see that annealing simply scales the latent space by $1/\sqrt{\beta}$, i.e. $f_\beta(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, \Sigma/\beta)$. Given this, it is easy to see that a VAE trained with the adjusted target $\mathcal{L}(\boldsymbol{x})\,(\pi_{\theta,\beta}, q_\phi)$, but appropriately scaling the latent space, will behave identically to a VAE trained with the original target $\mathcal{L}(\boldsymbol{x})$. They will also have identical ELBOs as the expected reconstruction is trivially the same, while the KL between Gaussians is invariant to scaling both equally.

In fact, including the entropy regulariser allows us to derive a specialisation of (2).

Corollary 1. *If $p_\theta(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, \Sigma)$ and $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}; \mu_\phi(\boldsymbol{x}), S_\phi(\boldsymbol{x})\right)$, then,*

$$\mathcal{L}_\beta(\boldsymbol{x}) = \mathcal{L}\left(p_{\theta'}(\boldsymbol{x} \mid \boldsymbol{z})p_\theta(\boldsymbol{z}), q_{\phi'}(\boldsymbol{z} \mid \boldsymbol{x})\right) + \frac{(\beta-1)}{2}\log|S_{\phi'}(\boldsymbol{x})| + c \tag{4}$$

where θ' and ϕ' represent rescaled networks such that

$$\begin{aligned}
p_{\theta'}(\boldsymbol{x} \mid \boldsymbol{z}) &= p_\theta\Big(\boldsymbol{x} \mid \boldsymbol{z}/\sqrt{\beta}\Big), & q_{\phi'}(\boldsymbol{z}|\boldsymbol{x}) &= \mathcal{N}\left(\boldsymbol{z}; \mu_{\phi'}(\boldsymbol{x}), S_{\phi'}(\boldsymbol{x})\right), \\
\mu_{\phi'}(\boldsymbol{x}) &= \sqrt{\beta}\mu_\phi(\boldsymbol{x}), & S_{\phi'}(\boldsymbol{x}) &= \beta S_\phi(\boldsymbol{x}),
\end{aligned}$$

and where $c \triangleq \frac{D(\beta-1)}{2}\left(1 + \log\frac{2\pi}{\beta}\right) + \log F_\beta$ is a constant, with D denoting the dimensionality of $\boldsymbol{z}$.

Proof. We start by noting that

$$\pi_{\theta,\beta}(\boldsymbol{x}) = \mathbb{E}_{f_\beta(\boldsymbol{z})}[p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] = \mathbb{E}_{p_\theta(\boldsymbol{z})}\Big[p_\theta\Big(\boldsymbol{x} \mid \boldsymbol{z}/\sqrt{\beta}\Big)\Big] = \mathbb{E}_{p_\theta(\boldsymbol{z})}[p_{\theta'}(\boldsymbol{x} \mid \boldsymbol{z})] = p_{\theta'}(\boldsymbol{x})$$

Now considering an alternate form of $\mathcal{L}(\boldsymbol{x})(\pi_{\theta,\beta}, q_\phi)$ in (2),

$$\begin{aligned}
\mathcal{L}(\boldsymbol{x})(\pi_{\theta,\beta}, q_\phi) &= \log \pi_{\theta,\beta}(\boldsymbol{x}) - \mathrm{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel \pi_{\theta,\beta}(\boldsymbol{z} \mid \boldsymbol{x})) \\
&= \log p_{\theta'}(\boldsymbol{x}) - \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log\left(\frac{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})p_{\theta'}(\boldsymbol{x})}{p_\theta(\boldsymbol{x} \mid \boldsymbol{z})f_\beta(\boldsymbol{z})}\right)\right] \\
&= \log p_{\theta'}(\boldsymbol{x}) - \mathbb{E}_{q_{\phi'}(\boldsymbol{z}|\boldsymbol{x})}\left[\log\left(\frac{q_\phi\big(\boldsymbol{z}/\sqrt{\beta} \mid \boldsymbol{x}\big)p_{\theta'}(\boldsymbol{x})}{p_{\theta'}(\boldsymbol{x} \mid \boldsymbol{z})f_\beta(\boldsymbol{z}/\sqrt{\beta})}\right)\right].
\end{aligned} \tag{5}$$

We first simplify $f_\beta(\boldsymbol{z}/\sqrt{\beta})$ as

$$f_\beta(\boldsymbol{z}/\sqrt{\beta}) = \frac{1}{\sqrt{2\pi|\Sigma/\beta|}} \exp\left(-\frac{1}{2}\boldsymbol{z}^T\Sigma^{-1}\boldsymbol{z}\right) = p(\boldsymbol{z})\beta^{(D/2)}.$$

Further, denoting $\boldsymbol{z}_\dagger = \boldsymbol{z} - \sqrt{\beta}\mu_{\phi'}(\boldsymbol{x})$, and $\boldsymbol{z}_\ddagger = \boldsymbol{z}_\dagger/\sqrt{\beta} = \boldsymbol{z}/\sqrt{\beta} - \mu_{\phi'}(\boldsymbol{x})$, we have

$$q_{\phi'}(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{1}{\sqrt{2\pi|S_\phi(\boldsymbol{x})\beta|}} \exp\left(-\frac{1}{2\beta}\boldsymbol{z}_\dagger^T S_\phi(\boldsymbol{x})^{-1}\boldsymbol{z}_\dagger\right),$$
$$q_\phi\left(\frac{\boldsymbol{z}}{\sqrt{\beta}} \mid \boldsymbol{x}\right) = \frac{1}{\sqrt{2\pi|S_\phi(\boldsymbol{x})|}} \exp\left(-\frac{1}{2}\boldsymbol{z}_\ddagger^T S_\phi(\boldsymbol{x})^{-1}\boldsymbol{z}_\ddagger\right)$$

$$\text{giving} \quad q_\phi\Big(\boldsymbol{z}/\sqrt{\beta} \mid \boldsymbol{x}\Big) = q_{\phi'}(\boldsymbol{z} \mid \boldsymbol{x})\beta^{(D/2)}.$$

Plugging these back in to (5), we have

$$\mathcal{L}(\boldsymbol{x})(\pi_{\theta,\beta}, q_\phi) = \log p_{\theta'}(\boldsymbol{x}) - \mathbb{E}_{q_{\phi'}(\boldsymbol{z}|\boldsymbol{x})}\left[\log\left(\frac{q_{\phi'}(\boldsymbol{z} \mid \boldsymbol{x})p_{\theta'}(\boldsymbol{x})}{p_{\theta'}(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{z})}\right)\right] = \mathcal{L}(\boldsymbol{x})\left(p_{\theta'}, q_{\phi'}\right),$$

showing that the ELBOs for the two setups are the same. For the entropy term, we note that

$$H_{q_\phi} = \frac{D}{2}\left(1 + \log 2\pi\right) + \frac{1}{2}\log|S_\phi(\boldsymbol{x})| = \frac{D}{2}\left(1 + \log\frac{2\pi}{\beta}\right) + \frac{1}{2}\log|S_{\phi'}(\boldsymbol{x})|.$$

Finally substituting for H_{q_ϕ} and $\mathcal{L}(\boldsymbol{x})\left(\pi_{\theta,\beta}, q_\phi\right)$ in (2) gives the desired result. □

Noting that c is inconsequential to the training process, this result demonstrates an equivalence, up to the scaling of the latent space, between training using the β-VAE objective and a maximum-entropy regularised version of the standard ELBO

$$\mathcal{L}_{H,\beta}(\boldsymbol{x}) \triangleq \mathcal{L}(\boldsymbol{x}) + \frac{(\beta - 1)}{2}\log|S_\phi(\boldsymbol{x})|, \tag{6}$$

whenever $p_\theta(\boldsymbol{z})$ and $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ are Gaussian. Note that we are here implicitly presuming suitable adjustment of neural-network hyper-parameters and the stochastic gradient scheme to account for the change of scaling in the optimal networks.

More formally we have the following, showing equivalence of all the local optima for the two objectives.

Corollary 2. *If* $\nabla_{\theta,\phi}\mathcal{L}_\beta(\boldsymbol{x}) = \mathbf{0}$ *then*

$$\nabla_{\theta',\phi'}\mathcal{L}_{H,\beta}\left(p_{\theta'}(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{z}), q_{\phi'}(\boldsymbol{z} \mid \boldsymbol{x})\right) = \mathbf{0}. \tag{7}$$

Provided that $\nabla_{\theta',\phi'}\theta$ *and* $\nabla_{\theta',\phi'}\phi$ *do not have any zeros distinct to those of* $\nabla_{\theta,\phi}\mathcal{L}_\beta(\boldsymbol{x})$*, then* (7) *holding also implies* $\nabla_{\theta,\phi}\mathcal{L}_\beta(\boldsymbol{x}) = \mathbf{0}$.

Proof. The proof follows directly from Corollary 1 and the chain rule. □

What we now see is that optimising for (6) leads to a pair of networks equivalent to those from training to the β-VAE target, except that encodings are all scaled by a factor of $\sqrt{\beta}$. While it would be easy to doubt any tangible effects from the rescaling of the β-VAE, closer inspection shows that it still plays an important role: it ensures the scaling of the encodings matches that of the prior. Just adding the entropy regularisation term will increase the scaling of the latent space as the higher variance it encourages will spread out the aggregate posterior $q_\phi(\boldsymbol{z}) = \mathbb{E}_{p_\theta(\boldsymbol{x})}[q_\phi(\boldsymbol{z} \mid \boldsymbol{x})]$. The rescaling of the β-VAE now cancels this effect, ensuring the scaling of $q_\phi(\boldsymbol{z})$ matches that of $p(\boldsymbol{z})$. This is perhaps easiest to see by considering what happens in the limit of large β for the two targets. With the β-VAE, we see from the original formulation that the encoder must provide embeddings equivalent to sampling from the prior. The entropy-regularised VAE on the other hand will produce an encoder with infinite variance. The equivalence between them is apparent when we scale the encodings of the latter by a factor of $1/\sqrt{\beta}$, and recover the encodings of the former, i.e. samples from the prior.

Encoder	Decoder
Input 64 x 64 binary image	Input $\in \mathbb{R}^{10}$
4x4 conv. 32 ReLU stride 2	FC. 128 ReLU
4x4 conv. 32 ReLU stride 2	FC. 4x4 x 64 ReLU
4x4 conv. 32 ReLU stride 2	4x4 upconv. 64 ReLU stride 2
4x4 conv. 64 ReLU stride 2	4x4 upconv. 64 ReLU stride 2
FC. 128	4x4 upconv. 32 ReLU stride 2
FC. 2x10	4x4 upconv. 1. stride 2

(a) 2D-shapes dataset.

Encoder	Decoder
Input $\in \mathbb{R}^2$	Input $\in \mathbb{R}^2$
FC. 100. ReLU	FC. 100 ReLU
FC. 2x2	FC. 2x2

(b) Pinwheel dataset.

Table 1: Encoder and decoder architectures.

B Experimental Details

2d-shapes: The experiments from ğ 3 on the impact of the prior in terms disentanglement are conducted on the **2D Shapes** [21] dataset, comprising of 737,280 binary 64 x 64 images of 2D shapes with ground truth factors [number of values]: shape[3], scale[6], orientation[40], x-position[32], y-position[32]. We use a convolutional neural network for the encoder and a deconvolutional neural network for the decoder, whose architectures are described in Table 1a. We use $[0, 1]$ normalised data as targets for the mean of a Bernoulli distribution, using negative cross-entropy for $\log p(\boldsymbol{x}|\boldsymbol{z})$. We rely on the Adam optimiser [17, 24] with learning rate $1e^{-4}$, β_1 = 0.9, β_2 = 0.999, to optimise the β-VAE objective from (2).

When $p_\theta(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \text{diag}(\sigma))$, experiments have been run with a batch size of 64 and for 20 epochs. When $p_\theta(\boldsymbol{z}) = \prod_i \text{STUDENT-T}(z_i; \nu)$, experiments have been run with a batch size of 256 and for 40 epochs. In Figure 1, the *PCA initialised anisotropic* prior is initialised so that its standard deviations are set to be the first D singular values computed on the observations dataset. These are then mapped through a softmax function to ensure that the β regularisation coefficient is not implicitly scaled compared to the isotropic case. For the *learned anisotropic* priors, standard deviations are first initialised as just described, and then learned along the model through a log-variance parametrisation.

We rely on the metric presented in Section (4) and Appendix (B) of [16] as a measure of axis-alignment of the latent encodings with respect to the true (known) generative factors. Confidence intervals in Figure 1 have been computed via the assumption of normally distributed samples with unknown mean and variance, with 100 runs of each model.

Pinwheel We generated spiral cluster data[1], with $n = 400$ observations, clustered in 4 spirals, with radial and tangential standard deviations respectively of 0.1 and 0.3, and a rate of 0.25. We use fully-connected neural networks for both the encoder and decoder, whose architectures are described in Table 1b. We minimise the objective from (3), with $\mathbb{D}$ chosen to be the inclusive KL, with $q_\phi(\boldsymbol{z})$ approximated by the aggregate encoding of the dataset

$$\begin{aligned}\mathbb{D}\left(q_\phi(\boldsymbol{z}), p(\boldsymbol{z})\right) = \text{KL}\left(p(\boldsymbol{z})||q_\phi(\boldsymbol{z})\right) &= \mathbb{E}_{p(\boldsymbol{z})}\left[\log(p(\boldsymbol{z})) - \log\left(\mathbb{E}_{p_\mathcal{D}(\boldsymbol{x})}[q_\phi(\boldsymbol{z} \mid \boldsymbol{x})]\right)\right] \\ &\approx \sum_{j=1}^{B}\left(\log p(\boldsymbol{z}_j) - \log\left(\sum_{i=1}^{n} q_\phi(\boldsymbol{z}_j \mid \boldsymbol{x}_i)\right)\right)\end{aligned}$$

with $\boldsymbol{z}_j \sim p(\boldsymbol{z})$. A Gaussian likelihood is used for the encoder. We trained the model for 500 epochs using the Adam optimiser [17, 24], with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and a learning rate of $1e^{-3}$. The batch size is set to $B = n$.

The mixture of Gaussian prior (c.f. Figure 3) is defined as

$$\begin{aligned}p(\boldsymbol{z}) &= \sum_{c=1}^{C} \pi_c \, \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \\ &= \sum_{c=1}^{C} \pi_c \prod_{d=1}^{D} \mathcal{N}(\boldsymbol{z}^d|\mu_c^d, \sigma_c^d)\end{aligned} \tag{8}$$

with $D = 2$, $C = 4$, $\boldsymbol{\Sigma}_c = .03\, I_D$, $\pi_c = 1/C$ and $\mu_c^d \in \{0, 1\}$.

[1] http://hips.seas.harvard.edu/content/synthetic-pinwheel-data-matlab.

Figure 3: PDF of Gaussian mixture model prior, i.e. $p(\boldsymbol{z})$ as per (8).

C Posterior regularisation

The aggregate posterior regulariser $\mathbb{D}(q(\boldsymbol{z}), p(\boldsymbol{z}))$ is a little more subtle to analyse than the entropy regulariser as it involves both the choice of divergence and potential difficulties in estimating that divergence. One possible choice is the exclusive Kullback-Leibler divergence $\mathrm{KL}(q(\boldsymbol{z}) \,\|\, p(\boldsymbol{z}))$, as previously used (without additional entropy regularisation) by [2, 11], but also implicitly by [7, 16], through the use of a total correlation (TC) term. We now highlight a shortfall with this choice of divergence due to difficulties in its empirical estimation.

In short, the approaches used to estimate the $\mathrm{H}[q(\boldsymbol{z})]$ (noting that $\mathrm{KL}(q(\boldsymbol{z}) \,\|\, p(\boldsymbol{z})) = -\mathrm{H}[q(\boldsymbol{z})] - \mathbb{E}_{q(\boldsymbol{z})}[\log p(\boldsymbol{z})]$, where the latter term can be estimated reliably by a simple Monte Carlo estimate) exhibit very large biases that result in quite different effects from what was intended. In fact, our results suggest they will exhibit behavior similar to the β-VAE. These biases arise from the effects of nesting estimators [22], where the variance in the nested (inner) estimator for $q(\boldsymbol{z})$ induces a bias in the overall estimator. Specifically, for any random variable $\hat{Z}$,

$$\mathbb{E}[\log(\hat{Z})] = \log(\mathbb{E}[\hat{Z}]) - \frac{\mathrm{Var}[\hat{Z}]}{2Z^2} + O(\varepsilon) \tag{9}$$

where $O(\varepsilon)$ represents higher-order moments that get dominated asymptotically if $\hat{Z}$ is a Monte-Carlo estimator (see Proposition 1c in Maddison et al. [20], Theorem 1 in Rainforth et al. [23], or Theorem 3 in Domke and Sheldon [9]). In this setting, $\hat{Z} = \hat{q}(\boldsymbol{z})$ is the estimate used for $q(\boldsymbol{z})$. We thus see that if the variance of $\hat{q}(\boldsymbol{z})$ is large, this will induce a significant bias in our KL estimator.

To make things precise, we consider the estimator used for $\mathrm{H}[q(\boldsymbol{z})]$ in Esmaeili et al. [11] and Anonymous [2] (noting that the analysis applies equally to those of Chen et al. [7]):

$$\mathrm{H}[q(\boldsymbol{z})] \approx \hat{\mathrm{H}} \triangleq -\frac{1}{B}\sum_{b=1}^{B} \log \hat{q}(\boldsymbol{z}_b), \tag{10a}$$

$$\text{where} \quad \hat{q}(\boldsymbol{z}_b) = \frac{q_\phi(\boldsymbol{z}_b|\boldsymbol{x}_b)}{n} + \frac{n-1}{n(B-1)}\sum_{b' \neq b} q_\phi(\boldsymbol{z}_b|\boldsymbol{x}'_b), \tag{10b}$$

each $\boldsymbol{z}_b \sim q_\phi(\boldsymbol{z}|\boldsymbol{x}_b)$, and $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_B\}$ is the mini-batch of data being used for the current iteration and n is the dataset size. Esmaeili et al. [11] correctly show that $\mathbb{E}[\hat{q}(\boldsymbol{z}_b)] = \tilde{q}(\boldsymbol{z}_b)$, with the first term of (10b) comprising an exact term in $\tilde{q}(\boldsymbol{z}_b)$ and the second term of (10b) being an unbiased Monte-Carlo estimate for the remaining terms in $\tilde{q}(\boldsymbol{z}_b)$.

To examine the practical behaviour of this estimator when $B \ll n$, we first note that the second term of (10b) is, in practice, usually very small and dominated by the first term. This is borne out empirically in our own experiments, and also noted in Kim and Mnih [16]. To see why this is the case, consider that given encodings of two independent data points, it is highly unlikely that the two encoding distributions will have any notable overlap (e.g. for a Gaussian encoder, the means will most likely be very many standard deviations apart), presuming a sensible latent space is being learned. Consequently, even though this second term is unbiased and may have an expectation comparable or even larger than the first, it is heavily skewed—it is usually negligible, but occasionally large in the rare instances where there is substantial overlap between encodings.

Let the second term of (10b) be denoted T_2 and the event that this it is significant be denoted E_S, such that $\mathbb{E}[T_2 \mid \neg E_s] \approx 0$. As explained above, it will typically be the case that $\mathbb{P}(E_S) \ll 1$. We now have

$$\begin{aligned}
\mathbb{E}\left[\hat{\mathrm{H}}\right] &= (1 - \mathbb{P}(E_S))\, \mathbb{E}\left[\hat{\mathrm{H}} \mid \neg E_S\right] + \mathbb{P}(E_S)\, \mathbb{E}\left[\hat{\mathrm{H}} \mid E_S\right] \\
&= (1 - \mathbb{P}(E_S)) \left(\log n - \frac{1}{B} \sum_{b=1}^{B} \mathbb{E}[\log q_\phi(\boldsymbol{z}_b \mid \boldsymbol{x}_b) \mid \neg E_S] - \mathbb{E}[T_2 \mid \neg E_S] \right) + \mathbb{P}(E_S)\, \mathbb{E}\left[\hat{\mathrm{H}} \mid E_S\right] \\
&= (1 - \mathbb{P}(E_S))\, (\log n - \mathbb{E}[\log q_\phi(\boldsymbol{z}_1 \mid \boldsymbol{x}_1) \mid \neg E_S] - \mathbb{E}[T_2 \mid \neg E_S]) + \mathbb{P}(E_S)\, \mathbb{E}\left[\hat{\mathrm{H}} \mid E_S\right] \\
&\approx (1 - \mathbb{P}(E_S))\, (\log n - \mathbb{E}[\log q_\phi(\boldsymbol{z}_1 \mid \boldsymbol{x}_1)]) + \mathbb{P}(E_S)\, \mathbb{E}\left[\hat{\mathrm{H}} \mid E_S\right]
\end{aligned}$$

where the approximation relies firstly on our previous assumption that $\mathbb{E}[T_2 \mid \neg E_S] \approx 0$ and also that $\mathbb{E}[\log q_\phi(\boldsymbol{z}_1|\boldsymbol{x}_1) \mid \neg E_S] \approx \mathbb{E}[\log q_\phi(\boldsymbol{z}_1|\boldsymbol{x}_1)]$. This second assumption will also generally hold in practice, firstly because the occurrence of E_S is dominated by whether two or not similar datapoints are drawn (rather than by the value of $\boldsymbol{x}_1$) and secondly because $\mathbb{P}(E_S) \ll 1$ implies that

$$\begin{aligned}
\mathbb{E}[\log q_\phi(\boldsymbol{z}_1 \mid \boldsymbol{x}_1)] &= (1 - \mathbb{P}(E_S))\, \mathbb{E}[\log q_\phi(\boldsymbol{z}_1 \mid \boldsymbol{x}_1) \mid \neg E_S] + \mathbb{P}(E_S)\, \mathbb{E}[\log q_\phi(\boldsymbol{z}_1 \mid \boldsymbol{x}_1) \mid E_S] \\
&\approx \mathbb{E}[\log q_\phi(\boldsymbol{z}_1 \mid \boldsymbol{x}_1) \mid \neg E_S].
\end{aligned}$$

Characterising $\mathbb{E}\left[\hat{\mathrm{H}} \mid E_S\right]$ precisely is a little more challenging, but it can safely be assumed to be smaller than $\mathbb{E}[\log q_\phi(\boldsymbol{z}_1 \mid \boldsymbol{x}_1)]$, which is approximately what would result from all the $\boldsymbol{x}'_b$ being the same as $\boldsymbol{x}_b$. We thus see that even when the event E_S does occur, the resulting gradients should still be on a comparable scale to when it does not. Consequently, whenever E_S is rare, the $(1 - \mathbb{P}(E_S))\, \mathbb{E}\left[\hat{\mathrm{H}} \mid \neg E_S\right]$ term should dominate and we thus have

$$\mathbb{E}\left[\hat{\mathrm{H}}\right] \approx \log n - \mathbb{E}[\log q_\phi(\boldsymbol{z}_1 \mid \boldsymbol{x}_1)] = \log n + \mathbb{E}_{p(\boldsymbol{x})}[\mathrm{H}[q_\phi(\boldsymbol{z} \mid \boldsymbol{x})]]. \tag{11}$$

More significantly, we see that the estimator mimics the $\beta-$VAE regularisation up to a constant factor $\log n$, as adding the $\mathbb{E}_{q(\boldsymbol{z})}[\log p(\boldsymbol{z})]$ back in gives

$$-\mathbb{E}\left[\hat{\mathrm{H}}\right] - \mathbb{E}_{q(\boldsymbol{z})}[\log p(\boldsymbol{z})] \approx \mathbb{E}_{p(\boldsymbol{x})}[\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \,\|\, p(\boldsymbol{z}))] - \log n. \tag{12}$$

We should thus expect to empirically see training with this estimator as a regulariser to behave similarly to the $\beta-$VAE with the same regularisation term whenever $B \ll n$. Note that the $\log n$ constant factor will not impact the gradients, but does mean that it is possible, even likely, that negative estimates for $\hat{\mathrm{KL}}$ will be generated, even though we know the true value is positive.