On the properties of high-capacity VAEs

Danilo J. Rezende [*] **Fabio Viola** [*]

{danilor, fviola}@google.com
DeepMind, London, UK

1 Introduction and related work

We present a detailed theoretical analysis of the behaviour of high-capacity β-VAEs [1], advancing our understanding of VAEs on multiple fronts: (i) We demonstrate that the posterior density of unconstrained VAEs will converge to an equiprobable partition of the latent-space; (ii) We provide a connection between β-VAEs and spectral clustering; (iii) Drawing from statistical mechanics we study phase-transitions in the reconstruction fixed-points of β-VAEs.

Our analysis extends the analysis performed in [2, 3] with an emphasis on the properties of the learned posterior densities in high-capacity VAEs. Our analysis also extends the idea of information bottleneck and geometrical clustering from [4] to the case where latent variables are continuous instead of discrete. A further extension of the analysis including the incorporation of Lipschitz constraints is also available in Appendix B. Complementary to the analysis from [5] with (semi-)affine assumptions on the VAE's decoder, we focus on non-linear aspects of high-capacity constrained VAEs. To obtain a tractable representation of both the posterior density and decoder for Proposition 2 and Proposition 4, we express them in a particular functional basis that is analogous to using a mixture posterior density as in [6] but where the modes of the mixture have non-overlapping supports. This assumption substantially simplify our analysis, allowing us to reach meaningful conclusions.

The remainder of the paper is structured as follows: in Section 2.1 we study the behavior of unconstrained VAEs at convergence; in Section 2.2 we draw links between β-VAEs, spectral clustering and statistical mechanic.

2 Analysis

We consider β-VAEs [1] of the form $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})\pi(\mathbf{z})$ that are trained by maximizing the modified variational objective $\mathcal{F}$, using a parametric variational posterior $q(\mathbf{z}|\mathbf{x})$,

$$\mathcal{F} = \mathbb{E}_{\rho(\mathbf{x})}\left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\ln p(\mathbf{x}|\mathbf{z})\right] - \beta \mathrm{KL}\left[q; \pi\right]\right], \tag{1}$$

where β is a positive scaling factor used to trade-off reconstruction error and compression rates and $\pi(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbb{I})$. The empirical data density is represented as $\rho(\mathbf{x}) = \frac{1}{n}\sum_i^n \delta(\mathbf{x} - \mathbf{x}_i)$, where $\mathbf{x}_i$ are individual data-points.

We focus on VAEs with Gaussian decoder density of the form $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|g(\mathbf{z}), \sigma^2\mathbb{I})$, where σ is a global parameter and $g(\mathbf{z})$ is referred to as a decoder or generator. Restricting to decoder densities where the components x_i of $\mathbf{x}$ are conditionally independent given a latent vector $\mathbf{z}$ eliminates a family of solutions in the infinite-capacity limit where the decoder density $p(\mathbf{x}|\mathbf{z})$ ignores the latent variables, i.e. $p(\mathbf{x}|\mathbf{z}) = p(\mathbf{x}) \approx \rho(\mathbf{x})$, as observed in [3]. This restriction is important because in this work we are interested in the behaviour of the latent variables.

[*] Both authors contributed equally to this work.

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

Proposition 1. *(Fixed-point equations) The extrema of the modified ELBO $\mathcal{F}$ with respect to the decoder $g(\mathbf{z})$ and encoder $q(\mathbf{z}|\mathbf{x})$ are solutions of the fixed-point Equations* (2) *and* (3)

$$q^t(\mathbf{z}|\mathbf{x}) \propto \pi(\mathbf{z}) e^{-\frac{\|\mathbf{x}-g_{\mathcal{F}}^{t-1}(\mathbf{z})\|^2}{2\beta\sigma^2}} \tag{2}$$

$$g^t(\mathbf{z}) = \sum_i w_i^t(\mathbf{z})\mathbf{x}_i \tag{3}$$

where $w_i^t(\mathbf{z}) = \frac{q^t(\mathbf{z}|\mathbf{x}_i)}{\sum_j q^t(\mathbf{z}|\mathbf{x}_j)}$. *The proof is sketched in Appendix C.1.*

In the following Sections 2.1 and 2.2 we analyze properties of Equations (2) and (3) constraints to gain a better understanding of the behaviour of VAEs. In Appendix B we also provide additional analysis of the effect of local and global Lipschitz constraints.

2.1 Unconstrained VAEs

In the unconstrained case, where we optimize the ELBO ($\beta = 1$), we can derive a few interesting conclusions from Equations (2) and (3): (i) The global optimal decoder $g(\mathbf{z})$ is a convex linear combination of the training data of the form $g(\mathbf{z}) = \sum_i w_i(\mathbf{z})\mathbf{x}_i$; (ii) If we optimize the standard-deviation σ jointly with the decoder and encoder, it will converge to zero; (iii) The posterior density $q(\mathbf{z}|\mathbf{x})$ will converge to a distribution with support corresponding to one element of a partition of the latent space. Moreover, the set of supports of the posterior density formed by each data-point constitutes a partition of the latent space that is equiprobable under the prior. These results are illustrated in Figure 1(left) and Figure 2, and formalized in Proposition 2, where we demonstrate that a solution satisfying all these properties is a fixed point of Equations (2) and (3).

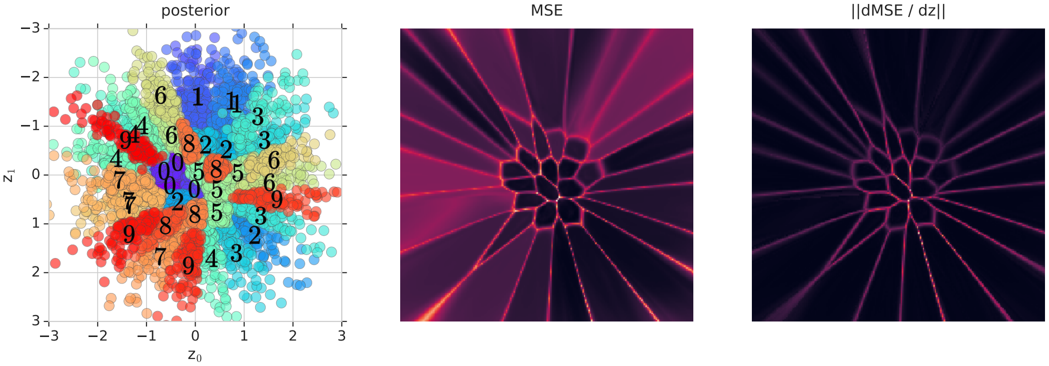

Figure 1: **Example of latent space tiling induced by the model posterior.** Visualization of properties of the tiling of the latent space induced by the posterior of a VAE with a Normal prior, NVP [7] encoder and 2 dimensional latent space, trained on a subset of MNIST comprising 4 exemplars per digit label. **Left**: Scatter plot of the posterior showing 100 z samples per exemplar, color-coded based on the exemplar id; labels are also overlaid to help identify the corresponding data classes. The posterior density $q(\mathbf{z}|\mathbf{x})$ partitions the latent space into tiles, equiprobable under the prior, each corresponding to one data point in the training set; refer to Proposition 2 for details. **Centre, Right**: Maps of mean-squared error (MSE) and norm of gradient of MSE *w.r.t.* z. For each z we first decode the latent, then identify the nearest neighbor (NN) in the training set, and finally compute the MSE and its gradient using the NN as target. The reconstruction error is approximately constant within a tile, and tile boundaries are aligned with regions of the latent space characterized by high reconstruction error gradient variance, caused by blurred reconstructions; refer to Proposition 3 for details.

Proposition 2. *(High-capacity VAEs learn an equiprobable partition of the latent space): Let* $q(\mathbf{z}|\mathbf{x}_i) = \pi(\mathbf{z})\mathbb{I}_{\mathbf{z}\in\Omega_i}/\pi_i$, *where* $\pi_i = \mathbb{E}_\pi\left[\mathbb{I}_{\mathbf{z}\in\Omega_i}\right]$ *is a normalization constant, be the variational posterior density, evaluated at a training point* x_i. *This density is equal to the restriction of the prior* $\pi(\mathbf{z})$ *to a limited support* $\Omega_i \subset \mathbb{R}^{d_z}$ *in the latent space.* $q(\mathbf{z}|\mathbf{x}_i)$ *is a fixed-point of equations Equations* (2) *and* (3) *for any set of volumes* Ω_i *that forms a partition of the latent space. Furthermore, the highest ELBO is achieved when the partition is equiprobable under the prior. That is,* $\mathbb{E}_\pi\left[\mathbb{I}_{\mathbf{z}\in\Omega_i}\right] = \mathbb{E}_\pi\left[\mathbb{I}_{\mathbf{z}\in\Omega_j}\right] = 1/n$, $\mathbb{R}^{d_z} = \cup_i\Omega_i$ *and* $\Omega_i \cap \Omega_j = \varnothing$ *if* $i \neq j$. *The proof is sketched in Appendix C.2.*

The fact that the standard deviation will converge to 0 results in a numerically ill-posed problem as observed in [8]. Nevertheless, the fixed-point equations still admits a stationary solution where the VAE becomes a mixture of Dirac-delta densities centered at the training data-points.

It is known empirically that low-capacity VAEs tend to produce blurred reconstructions and samples [2, 1]. Contrary to a popularly held belief (e.g. [9]), this phenomenon is not caused by using a Gaussian likelihood alone: as observed in [2], this is primarily caused by a sub-optimal variational posterior. The fixed-point equation (3) provides a mathematical explanation for this phenomenon, generalizing the result from [2]: the optimal decoder $g(\mathbf{z})$ for a given encoder $q(\mathbf{z}|\mathbf{x})$ is a convex linear combination of the training data. If the VAE's encoder cannot accurately distinguish between multiple training data-points, the resulting weights $w_i(\mathbf{z})$ in the VAE's decoder will be spread across the same data-points, resulting in a blurred reconstruction. This is formalized in Proposition 3 and empirically illustrated in Figure 1(middle).

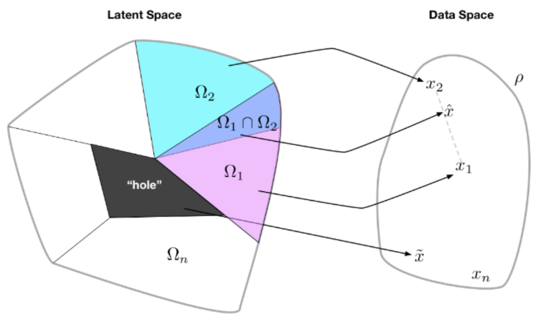

Figure 2: **Illustration of the "blurred reconstructions" and the "holes" problems. Left**: Latent space with a posterior with support in a tiling $\{\Omega_i\}$, where each tile Ω_i represents the support of the posterior for the data-point x_i. **Right**: Data space. In the region of the latent space where the posteriors of the data-points x_1 and x_2 overlap, $\Omega_1 \cap \Omega_2$, the optimal reconstruction $\hat{x}$ is a weighted average of the corresponding data-points, resulting in a blurred sample. In a region of low density under the marginal posterior, a "hole" (represented by the black area in the figure), the optimal reconstructions from these regions $\tilde{x}$ are unconstrained by the ELBO objective function.

Proposition 3. *(Blurred reconstructions) If the supports Ω_i from Proposition 2 are overlapping (i.e. $\Omega_i \cap \Omega_j \neq \varnothing$ for $i \neq j$), then the optimal reconstruction at a latent point $\mathbf{z}$ for a fixed encoder $q(\mathbf{z}|x_i) = \pi(\mathbf{z})\mathbb{I}_{z \in \Omega_i}/\pi_i$ will be the average of all data-points mapping to any of the overlapping basis weighted by the inverse prior probability of the respective basis. See proof in Appendix C.3.*

Another striking conclusion we can derive from Propositions 1 and 3 is that the support of the optimal decoder as a function of the latent vector will be concentrated in the support of the marginal posterior. In fact, if we revisit the proof of Proposition 1 in Appendix C.1 when there are regions in the latent space where $q(\mathbf{z}|\mathbf{x}) \approx 0$ we notice that the decoder is completely unconstrained by the ELBO in these regions. We refer to this as the "holes problem" in VAEs. This problem is commonly encountered when using simple Gaussian posteriors (*e.g.* [10]).

2.2 High-capacity β-VAEs and spectral methods

The β-coefficient in a β-VAE [1] can be interpreted as the Lagrange multiplier of an inequality constraint imposing either a restriction on the value of the KL-term or a constraint on the reconstruction error [11, 3]. When using the reconstruction error constraint $\mathcal{C}(\mathbf{x}, g(\mathbf{z})) = \|\mathbf{x} - g(\mathbf{z})\|^2 - \kappa^2$ in Equation (1), the Lagrange multiplier $\boldsymbol{\lambda}$ is related to the β from [1] by $\boldsymbol{\lambda} = \frac{1}{\beta}$.

While VAEs with simple linear decoders can be related to a form of robust PCA, [5], we demonstrate a relation between β-VAEs with high-capacity decoders and kernel methods such as spectral clustering. More precisely, we show in Proposition 4 that the fixed-point equations of a high-capacity decoder expressed in a particular orthogonal basis are analogous to the reconstruction fixed point equations used in spectral clustering.

In the literature of spectral clustering and kernel PCA it is known that reconstructions based on a Gaussian Gram matrix may suffer phase-transitions (sudden change in eigen-values or reconstruction fixed-points) as a function of the scale parameter [12, 13, 14, 15, 16]. Making a bridge between VAEs and these methods allows us to investigate phase-transitions in high-capacity β-VAEs, where the

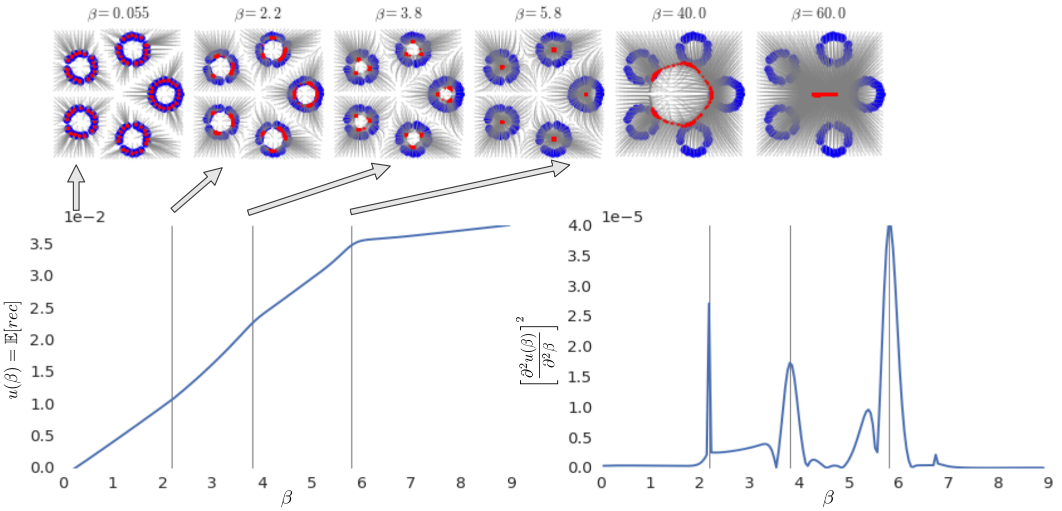

Figure 3: **Effect of β on the reconstruction fixed-points and phase-transitions**. **Top images** Grey curves indicate the trajectories of the vectors ψ. Red points are the fixed-points. Blue points are the data points; **Bottom left** Expected reconstruction error as a function of β. Vertical grey lines indicate the detected critical-temperatures β_c^k; **Bottom right** Expected second-order derivative of the reconstruction error as a function of β. At critical temperatures reconstruction fixed-points will merge with each other, resulting in sudden changes in the slope of the reconstruction error with respect to the temperature β, these points correspond to spikes in the second-order derivatives. For analysis, we sorted the local maxima according to their height and restricted the analysis to the top-3 points, $\beta_c^{k=1,2,3}$. Details of this simulation are explained in Appendix A.

expected reconstruction error is treated as an order parameter $u(\beta) = \mathbb{E}\left[\|\mathbf{x} - g(\mathbf{z})\|^2\right]$. In this case, phase transitions will occur at critical temperature points β_c, which can be detected by analyzing regions of high-curvature (high absolute second-order derivative $|\frac{\partial^2 u(\beta)}{\partial^2 \beta}|$). As in spectral clustering, β-VAEs phase transitions correspond to the merging of neighboring data-clusters at different spatial scales. This is illustrated in the experiment shown in Figure 3, where we look at the merging of reconstruction fixed-points as we increase β. Interestingly, the phase-transitions that we observe are similar to what is known as first-order phase transitions in statistical mechanics [17], but further analysis is necessary to clarify this connection.

Proposition 4. *(High-capacity β-VAE and spectral methods) Let $\phi_a : \mathbb{R}^{d_z} \rightarrow \{0, 1\}$ be an orthogonal basis in the latent space. If we express the posterior density and generator using this basis respectively as $q(\mathbf{z}|x_i) = \pi(\mathbf{z}) \sum_a m_{ia} \phi_a(\mathbf{z})$ and $g(\mathbf{z}) = \psi^T \phi(\mathbf{z})$, where m_{ia} is a matrix with positive entries satisfying the constraint $\sum_a m_{ia} \pi_a = 1$ where $\pi_a = \mathbb{E}_{\pi(\mathbf{z})}[\phi_a(\mathbf{z})]$. The fixed-point equations that maximize the ELBO with respect to ψ are convergent under the appropriate initial conditions and are equivalent to computing the pre-images (reconstructions) of a Kernel-PCA model with a normalized Gaussian Kernel with scale parameter $\sqrt{\beta}$. The proof is sketched in Appendix C.4.*

In statistical mechanics there is an important result, known as *equipartition of energy theorem*. It states that, for a system in thermodynamic equilibrium, energy is shared equally amongst all accessible degrees of freedom at a fixed energy level, [18]. We demonstrate that a similar theorem holds for VAEs in Proposition 5. In Proposition 2 we have seen that the optimal posterior for each data point in a high-capacity VAE has its support in the elements of a partition of the latent space and that this partition is equiprobable under the prior. We can generalize this result to β-VAEs by noticing that, as a result of the existence of reconstruction fixed points from Proposition 4, there will be regions in the latent space where the Hamiltonian $H(\mathbf{x}, \mathbf{z})$ is approximately constant for a given $\mathbf{x}$. At these regions, the posterior will be proportional to the prior and they will work as a discrete partition of the latent space, as in Proposition 2. The concept that VAE encoders learn a tiling of the latent space, each tile corresponding to a different level of the function $H(\mathbf{x}, \mathbf{z})$, can be a guiding principle to evaluate generative models as well as to construct more meaningful constraints.

Proposition 5. *(Equipartition of Energy for high-capacity VAEs). Let $H(\mathbf{x}, \mathbf{z})$ be the "Hamiltonian" function from Proposition 1 for a VAE trained in a dataset with n data-points. For a given data-point*

$\mathbf{x} \in \mathbb{R}^{d_x}$*, latent point* $\mathbf{z} \in \mathbb{R}^{d_z}$ *and precision* $\epsilon > 0$*, let* $\Omega(\mathbf{x}, \mathbf{z}_0) = \{z' || H(\mathbf{x}, \mathbf{z}') - H(\mathbf{x}, \mathbf{z}_0)| \leqslant \epsilon\} \subseteq \mathbb{R}^{d_z}$*. That is,* $\Omega(\mathbf{x}, \mathbf{z}_0)$ *is the set of latent points where the Hamiltonian is approximately constant. As we vary* $\mathbf{x}$ *and* $\mathbf{z}_0$*, each set* $\Omega(\mathbf{x}, \mathbf{z}_0)$ *will be one element of a discrete set of disjoint sets, which we enumerate as* Ω_a*. The encoder density* $q(\mathbf{z}|\mathbf{x}_i)$ *will converge to a mixture of the restrictions of the prior to the basis elements* Ω_a*. Moreover, the probability* γ_a *of a sample from the prior falling in the partition element* Ω_a *is a solution of* $\sum_i \frac{e^{-\frac{H_{ia}}{\beta}}}{\sum_b e^{-\frac{H_{ib}}{\beta}} \gamma_b} = n$ *where* H_{ia} *is the value of* $H(\mathbf{x}_i, \mathbf{z})$ *for* $\mathbf{z} \in \Omega_a$*. The proof is sketched in Appendix C.5.*

3 Discussion

We have provided a detailed theoretical analysis of the behavior of high-capacity VAEs and variants of β-VAEs. We have made connections between VAEs, spectral clustering methods and statistical mechanics (phase transitions). Our analysis provides novel insights to the two most common problems with VAEs: blurred reconstructions/samples, and the "holes problem".

References

[1] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. b-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[2] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Towards deeper understanding of variational autoencoding models. *arXiv preprint arXiv:1702.08658*, 2017.

[3] Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a broken elbo. 2017.

[4] DJ Strouse and David J Schwab. The information bottleneck and geometric clustering. *arXiv preprint arXiv:1712.09657*, 2017.

[5] Bin Dai, Yu Wang, John Aston, Gang Hua, and David O Wipf. Hidden talents of the variational autoencoder. 2017.

[6] Eric T. Nalisnick, Lars Hertel, and Padhraic Smyth. Approximate inference for deep latent gaussian mixtures. 2016.

[7] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[8] Pierre-Alexandre Mattei and Jes Frellsen. Leveraging the exact likelihood of deep latent variable models. *CoRR*, abs/1802.04826, 2018.

[9] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016.

[10] Diederik P. Kingma, Tim Salimans, and Max Welling. Improved variational inference with inverse autoregressive flow. 2016.

[11] Christopher Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. 2018.

[12] David Hoyle and Magnus Rattray. Limiting form of the sample covariance eigenspectrum in pca and kernel pca. In *Advances in Neural Information Processing Systems*, pages 1181–1188, 2004.

[13] Boaz Nadler, Stephane Lafon, Ronald Coifman, and Ioannis G Kevrekidis. Diffusion maps-a probabilistic interpretation for spectral embedding and clustering algorithms. In *Principal manifolds for data visualization and dimension reduction*, pages 238–260. Springer, 2008.

[14] Quan Wang. Kernel principal component analysis and its applications in face recognition and active shape models. *arXiv preprint arXiv:1207.3538*, 2012.

[15] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, pages 583–588. Springer, 1997.

[16] Sebastian Mika, Bernhard Schölkopf, Alex J Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. Kernel pca and de-noising in feature spaces. In *Advances in neural information processing systems*, pages 536–542, 1999.

[17] Stephen J Blundell and Katherine M Blundell. *Concepts in thermal physics*. OUP Oxford, 2009.

[18] Richard Chace Tolman. *The principles of statistical mechanics*. Courier Corporation, 1938.

[19] Mícheál O'Searcoid. *Metric spaces*. Springer Science & Business Media, 2006.

[20] Maxwell Rosenlicht. *Introduction to analysis*. Courier Corporation, 1968.

Appendix A Reconstruction fixed-points experiment details

To produce the results in Figure 3, we iterated the fixed-point equations for the matrix ψ^t using exponential smoothing with $\alpha = 0.9$. For each experiment, we iterated the smoothed fixed-point equations until either the number of iterations exceeded 400 steps or the Euclidean distance between two successive steps became smaller than 1e-3. The basis functions $\phi_i(z)$ were arranged as a 32x32 grid in a compact latent space $\mathbf{z} \in [-\frac{1}{2}, \frac{1}{2}]^2$, the prior was chosen to be uniform, $\pi(\mathbf{z}) = 1$. The matrix ψ was initialized at the center-points of the grid tiles with small uniform noise in $[-0.1, 0.1]$.

Appendix B High-capacity β-VAEs and Lipschitz constraints

The analysis of high-capacity VAEs in Sections 2.1 and 2.2 reveal interesting aspects of VAEs near convergence, but the type of solutions implied by Propositions 2 and 4 may seem unrealistic for VAEs with smooth decoders parametrized by deep neural networks. For instance, the solutions of the fixed point-equations from Propositions 2 and 4 have no notion that the outputs of the decoder $g(\mathbf{z})$ and $g(\mathbf{z}')$ for two similar latent vectors $\mathbf{z}$ and $\mathbf{z}'$ should also be similar. That is, these solutions are not sensitive to the metric and topological properties of the latent space. This implies that, if we want to work in a more realistic high-capacity regime, we must further constrain the solutions so that they have at least a continuous limit as we grow the number of basis functions to infinity.

A sufficient condition for a function $g(\mathbf{z})$ to be continuous, is that it is a locally-Lipschitz function of $\mathbf{z}$ [19]. Thus, to bring our analysis closer to more realistic VAEs, we consider high-capacity β-VAEs with an extra L-Lipschitz inequality constraint $\mathcal{C}$ in the decoder function $g(\mathbf{z})$. The new term that we add to the augmented Lagrangian, with a functional Lagrange multiplier $\Omega(\mathbf{z}, \mathbf{z}') \geq 0$ is given by

$$\mathcal{C}[g] = \frac{1}{2} \int d\mathbf{z} d\mathbf{z}' \pi(\mathbf{z}) \pi(\mathbf{z}') \Omega(\mathbf{z}, \mathbf{z}') \left[\|g(\mathbf{z}) - g(\mathbf{z}')\|^2 - L^2 \|\mathbf{z} - \mathbf{z}'\|^2 \right]. \tag{4}$$

By expressing $g(\mathbf{z})$ and $\Omega(\mathbf{z}, \mathbf{z}')$ in the functional basis $\phi(\mathbf{z})$ of Proposition 4, $g(\mathbf{z}) = \sum_a \psi_a \phi_a(\mathbf{z})$ and $\Omega(\mathbf{z}, \mathbf{z}') = \sum_{a,b} \hat{\Omega}_{ab} \phi_a(\mathbf{z}) \phi_b(\mathbf{z})$, we can rewrite the constraint term as a quadratic inequality constraint in the matrix ψ,

$$\mathcal{C}[g] = \frac{1}{2} \sum_{a,b} \tilde{\Omega}_{ab} \left[C_{ab} \|\psi_a - \psi_b\|^2 - 1 \right], \tag{5}$$

where $\tilde{\Omega}_{ab} = L^2 K_{ab} \hat{\Omega}_{ab}$ are new Lagrange multipliers, $C_{ab} = \pi_a \pi_b / (L^2 K_{ab})$ and $K_{ab} = \int d\mathbf{z} d\mathbf{z}' \phi_a(\mathbf{z}) \phi_b(\mathbf{z}') \|\mathbf{z} - \mathbf{z}'\|^2$. The matrices C_{ab} and K_{ab} embed the metric and topological properties of the latent space but are otherwise independent of the rest of the model.

The constraint (4) can be used to enforce both global and local Lipschitz constraints by controlling the size of the support of the Lagrange multiplier function $\Omega(\mathbf{z}, \mathbf{z}')$. If $\Omega(\mathbf{z}, \mathbf{z}') = 0$ for $\|\mathbf{z} - \mathbf{z}'\| >= r$, we will be constraining the decoder function to be locally Lipschitz within a radius r in the latent space.

Note that the Lipschitz constraint is not a reconstruction constraint as it only constrains the VAE's decoder at arbitrary points in the latent space. For this reason, it can be implemented as a projection step just after the iteration from Equation (15). This is formalized in Proposition 6. We illustrate the combined effect of β, local and global Lipschitz constraints on VAEs in Figure 4.

Proposition 6. *The Lipschitz constraint from Equation* (5) *can be incorporated to the fixed-point Equation* (15) *as a projection of the form $\psi^{t+1} = F(\psi^t) P^t$, where F is the transition operator without Lipschitz constraints. See proof in Appendix C.6.*

Appendix C Proofs

C.1 Derivation of Proposition 1

Proof. We can obtain these equations by taking the functional derivatives $\frac{\delta \mathcal{F}}{\delta g(\mathbf{z})}$, $\frac{\delta \mathcal{F}}{\delta q(\mathbf{z}|\mathbf{x})}$ of $\mathcal{F}$ with respect to $g(\mathbf{z})$ and $q(\mathbf{z}|\mathbf{x})$ respectively and re-arranging the terms of the equations $\frac{\delta \mathcal{F}}{\delta q(\mathbf{z}|\mathbf{x})} = 0$, $\frac{\delta \mathcal{F}}{\delta g(\mathbf{z})} =$

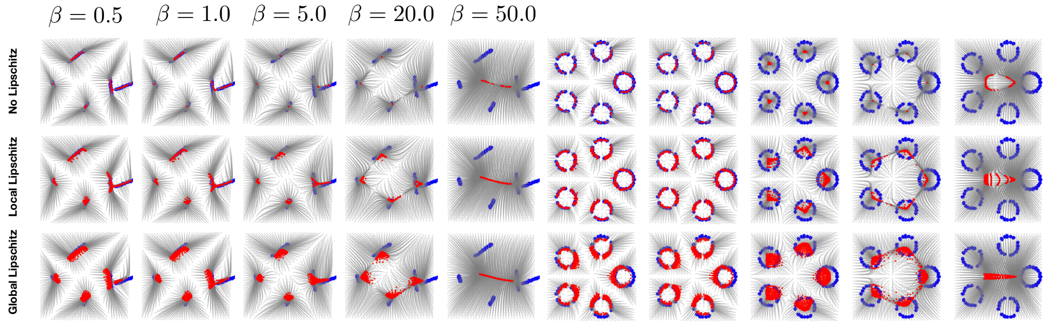

Figure 4: **Combined effect of β-VAEs and Lipschitz constraints**. Blue points are data-points for a mixture of lines (left) and a mixture of circles (right). Grey curves indicate the trajectories of the reconstruction vectors ψ from initial conditions on an uniform 2D grid. Red points are the found fixed-points. **Top row** Reconstruction fixed-points without Lipschitz constraints. An increase in β factor causes the reconstruction fixed-points to collapse, effectively clustering the data at different spatial resolutions; **Middle row** Reconstruction fixed-points with local Lipschitz constraints ($r = 0.2$). As we increase the strength of local Lipschitz constraints, the fixed-points tend organize in thin manifolds connecting regions of high density. **Bottom row** Reconstruction fixed-points with global Lipschitz constraints. As we increase the strength of global Lipschitz constraints ($r = 1.0$), the fixed-points tend organize in manifolds covering regions of high density.

0. For the density $q(\mathbf{z}|\mathbf{x})$ we must also take the normalization constraint $\int d\mathbf{x}d\mathbf{z}\lambda(\mathbf{x})(q(\mathbf{z}|\mathbf{x}) - 1)$ into account:

$$\frac{\delta \mathcal{F}}{\delta g(\mathbf{z})} = \frac{1}{\sigma^2} \left(\mathbb{E}_{\rho(\mathbf{x})} \left[q(\mathbf{z}|\mathbf{x})\mathbf{x} \right] - \mathbb{E}_{\rho(\mathbf{x})} \left[q(\mathbf{z}|\mathbf{x}) \right] g(\mathbf{z}) \right) = 0,$$

$$\implies g(\mathbf{z}) = \frac{\mathbb{E}_{\rho(\mathbf{x})} \left[q(\mathbf{z}|\mathbf{x})\mathbf{x} \right]}{\mathbb{E}_{\rho(\mathbf{x})} \left[q(\mathbf{z}|\mathbf{x}) \right]} = \sum_i w_i^t(\mathbf{z}) x_i \tag{6}$$

$$\frac{\delta \mathcal{F}}{\delta q(\mathbf{z}|\mathbf{x})} = \rho(\mathbf{x}) \left[\frac{\|\mathbf{x} - g(\mathbf{z})\|^2}{2\sigma^2} + \beta \ln \frac{q(\mathbf{z}|\mathbf{x})}{\pi(\mathbf{z})} + 1 + \lambda(\mathbf{x}) \right] = 0,$$

$$\implies q(\mathbf{z}|\mathbf{x}) = \pi(\mathbf{z}) e^{-\frac{\|\mathbf{x} - g(\mathbf{z})\|^2}{2\beta\sigma^2} - \lambda(\mathbf{x})} \propto \pi(\mathbf{z}) e^{-\frac{\|\mathbf{x} - g(\mathbf{z})\|^2}{2\beta\sigma^2}}, \tag{7}$$

where $w_i^t(\mathbf{z}) = \frac{q^t(\mathbf{z}|\mathbf{x}_i)}{\sum_j q^t(\mathbf{z}|\mathbf{x}_j)}$. □

C.2 Derivation of Proposition 2

Proof. First, we note that for any given $q(\mathbf{z}|x)$, the ELBO is a convex quadratic functional of the decoder $g(\mathbf{z})$ and, for a fixed $g(\mathbf{z})$, it is a convex functional of the encoder $q(\mathbf{z}|x)$. Second, for a fixed partition Ω_i, replacing the solution $q^t(\mathbf{z}|x_i) = \pi(\mathbf{z})\mathbb{I}_{z\in\Omega_i}/\pi_i$, where $\pi_i = \mathbb{E}_\pi \left[\mathbb{I}_{z\in\Omega_i}\right]$, in the fixed-point equations, results in itself. That is,

$$w_i^t(\mathbf{z}) = \frac{q^t(\mathbf{z}|x_i)}{\sum_j q^t(\mathbf{z}|\mathbf{x}_j)} = \mathbb{I}_{z\in\Omega_i} \tag{8}$$

$$g^t(\mathbf{z}) = \sum_i w_i^t(\mathbf{z}) x_i = \sum_i x_i \mathbb{I}_{z\in\Omega_i} \tag{9}$$

$$\sigma^t = \sqrt{\mathbb{E}_{\rho(\mathbf{x})q^t(\mathbf{z}|\mathbf{x})}[\|\mathbf{x} - g^t(\mathbf{z})\|^2]} = 0 \tag{10}$$

$$q^{t+1}(\mathbf{z}|x_i) = \lim_{\sigma\to 0} \frac{\pi(\mathbf{z}) e^{-\frac{\|x_i - g^t(\mathbf{z})\|^2}{2\sigma^2}}}{c(x_i)} = \lim_{\sigma\to 0} \frac{\pi(\mathbf{z}) \sum_j e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \mathbb{I}_{z\in\Omega_j}}{\int dz' \pi(\mathbf{z}') \sum_k e^{-\frac{\|x_i - x_k\|^2}{2\sigma^2}} \mathbb{I}_{z\in\Omega_k}} = \frac{1}{\pi_i} \pi(\mathbf{z}) \mathbb{I}_{z\in\Omega_i}. \tag{11}$$

Therefore, $q^t(\mathbf{z}|x_i) = \pi(\mathbf{z})\mathbb{I}_{z\in\Omega_i}/\pi_i$ is a fixed-point in the family of densities constrained by the partition Ω_i. We observe that the negative ELBO reduces to the expected KL term only, $\mathbb{E}_p[\mathrm{KL}(q;\pi)] = \frac{1}{n}\sum_j \int dz \frac{\pi(\mathbf{z})\mathbb{I}_{z\in\Omega_j}}{\pi_j}(-\ln\pi_j) = -\frac{1}{n}\sum_j \ln\pi_j$. We can now optimize the partition to further maximize the ELBO. This results in $\pi_j = 1/n$. That is, the tiles Ω_i must be equiprobable. At this point, we have $\mathbb{E}_p[\mathrm{KL}(q;\pi)] = \ln n$. □

C.3 Derivation of Proposition 3

Proof. From (3) we have that $g(\mathbf{z}) = \sum_i w_i^t(\mathbf{z})\mathbf{x}_i = \sum_i \mathbf{x}_i \frac{q(\mathbf{z}|x_i)}{\sum_j q(\mathbf{z}|\mathbf{x}_j)}$. Substituting $q(\mathbf{z}|x_i) = \pi(\mathbf{z})\mathbb{I}_{z\in\Omega_i}/\pi_i$ we have,

$$g(\mathbf{z}) = \sum_i \mathbf{x}_i \frac{q(\mathbf{z}|x_i)}{\sum_j q(\mathbf{z}|\mathbf{x}_j)} = \sum_i \mathbf{x}_i \frac{\mathbb{I}_{z\in\Omega_i}/\pi_i}{\sum_j \mathbb{I}_{z\in\Omega_j}/\pi_j} = \sum_{i|\mathbf{z}\in\Omega_i} \mathbf{x}_i \frac{1/\pi_i}{\sum_{j|\mathbf{z}\in\Omega_j} 1/\pi_j}, \tag{12}$$

where $\pi_i = \mathbb{E}_\pi\left[\mathbb{I}_{z\in\Omega_i}\right]$. □

C.4 Derivation of Proposition 4

Proof. The expression for the generator can be obtained by substitution and algebraic simplification using the fact that ϕ_a form an orthogonal basis and that $\phi_a(\mathbf{z}) \in \{0, 1\}$:

$$g(\mathbf{z}) = \sum_{ib} \frac{q^t(\mathbf{z}|\mathbf{x}_i)}{\sum_j q^t(\mathbf{z}|\mathbf{x}_j)}\mathbf{x}_i = \sum_{ia} \frac{m_{ia}\phi_a(\mathbf{z})}{\sum_{jb} m_{jb}\phi_b(\mathbf{z})}\mathbf{x}_i = \sum_{ib} \frac{m_{ib}}{\sum_j m_{jb}}\phi_b(\mathbf{z})\mathbf{x}_i = \psi^T\phi(\mathbf{z}), \tag{13}$$

with $\psi_b = \sum_i \frac{m_{ib}}{\sum_j m_{jb}}\mathbf{x}_i$. Similarly, we can compute the fixed point equations for ψ_a by substitution on equations (2) and (3),

$$q_i^{t+1}(\mathbf{z}) = \pi(\mathbf{z})\sum_b \frac{e^{-\frac{\|x_i-\psi_b^t\|^2}{2\beta}}}{c_i^t}\phi_b(\mathbf{z}) = \pi(\mathbf{z})\sum_a m_{ia}^{t+1}\phi_a(\mathbf{z}) \tag{14}$$

$$\psi_b^{t+1} = \sum_i x_i \frac{m_{ib}^{t+1}}{\sum_j m_{jb}^{t+1}} = \frac{\sum_i x_i \frac{e^{-\frac{\|x_i-\psi_b^t\|^2}{2\beta}}}{c_i^t}}{\sum_i \frac{e^{-\frac{\|x_i-\psi_b^t\|^2}{2\beta}}}{c_i^t}}, \tag{15}$$

where $c_i = \sum_b e^{-\frac{\|x_i-\psi_b^t\|^2}{2\beta}}\pi_b$ and $m_{ib}^{t+1} = \frac{e^{-\frac{\|x_i-\psi_b^t\|^2}{2\beta}}}{c_i^t}$. If the initial reconstruction vectors ψ_a^t are in the convex-hull of the training data-points, then equations (15) will map them to another set of points ψ_a^{t+1} in the convex-hull of the training data. Since, these equations are also smooth with respect to ψ, they are guaranteed to converge as a consequence of the fixed-point theorem [20]. Importantly, equation (15) corresponds to the fixed point iterations for computing the pre-image (reconstructions) of Kernel-PCA but using a normalized Gaussian kernel. □

C.5 Derivation of Proposition 5

Proof. From Proposition 4, we have seen that the reconstruction vectors ψ_a of a high-capacity β-VAE will converge to a set of m fixed-points. This means that ψ will map the set basis-functions to a smaller subset of points. As a consequence, all the latent-vectors falling in the support of these basis elements will also map to the same reconstruction and for a fixed $\mathbf{x}$, the function $H(\mathbf{x}, \mathbf{z})$ will only have m possible distinct values, which we can enumerate as H_{ia}. If we enumerate all distinct supports as Ω_a, then from Equation (2) we have that $q_i(\mathbf{z}) = \pi(\mathbf{z})\sum_a \frac{e^{-\frac{H_{ia}}{\beta}}}{\sum_b e^{-\frac{H_{ib}}{\beta}}\gamma_b}\mathbb{I}_{\mathbf{z}\in\Omega_a}$. Replacing $q_i(\mathbf{z})$ in Equation (1) and maximizing with respect to γ_a results in the condition $\sum_i \frac{e^{-\frac{H_{ia}}{\beta}}}{\sum_b e^{-\frac{H_{ib}}{\beta}}\gamma_b} = n$. □

C.6 Derivation of Proposition 6

Proof. We first write the relevant terms of the augmented Lagrangian $\mathcal{L}_{\boldsymbol{\lambda},\Omega}$ in the basis $\phi(\mathbf{z})$ from Proposition 4,

$$\begin{aligned}
\mathcal{L}_{\boldsymbol{\lambda},\Omega} &= \mathbb{E}_{\rho(\mathbf{x})}\left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\frac{\|\mathbf{x}-\psi^T\phi(\mathbf{z})\|^2}{2\sigma^2}\right]\right] + \frac{1}{2n}\sum_{a,b}\tilde{\Omega}_{ab}\left[C_{ab}\|\psi_a-\psi_b\|^2-1\right] + \text{cst} && (16)\\
&= \mathbb{E}_{\rho(\mathbf{x})}\left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\phi_a(\mathbf{z})\right]\frac{\|\mathbf{x}-\psi_a\|^2}{2\sigma^2}\right] + \frac{1}{2n}\sum_{a,b}\tilde{\Omega}_{ab}\left[C_{ab}\|\psi_a-\psi_b\|^2-1\right] + \text{cst} && (17)\\
&= \frac{1}{n}\sum_{ia} m_{ia}\pi_a\frac{\|\mathbf{x}_i-\psi_a\|^2}{2\sigma^2} + \frac{1}{2n}\sum_{a,b}\tilde{\Omega}_{ab}C_{ab}\|\psi_a-\psi_b\|^2 + \text{cst}, && (18)
\end{aligned}$$

where cst are terms that do not depend on ψ for a fixed q. Solving $\frac{\partial\mathcal{L}_{\boldsymbol{\lambda},\Omega}}{\partial\psi_a} = 0$ with respect to ψ results in

$$\begin{aligned}
\frac{\partial\mathcal{L}_{\boldsymbol{\lambda},\Omega}}{\partial\psi_a} &= \frac{\pi_a\sum_i m_{ia}}{\sigma^2}\left[\sum_i \mathbf{x}_i\frac{m_{ia}}{\sum_i m_{ia}} - \psi_a + \frac{\sigma^2}{\pi_a\sum_i m_{ia}}\sum_b\tilde{\Omega}_{ab}C_{ab}(\psi_a-\psi_b)\right] = 0 && (19)\\
\psi_b &= \sum_{i,a}\mathbf{x}_i\frac{m_{ia}}{\sum_i m_{ia}}P_{ab}, && (20)
\end{aligned}$$

where $P = [\mathbb{I} - \text{diag}(\frac{n\sigma^2}{1^T m \bigodot \pi})(\text{diag}(1^T(\tilde{\Omega}\bigodot C)) - \tilde{\Omega}\bigodot C)]^{-1}$. □