WHEN DO VARIATIONAL AUTOENCODERS KNOW WHAT THEY DON'T KNOW?

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, the ability of deep generative models to detect outliers has been called into question because of the demonstration that they frequently assign higher probability density to samples from completely different data sets than were used for training. For example, a model trained on CIFAR-10 may counter-intuitively attribute higher likelihood to samples obtained from SVHN. In this work, we closely examine this phenomena in the specific context of variational autoencoders, a commonly-used approach for anomaly detection. In particular, we demonstrate that VAEs, when appropriately designed and trained, are in fact often proficient in differentiating inlier and outlier distributions, e.g., FashionMNIST vs MNIST, CIFAR-10 vs SVHN and CelebA. We describe various mechanisms that mitigate this capability, including the paradoxical necessity of large or unbounded gradients, which have sometimes been observed to occur during training of VAE models.

1 INTRODUCTION

Suppose we have access to continuous variables $\boldsymbol{x} \in \boldsymbol{\chi}$ that are assumed to be drawn from ground-truth measure μ_{gt}. This measure assigns probability mass $\mu_{gt}(d\boldsymbol{x})$ to the infinitesimal $d\boldsymbol{x}$ residing within $\boldsymbol{\chi} \subseteq \mathbb{R}^d$ such that we have $\int_{\boldsymbol{\chi}} \mu_{gt}(d\boldsymbol{x}) = 1$. This formalism allows us to consider data that may lie on or near an r-dimensional manifold embedded in $\mathbb{R}^d$ (implying $r \leq d$), capturing the notion of low-dimensional structure relative to the high-dimensional ambient space.

Because of the possibility of low-dimensional latent structure, it is common to approximate the unknown ground-truth measure via a density model parameterized as $p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$. In this expression θ are trainable parameters and $\boldsymbol{z} \in \mathbb{R}^\kappa$ serves as a low-dimensional latent representation, with fixed prior $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$ and ideally $\kappa \approx r$. If some θ^* were available such that $\int_A p_{\theta^*}(\boldsymbol{x})d\boldsymbol{x} \approx \int_A \mu_{gt}(d\boldsymbol{x})$ for any measurable $A \subseteq \boldsymbol{\chi}$, then the model would adequately reflect the intrinsic underlying distribution.

In practice then, we could generate new samples from $p_{\theta^*}(\boldsymbol{x})$ by drawing $\boldsymbol{z}^{new} \sim \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$ and then $\boldsymbol{x}^{new} \sim p_{\theta^*}(\boldsymbol{x}|\boldsymbol{z}^{new})$. Alternatively, we could evaluate the negative log-likelihood (NLL) of a test sample $\boldsymbol{x}^{test}$, meaning $-\log p_{\theta^*}(\boldsymbol{x}^{test})$, which could in turn be applied to various tasks such as outlier/anomaly detection. In terms of the latter, a low NLL should ostensibly be associated with inliers, while high values would distinguish outliers.

Of course we will generally not know in advance the value of θ^*, but in principle we might consider minimizing $-\log p_\theta(\boldsymbol{x})$ averaged across a set of training samples $\{\boldsymbol{x}^{(i)}\}_{i=1}^n$ drawn from μ_{gt}, i.e., minimize $\frac{1}{n}\sum_i -\log\left[p_\theta\left(\boldsymbol{x}^{(i)}\right)\right] \approx \int -\log\left[p_\theta(\boldsymbol{x})\right]\mu_{gt}(d\boldsymbol{x})$ over θ. Unfortunately though, the marginalization required to produce $p_\theta\left(\boldsymbol{x}^{(i)}\right)$ is generally intractable for models of sufficient representational power. To circumvent this issue, the variational autoencoder (VAE) (Kingma & Welling, 2014; Rezende et al., 2014) instead optimizes the tractable variational bound $\mathcal{L}(\theta, \phi) \triangleq$

$$\frac{1}{n}\sum_{i=1}^{n}\left\{-\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[\log p_\theta\left(\boldsymbol{x}^{(i)}|\boldsymbol{z}\right)\right] + \mathbb{KL}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})||p(\boldsymbol{z})\right]\right\} \geq \frac{1}{n}\sum_{i-1}^{n} -\log\left[p_\theta\left(\boldsymbol{x}^{(i)}\right)\right]. \tag{1}$$

Here $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ represents a tractable variational approximation to $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ with additional parameters ϕ governing the tightness of the bound. It is commonly referred to as an *encoder* distribution since

it quantifies the mapping from $\boldsymbol{x}$ to the latent code $\boldsymbol{z}$. For analogous reasons, $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ is labeled as the *decoder* distribution. When combined, the data-dependent factor $-\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta\left(\boldsymbol{x}|\boldsymbol{z}\right)\right]$ can be viewed as instantiating a form of stochastic autoencoder (AE) structure, which attempts to assign high probability to accurate reconstructions of each $\boldsymbol{x}$; if $q_\phi\left(\boldsymbol{z}|\boldsymbol{x}\right)$ is Dirac delta function, then a regular deterministic AE emerges with loss dictated by the decoder negative log-likelihood $-\log p_\theta(\boldsymbol{x}|\boldsymbol{z})$. Beyond this, $\mathbb{KL}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right]$ serves as a regularization factor that pushes the encoder distribution towards the prior. The bound (1) can be minimized over $\{\theta, \phi\}$ using SGD and a simple reparameterization trick (Kingma & Welling, 2014; Rezende et al., 2014).

The latter requires that we assume specific function forms for the encoder and decoder distributions. In this regard, it is common to select $q_\phi\left(\boldsymbol{z}|\boldsymbol{x}\right) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_z, \text{diag}[\boldsymbol{\sigma}_z]^2)$, where the Gaussian moment vectors $\boldsymbol{\mu}_z$ and $\boldsymbol{\sigma}_z$ are functions of model parameters ϕ and the random variable $\boldsymbol{x}$, i.e., $\boldsymbol{\mu}_z \equiv \boldsymbol{\mu}_z\left(\boldsymbol{x};\phi\right)$, and $\boldsymbol{\sigma}_z \equiv \boldsymbol{\sigma}_z\left(\boldsymbol{x};\phi\right)$. Similarly, for continuous data the decoder model is conventionally parameterized as $p_\theta\left(\boldsymbol{x}|\boldsymbol{z}\right) = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_x, \gamma\boldsymbol{I})$, with mean defined analogously as $\boldsymbol{\mu}_x \equiv \boldsymbol{\mu}_x\left(\boldsymbol{z};\theta\right)$ and scalar, trainable variance parameter $\gamma > 0$. The functions $\boldsymbol{\mu}_z\left(\boldsymbol{x};\phi\right)$, $\boldsymbol{\sigma}_z\left(\boldsymbol{x};\phi\right)$, and $\boldsymbol{\mu}_x\left(\boldsymbol{z};\theta\right)$ are all instantiated using deep neural network layers.

Although VAEs can be applied to a variety of practical problems (Li & She, 2017; Schott et al., 2018; Walker et al., 2016), our focus herein will be on accurately detecting anomalous out-of-distribution samples. This endeavor serves two interrelated purposes. First, isolating and removing outliers represents an important component of many real-world machine learning pipelines in and of itself. For example, as pointed out in Nalisnick et al. (2019), generative models or related have frequently been proposed as a means of removing problematic outlier samples that may prove difficult for downstream classifiers (Bishop, 1994).

Secondly though, the ability of deep generative models to successfully detect samples completely different from the training set has recently been called into question by multiple studies (Choi & Jang, 2018; Hendrycks et al., 2018; Nalisnick et al., 2019; Shafaei et al., 2018; Škvára et al., 2018). In particular, when trained on CIFAR-10 data, VAEs, as well as alternative autoregressive (van den Oord et al., 2016) and flow-based generative models (Kingma & Dhariwal, 2018), were all shown to assign a lower NLL (i.e., higher density) to SVHN samples than the actual training set (Nalisnick et al., 2019). This obviously undercuts the value of these models in detecting outlier distributions, but it is also quite counter-intuitive given that SVHN house number images are visually different from those contained within CIFAR-10 (e.g., cats, dogs, cars, etc.) and should therefore ostensibly be easy to differentiate. Some compelling, plausible arguments also exist as to how this might happen within the specific context of invertible flow-based models (Nalisnick et al., 2019).

We attempt to expand on this work by providing complementary analysis and perspective specifically regarding the VAE within the context of robust detection of out-of-distribution data. Beginning in Section 2, we discuss the importance of closely aligning the support of $p_\theta(\boldsymbol{x})$ to the ground-truth μ_{gt} when low-dimensional manifold structure is present as is typical with natural image data. Section 3 then analyzes VAE capabilities in this regard as the dimensionality of $\boldsymbol{z}$ and the capacity of $\boldsymbol{\mu}_x\left(\boldsymbol{z};\theta\right)$ and therefore $p_\theta(\boldsymbol{x})$ are varied, differentiating likely success and failure regimes. We also formally demonstrate the underappreciated yet unavoidable emergence of unbounded gradients when training a broad class of autoencoder-based models, VAE or otherwise, designed to match ground-truth manifolds. Finally, we corroborate the analysis from Section 3 with a series of experiments in Section 4, demonstrating that VAE strengths can in fact often be leveraged to differentiate samples from an outlier distribution, i.e., know what they don't know. Overall, the results we present contribute to a better understanding of relevant capabilities such that VAE models are not underutilized within common practical application domains such as anomaly detection (An & Cho, 2015; Xu et al., 2018).

2 Proper Alignment of Inlier Models with the Data Manifold

In this section we argue that the *support* of $p_\theta(\boldsymbol{x})$, meaning the set $\mathcal{S}_x \triangleq \{\boldsymbol{x} : p_\theta(\boldsymbol{x}) > 0\}$ is critical in differentiating outliers.[1] To begin, consider the illustrative example where μ_{gt} assigns all of its probability mass uniformly to the perimeter of a circle centered at zero. If model capacity is limited such that $p_\theta(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ assumed to be fixed parameters for present

[1]The same basic intuitions would also hold if we relax this definition to $\mathcal{S}_x \triangleq \{\boldsymbol{x} : p_\theta(\boldsymbol{x}) > \epsilon\}$, for ϵ sufficiently small.

purposes, then the maximum likelihood values are $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$. But now suppose there exists an arbitrary outlier distribution with nonzero support *within* the aforementioned circle. Any such outlier sample will necessarily have a higher probability under the stated Gaussian model than any of the true inlier data spread around the circumference of the circle. The problem is that the assumed inlier distribution does not have sufficient capacity to learn the ground-truth support pattern of the data, and is therefore susceptible to inadvertently assigning higher probability to regions that do not actually contain inliers (in this case, the center of the circle).

In differentiating outliers then, we must ultimately navigate between two extremes. First, as motivated above, we should supply our model with sufficient capacity such that probability mass can be confined to regions of $\boldsymbol{x}$-space that closely surround the training data manifold, and avoid leaving probability mass in regions vulnerable to outliers. But secondly, we must also ensure that we do not grossly include superfluous capacity such that we simply memorize the data by parking a delta function with infinite density at each training sample. In this case, both outliers and inlier test samples would occupy areas of zero density and would therefore be indistinguishable. Fortunately, within these extremes there exists ample opportunity for an effective balance via an appropriate modeling framework.

That being said, flow-based likelihood models have no mechanism for explicitly reflecting low-dimensional support structure because the assumed class of distributions must be homeomorphic to $\mathbb{R}^d$. More precisely, these methods transform a latent code $\boldsymbol{z} \in \mathbb{R}^d$ drawn from $\mathcal{N}(\boldsymbol{z}|\mathbf{0}, \boldsymbol{I})$ to a more complex distribution $p_\theta(\boldsymbol{x})$ through a series of invertible transformations (Dinh et al., 2016; Kingma & Dhariwal, 2018). However, there is no invertible mapping between $\mathcal{N}(\boldsymbol{z}|\mathbf{0}, \boldsymbol{I})$ and μ_{gt} if the manifold $\boldsymbol{\chi}$ has zero Lebesgue measure in $\mathbb{R}^d$ (e.g., low-dimensional manifold structure exists). This is not to say that flow-based models can never at least approximate a low-dimensional manifold, but the basic parameterization is counter to this assumption, and there is no clear indicator within the model of which latent degrees-of-freedom are superfluous. In this regard though, the VAE is decidedly different as discussed next.

3 Implications for the VAE

With a typical Gaussian VAE model of continuous data as we have adopted, the support of $p_\theta(\boldsymbol{x})$ is explicitly controlled by two interpretable factors: (i) the variance γ of the decoder distribution $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, and (ii) the number of dimensions within $\boldsymbol{z}$ that contain useful information about $\boldsymbol{x}$. Both of these complementary factors can be monitored to evaluate the intrinsic data dimensionality. For example, if during training $\gamma \rightarrow 0$, by design $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ collapses to a deterministic mapping $\boldsymbol{\mu}_x(\boldsymbol{z}; \theta)$ from $\mathbb{R}^\kappa$ to $\mathbb{R}^d$. The resulting probability mass assigned by the model will thus be caged within a κ-dimensional manifold assuming $\kappa < d$. But suppose also that the ground-truth data manifold is such that $r < \kappa$, implying that all degrees-of-freedom within $\boldsymbol{z}$ are not needed. If $\mu_z(\boldsymbol{x}^{(i)}; \phi)_k^2 / \sigma_z(\boldsymbol{x}^{(i)}; \phi)_k^2 \rightarrow 0$ during training, then no information pertaining to $\boldsymbol{x}^{(i)}$ is preserved by the k-th latent dimension of $\boldsymbol{z}$.

Per these two considerations, it has been demonstrated in Dai & Wipf (2019) that indeed, when granted sufficient capacity, Gaussian VAEs will produce near perfect reconstructions of the training data, while pushing $\gamma \rightarrow 0$ and $\mu_z(\boldsymbol{x}^{(i)}; \phi)_k^2 / \sigma_z(\boldsymbol{x}^{(i)}; \phi)_k^2 \rightarrow 0$ for useless dimensions, both of which allow the model to tighten its probability assignment to a minimal support pattern containing the training data. We conjecture that this mechanism should be helpful in allowing the VAE to isolate outliers.

Figure 1 illustrates this claim via a series of hypothetical scenarios as the VAE capacity and latent dimensionality κ are varied. In this narrow context, capacity refers to the complexity of the decoder mean network $\boldsymbol{\mu}_x(\boldsymbol{z}; \theta)$ which controls the flexibility of $p_\theta(\boldsymbol{x})$; we assume that the encoder network maintains sufficient complexity to produce a reasonably tight variational bound given the decoder. The predicted behavior can be described and contrasted in the following three regimes: (i) insufficient capacity, (ii) suitable capacity, and (iii) excessive capacity. In each case the impact of the latent dimensionality will be different as described next.

3.1 Analyses with Respect to Varying Network Architecture

Insufficient Network Capacity: If the latent dimension is small relative to the data manifold, meaning $\dim[\boldsymbol{z}] = \kappa < \dim[\boldsymbol{\chi}] = r$, then there is no way for the VAE to produce small reconstruction errors with a low-capacity decoder mean $\boldsymbol{\mu}_x(\boldsymbol{z}; \theta)$. This ensures that γ will necessarily become large

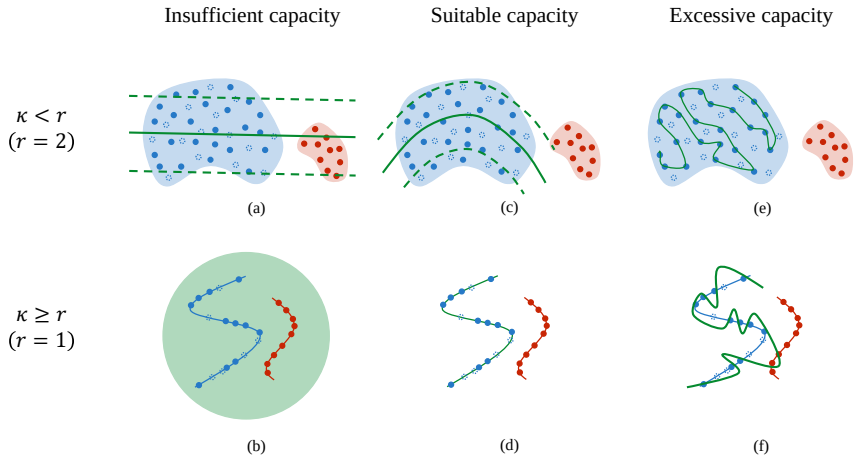


Figure 1: Illustration of VAE modeling behavior as $\dim[\boldsymbol{z}] = \kappa$ and the capacity of $\boldsymbol{\mu}_x(\boldsymbol{z}; \theta^*)$ are varied. Blue dots and dashed blue circles represent training and testing samples located within the ground-truth data manifold shown in blue. Analogously, red dots are outliers. The learned manifold as produced by $\boldsymbol{\mu}_x(\boldsymbol{z}; \theta^*)$ is shown in green, with dashed green lines denoting spread around the manifold from $\gamma \gg 0$ in subplots (a) and (b). In the top row, the support of μ_{gt} is a 2D blob (i.e., $r = 2$); for the bottom row it is 1D curve (i.e., $r = 1$). See Section 3.1 for a detailed explanation.

during training given that, with all other parameters fixed, the optimal value of γ is related to the reconstruction fidelity via[2]

$$\gamma = -\tfrac{1}{nd} \sum_{i=1}^{n} \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})} \left[\log p_\theta \left(\boldsymbol{x}^{(i)} | \boldsymbol{z} \right) \right] = \tfrac{1}{nd} \sum_{i=1}^{n} \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})} \left[\left\| \boldsymbol{x}^{(i)} - \boldsymbol{\mu}_x(\boldsymbol{z}; \theta) \right\|_2^2 \right]. \quad (2)$$

Consequently, the resulting density estimate

$$p_{\theta^*}(\boldsymbol{x}) = \int p_{\theta^*}(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} = \int \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_x(\boldsymbol{z}; \theta^*), \gamma \boldsymbol{I}) \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I}) d\boldsymbol{z}, \quad (3)$$

with optimized parameters θ^*, will be spread broadly around $\boldsymbol{\mu}_x(\boldsymbol{z}; \theta^*)$ between the dashed green lines in Figure 1(a). It then follows that outlier samples may well receive high likelihood or a low NLL. Additionally, if we force $\kappa = 0$, then $\boldsymbol{\mu}_x(\boldsymbol{z}; \theta^*)$ becomes a constant and the situation mirrors the toy example described in Section 2.

In contrast, with $\kappa > r$, the VAE with limited capacity can just assign density to a simpler but higher-dimensional approximate manifold or subspace that subsumes the more complex, yet lower-dimensional ground-truth manifold. This allows for low reconstruction errors and $\gamma \to 0$. However, as shown in Figure 1(b), the resulting approximate density will also be incapable of properly differentiating outliers. So in the end, without adequate capacity, the VAE is likely to fail in consistently distinguishing outliers regardless of κ.

Suitable Network Capacity: When the network capacity is well-calibrated to the complexity of the ground-truth data manifold, the situation may not actually change dramatically if κ specifically is still too small. For example, as shown in Figure 1(c), with $\kappa < r$, the decoder mean $\boldsymbol{\mu}_x(\boldsymbol{z}; \theta^*)$ can better reflect the shape of the inlier data distribution than was possible in 1(a). However, without sufficient degrees-of-freedom within the latent code, the reconstruction errors will still be high, and γ will remain large. Consequently, $p_{\theta^*}(\boldsymbol{x})$ may not be fully segregated from regions that contain outliers.

But with $\kappa \geq r$, the situation is much more favorable. The available network capacity can be leveraged to produce near perfect reconstructions using a minimal number of latent degrees-of-freedom per the arguments from Dai & Wipf (2019). This drives $\gamma \to 0$, restricting the assigned probability mass to a narrow manifold surrounding the training data. This mitigates the risk of either

[2]This relationship is obtained by simply differentiating the VAE cost with respect to γ, equating to zero, and rearranging terms.

overfitting or co-mingling with the outlier distribution as shown in Figure 1(d). The low NLL from inlier test samples will therefore be easily distinguishable from the much higher values produced by outliers.

Excessive Network Capacity With excessive model capacity, the VAE can overfit to the training data. Even when κ is much smaller than r, it is still possible to perfectly fit the training samples and drive $\gamma \rightarrow 0$ with a sufficiently complex $\boldsymbol{\mu}_x(\boldsymbol{z};\theta)$ as depicted in Figure 1(e). Although the NLL for the training set will be low, on novel test samples the NLL for inliers could be higher than for outliers, making proper anomaly detection problematic. Likewise, when we have $\kappa \geq r$, overfitting leads to inlier likelihood assignments that deviate from the ground-truth manifold with similarly deleterious effects; see Figure 1(f). Note that while the VAE has the ability to automatically prune a superfluous latent dimension k by setting $\mu_z(\boldsymbol{x}^{(i)};\phi)_k^2/\sigma_z(\boldsymbol{x}^{(i)};\phi)_k^2 \rightarrow 0$, it does *not* have any mechanism for regularizing an excessively complex decoder mean function $\boldsymbol{\mu}_x(\boldsymbol{z};\theta)$ (Dai et al., 2018).

3.2 Practical Considerations

The analysis thus far suggests that learning γ, instead of choosing a fixed value such as $\gamma = 1$ (as is frequently done in practice), while choosing $\kappa \geq r$ is a useful prescription for applying the VAE to detecting out-of-distribution anomalies. And at least in principle κ need not be carefully calibrated so long as it is sufficiently large because excessive latent dimensions can be pruned. As for the network capacity/depth, it can be increased until there is a significant gap between train and test NLL values on the inlier data (access to unknown outliers is not needed to avoid over-fitting).

Beyond this, we should also mention the following caveats. First, the discussion in Section 3.1 implicitly assumed that VAE training was more-or-less successful, meaning that over-regularized, degenerate local minima were avoided. While obviously this cannot be always guaranteed, several steps have been proposed in the literature to help ensure favorable training conditions (Bowman et al., 2015; Cai et al., 2017; Dieng et al., 2018; Sønderby et al., 2016). We did not however require any of these methods for the results reported in Section 4. Secondly, constraining probability mass to a narrow manifold necessarily involves $\gamma \rightarrow 0$ as well as partitioning $\mu_z(\boldsymbol{x}^{(i)};\phi)_k^2/\sigma_z(\boldsymbol{x}^{(i)};\phi)_k^2 \rightarrow 0$ for useless dimensions and $\mu_z(\boldsymbol{x}^{(i)};\phi)_k^2/\sigma_z(\boldsymbol{x}^{(i)};\phi)_k^2 \rightarrow \infty$ along informative ones. We may therefore expect to encounter large or unbounded gradients during training of the VAE cost. This can occur within the data term from (1) because of $\gamma \rightarrow 0$, and within the KL term because of $\sigma_z(\boldsymbol{x}^{(i)};\phi)_k^2 \rightarrow 0$ along an active dimension. This issue has previously been raised as a potential concern, but we will now reframe such gradients as more of a necessary risk.

3.3 Unbounded Gradients as a Necessary Risk

As motivated in Dai & Wipf (2019) and explicitly contextualized with respect to outlier detection in Section 3.1, the VAE is capable of providing accurate reconstructions of the training data using a minimal number of active latent dimensions containing information about each $\boldsymbol{x}^{(i)}$, a construct that we will henceforth refer to as an *optimal sparse reconstruction*. However, an unintended consequence of this phenomena is the potential for divergent gradients as discussed in Section 3.2. This then naturally begs the question: Could we not just simply train a deterministic autoencoder (AE) to obtain such optimal sparse reconstructions and avoid this issue altogether? Interestingly, the answer turns out to be unequivocally no, at least in the sense formalized by the following analysis.

Consider the constrained objective function $\mathcal{L}_h(\theta,\phi) \triangleq$

$$h\left(\frac{1}{dn}\sum_{i=1}^{n}\left\|\boldsymbol{x}^{(i)}-\boldsymbol{\mu}_x\left(\boldsymbol{z}^{(i)};\theta\right)\right\|_2^2\right)+\frac{1}{d}\sum_{k=1}^{\kappa}h\left(\frac{1}{n}\left\|\boldsymbol{z}_k\right\|_2^2\right), \;\; \text{s.t. } \boldsymbol{z}^{(i)}=\boldsymbol{\mu}_z\left(\boldsymbol{x}^{(i)};\phi\right) \;\forall i,\; \theta\in\Theta, \tag{4}$$

where $\boldsymbol{Z} \triangleq \{\boldsymbol{z}^{(i)}\}_{i=1}^n \in \mathbb{R}^{\kappa\times n}$ and $\boldsymbol{z}_k$ denotes the k-th row of $\boldsymbol{Z}$. This expression can be viewed as characterizing a typical regularized AE with a generic penalty function $h : \mathbb{R}^+ \rightarrow \mathbb{R}$ on the norm across training samples of each latent dimension. The multipliers $1/n$, $1/d$, and $1/(dn)$ ensure a form of proportional regularization as within energy functions composed of multiple penalty factors of varying dimension designed to favor sparsity (Wipf & Wu, 2012). The square-root Lasso can be viewed as a special case of this strategy that emerges when h is a square-root function (Belloni et al., 2011). We adopt this formalism to avoid distracting complications from tunable trade-off parameters; however, our central conclusions still hold even when such a parameter is introduced. And finally, the constraint $\theta \in \Theta$ is included to prevent the trivial solution $\boldsymbol{Z} \rightarrow \boldsymbol{0}$, which could

occur if each $\boldsymbol{z}^{(i)}$ is pushed to zero while $\boldsymbol{\mu}_x$ includes an unconstrained compensatory factor that grows towards infinity such that the error $\left\|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_x\left(\boldsymbol{z}^{(i)};\theta\right)\right\|_2$ can still be minimized to zero. Any regularized AE must include such constraints to avoid trivial solutions, or else additional penalty terms on θ that serve a similar purpose.

Given a generic AE architecture as in (4), it is natural to examine what possible functions h are such that any global minimum of $\mathcal{L}_h(\theta,\phi)$ is guaranteed to produce an optimal sparse representation. This can be addressed as follows:

Theorem 1 *Assume the constraint $\theta \in \Theta$ and data $\boldsymbol{X} = \{\boldsymbol{x}^{(i)}\}_{i=1}^n \in \mathbb{R}^{d\times n}$ are such that to achieve $\boldsymbol{x}^{(i)} = \boldsymbol{\mu}_x\left(\boldsymbol{z}^{(i)};\theta\right)\ \forall i$ (i.e., perfect reconstruction) requires that $\|\boldsymbol{z}_k\|_2 > 0$ for at least $r < d$ rows of $\boldsymbol{Z}$. Then to guarantee that minimization of $\mathcal{L}_h(\theta,\phi)$ achieves zero reconstruction error using at most r nonzero rows of $\boldsymbol{Z}$ (i.e., active dimensions), h must have an unbounded gradient around zero.*

The proof is deferred to the supplementary. Note that a similar result can be obtained by replacing the reconstruction penalty with the additional constraint $\sum_{i=1}^n \left\|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_x\left(\boldsymbol{z}^{(i)};\theta\right)\right\|_2^2 = 0$, in which case no trade-off parameter, fixed or otherwise, need be included. We also emphasize that Theorem 1 effectively implies that, to guarantee every global minima corresponds with an optimal sparse reconstruction per our definition, the constituent penalty functions must have an infinite gradient around zero. Given that we may readily introduce arbitrary scalings and translations, this condition is tantamount to requiring penalty functions with an energy gap that is unbounded about zero. For example, the selection $h(u) = \mathcal{I}\left[u > 0\right]$, i.e., an indicator function that equals zero if $u = 0$ and one for all $u > 0$, will guarantee that any global minimum of (4) produces an optimal sparse reconstruction under the stated conditions. However, given that $\mathcal{I}\left[u > 0\right] \equiv \lim_{p\to 0} u^p$ and $\lim_{p\to 0} \frac{1}{p}\left(u^p - 1\right) = \log u$, we see that an unbounded log function can essentially achieve the same result in the limit.

The VAE can be interpreted as a form of stochastic AE, with subtle regularization effects introduced via the interplay between the reconstruction and KL terms. A number of recent works have mentioned that if a flexible decoder variance parameter γ is included within a Gaussian VAE, then the optimal value may converge to zero, resulting in infinite gradients and potential instabilities (Dai & Wipf, 2019; Mattei & Frellsen, 2018; Takahashi et al., 2018). While unbounded gradients may indeed be troublesome from an optimization perspective, based on the analysis of this section, we frame such gradients as a necessary component of any model that attempts to produce optimal sparse reconstructions. In this regard, it has been argued that as $\gamma \to 0$, the VAE can achieve zero reconstruction error at the global optimum by selectively pushing $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ towards a degenerate Gaussian, with zero variance along the minimal number of directions needed for reconstructing $\boldsymbol{x}$, and unit variance elsewhere so as to reduce the KL regularization factor (Dai & Wipf, 2019). This is exactly a stochastic version of an optimal sparse reconstruction, which can be exploited for outlier detection per the discussion from Section 3.1. See the supplementary for more details on this topic, including related empirical results.

4 Experimental Validation

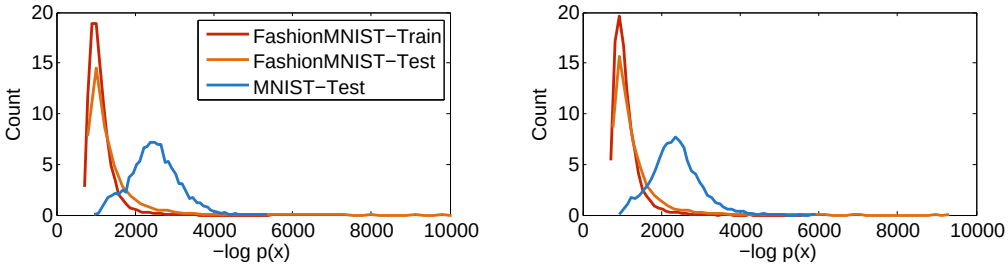
In this section we empirically corroborate the analysis from Section 3, demonstrating that in the predicted operating regimes, the VAE can indeed differentiate inlier and outlier distributions. For this purpose, we train a variety VAE models differing in latent dimensionality and network capacity in an attempt to isolate the ground-truth inlier manifold. We adopt the network structure from Bińkowski et al. (2018) as our baseline and include γ as a trainable parameter for reasons given in Section 3. To manipulate capacity we multiply the number of channels in all the layers (except the final encoder layer producing $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$) by a factor of $\alpha \in [\frac{1}{4}, \frac{1}{2}, 1, 2, 3, 4]$. We apply the notation $\times\alpha$ to represent the network capacity across varying test conditions. For all the experiments, we use the Adam optimizer and train the network for 200K iterations with a fixed learning rate of 10^{-4}. Please see the supplementary for further details, as well as additional analysis and experiments.

4.1 Evaluations across Varying Latent Dimensionality and Network Capacity

We employ the basic experimental paradigm from Nalisnick et al. (2019), training models on a given inlier set and then comparing evaluation metrics applied to both inlier train/test samples and distinct outlier samples. Consistent with Nalisnick et al. (2019) and convention elsewhere, we use

	Glow	16	32	64	128	256	512
FashionMNIST-Train	2.902	2.375	2.208	2.084	2.009	2.054	2.020
FashionMNIST-Test	2.958	2.805	2.690	2.562	2.424	2.397	2.301
MNIST-Test	1.833	9.598	8.618	6.294	4.958	4.578	4.351
γ	–	0.0055	0.0040	0.0030	0.0024	0.0023	0.0024

Table 1: BPD values for VAEs trained on FashionMNIST with capacity $\times 1$ as κ is varied from 16 to 512. When κ increases, the BPD saturates while robustly differentiating inliers and outliers.



(a) Capacity$\times 1$, $\kappa = 256$ (b) Capacity$\times 1$, $\kappa = 512$

Figure 2: Stable histogram of VAE NLL values for FashionMNIST vs MNIST for different κ.

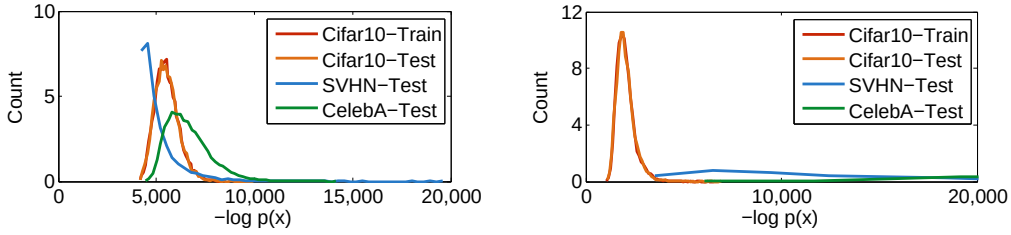
the bits-per-dimension (BPD) metric for expressing aggregate results in a convenient range for tables, and histograms of NLL values for presenting further details. Both metrics reflect essentially the same thing, with lower values indicating a higher likelihood (see Theis et al. (2016) for information regarding these metrics and how they are computed). Furthermore, unlike Nalisnick et al. (2019) which varied the generative model type (the flow-based Glow method (Kingma & Dhariwal, 2018), autoregressive PixelCNN (van den Oord et al., 2016), and a basic VAE) but not the network architecture within each type, we focus on quantifying VAE capabilities as latent dimensionality and network capacity vary per our prior analysis.

We first consider FashionMNIST (Xiao et al., 2017) (inlier) vs MNIST (LeCun et al., 1998) (outlier) as κ is varied. It was reported in Nalisnick et al. (2019) that a Glow model trained on FashionMNIST produced lower BPD scores on MNIST, indicating that out-of-distribution data was preferred. However, when we trained our VAE baseline with capacity $\times 1$, we observe that FashionMNIST has a clearly lower BPD score across κ values ranging from 16 to 512 as shown in Table 1. This indicates that the model has in some sense correctly rejected the outliers independently of κ as would be expected for a model with suitable capacity. Note also that the BPD values for the MNIST testing data stabilize such that even with what would seem to be excessively high κ values, the extra degrees-of-freedom do not provide an inadvertent pathway for the out-of-distribution samples to receive undesirable preference (presumably because of pruning; see Section 4.2). Furthermore γ remains small and stable given that additional latent degrees-of-freedom are not needed to improve the data fit. We also plot a histogram of NLL values from all training and testing data in Figure 2. Most MNIST samples have much larger NLL as desired.

We next move on to a more challenging CIFAR-10 (Krizhevsky & Hinton, 2009) (inlier) vs SVHN (Netzer et al., 2011)/ CelebA (Liu et al., 2015) (outlier) case where changing network capacity can play a significant role. The VAE model trained with capacity $\times 1$ and $\kappa = 32$ now computes a lower BPD score for the outlier data because CIFAR-10 data is more complex than FashionMNIST such that the baseline capacity is inadequate. As we increase the capacity, we would therefore expect the model to better learn the data manifold, push γ to even smaller values, and eventually assign a lower BPD to the inlier samples. Table 2 indicates that this is in fact the case, and ultimately the outlier data is assigned a much higher BPD synced to lower γ values. Again we show the NLL histograms from all training and testing data in Figure 3. When the capacity is lower ($\times 1$, Figure 3(a)), most of the SVHN outlier samples have smaller NLL than the inlier data, implying that the VAE model fails to precisely learn the correct inlier manifold. The CelebA set also has just slightly higher NLL values than the inlier data even though the appearance of CelebA and CIFAR-10 are distinct. But when the network capacity is increased to $\times 8$ (Figure 3(b)), the inlier NLL values become smaller while the outlier NLL becomes larger as expected. Note that the CIFAR-10 train and test BPD and NLL are virtually the same, indicating that no overfitting has occurred. Therefore, we can safely apply even

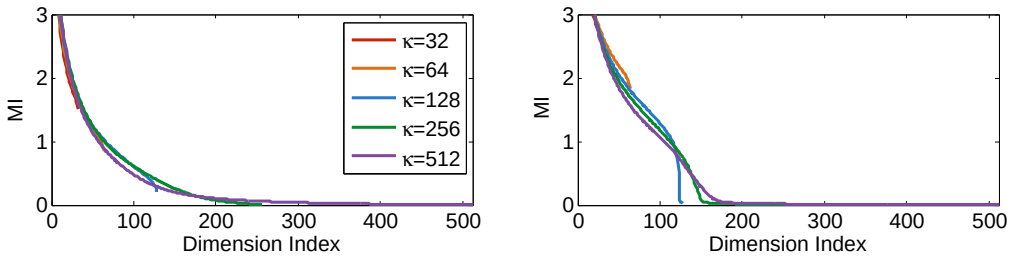
Capacity	Glow	$\times 1/4$	$\times 1/2$	$\times 1$	$\times 2$	$\times 3$	$\times 4$
CIFAR-10-Train	3.386	2.828	2.766	2.602	2.275	1.981	1.728
CIFAR-10-Test	3.464	2.824	2.761	2.598	2.272	1.978	1.726
SVHN-Test	2.389	2.404	2.394	2.406	2.579	2.890	3.367
CelebA-Test	–	2.844	2.882	3.106	3.965	5.053	6.424
γ	–	0.0110	0.0101	0.0082	0.0051	0.0035	0.0024

Table 2: BPD values for VAEs trained on CIFAR-10 using $\kappa = 32$ as capacity is varied. When the capacity is small, the outlier data has lower BPD as Nalisnick et al. (2019) has shown. However, as capacity increases, BPD for both training and testing sets decreases while that of the outlier data increases consistent with Section 3. Additionally, γ becomes significantly smaller with increased capacity as expected.



(a) Capacity$\times 1$,$\kappa = 32$ (b) Capacity$\times 8$,$\kappa = 32$

Figure 3: Histogram of VAE NLL values. Training on CIFAR-10, testing on CIFAR-10, CelebA, and SVHN. Increasing the network capacity makes the inlier NLL smaller but the outlier NLL larger as expected; larger capacity can potentially separate them even further.



(a) Capacity$\times 1$ (b) Capacity$\times 4$

Figure 4: MI values between each latent dimension and the data (sorted in descending order).

larger capacity models, which would likely improve the situation further, pushing inlier/outlier NLL curves more unmistakeably apart.

4.2 Pruning Superfluous Latent Dimensions

The ability to reconstruct the training data using the fewest number of active latent dimensions is critical to obtaining a tight manifold fit as argued in Section 3. We empirically demonstrate the VAE performance in this context by evaluating the mutual information (MI) between each latent dimension of $\boldsymbol{z}$ and the input $\boldsymbol{x}$ using FashionMNIST training data. If one such dimension is shut off during training, then the MI should be close to zero. For visualization purposes, we approximate true MI values using the VAE KL loss and then sort them in descending order. Results are shown in Figure 4. When κ is relatively small (e.g., $\kappa = 32$), all the dimensions are informative and display relatively large MI values. As κ is increased, there is diminishing information to transmit; however, the VAE model does not redistribute the mutual information across all the latent dimensions. Rather it is more likely to aggregate the useful information in roughly the same percentage of active dimensions, while ignoring superfluous dimensions when their cardinality increases. Additionally, increasing the capacity from $\times 1$ in 4(a) to $\times 4$ in 4(b) encourages the model to rely on even fewer latent degrees-of-freedom, producing a sharper cut-off between active and inactive latent dimensions. And for the larger κ values, this cut-off would likely be even sharper with additional training epochs.

REFERENCES

Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2:1–18, 2015.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401*, 2018.

Christopher Bishop. Novelty detection and neural network validation. *IEE Proceedings-Vision, Image and Signal Processing*, 141(4):217–222, 1994.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Lei Cai, Hongyang Gao, and Shuiwang Ji. Multi-stage variational auto-encoders for coarse-to-fine image generation. *arXiv preprint arXiv:1705.07202*, 2017.

Hyunsun Choi and Eric Jang. Generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

Bin Dai and David Wipf. Diagnosing and enhancing VAE models. *International Conference on Learning Representations*, 2019.

Bin Dai, Yu Wang, John Aston, Gang Hua, and David Wipf. Connections with robust PCA and the role of emergent sparsity in variational autoencoder models. *Journal of Machine Learning Research*, 2018.

Adji Dieng, Yoon Kim, Alexander Rush, and David Blei. Avoiding latent variable collapse with generative skip models. *arXiv preprint arXiv:1807.04863*, 2018.

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

Dan Hendrycks, Mantas Mazeika, and Thomas G Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018.

Diederik Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pp. 10215–10224, 2018.

Diederik Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Xiaopeng Li and James She. Collaborative variational autoencoder for recommender systems. In *International Conference on Knowledge Discovery and Data Mining*, pp. 305–314, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Pierre-Alexandre Mattei and Jes Frellsen. Leveraging the exact likelihood of deep latent variables models. *arXiv preprint arXiv:1802.04826*, 2018.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Ng. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on MNIST. In *International Conference on Learning Representations*, 2018.

Alireza Shafaei, Mark Schmidt, and James Little. Does your model know the digit 6 is not a cat? A less biased evaluation of "outlier" detectors. *arXiv preprint arXiv:1809.04729*, 2018.

Vít Škvára, Tomáš Pevnỳ, and Václav Šmídl. Are generative deep models for novelty detection truly better? *arXiv preprint arXiv:1807.05027*, 2018.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. How to train deep variational autoencoders and probabilistic ladder networks. *arXiv preprint arXiv:1602.02282*, 2016.

Hiroshi Takahashi, Tomoharu Iwata, Yuki Yamanaka, Masanori Yamada, and Satoshi Yagi. Student-t variational autoencoder for robust density estimation. In *International Joint Conference on Artificial Intelligence*, pp. 2696–2702, 2018.

Lucas Theis, Aaron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations*, pp. 1–10, 2016.

Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with PixelCNN decoders. In *Advances in Neural Information Processing Systems*, pp. 4790–4798, 2016.

Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *European Conference on Computer Vision*, pp. 835–851, 2016.

David Wipf and Yi Wu. Dual-space analysis of the sparse linear model. In *Advances in Neural Information Processing Systems*, pp. 1745–1753, 2012.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Haowen Xu, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, Dan Pei, Yang Feng, et al. Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. In *International World Wide Web Conference*, pp. 187–196, 2018.

Supplementary File
When Do Variational Autoencoder Know What They Don't Know?

1. Network Structure and Experimental Settings

The baseline encoder we adopt is composed of three convolutional layers of 64, 128, 256 channels with stride 2. The feature map is then flattened and fed into two FC layers producing $\boldsymbol{\mu}_z$ and $\log \boldsymbol{\sigma}_z$ respectively. The baseline decoder is composed of one linear layer with $4 \times 4 \times 256$ hidden units, a reshaping layer, and three transposed convolution layers with 128, 64 and C channels where C is the number of channels of the image.

2. Further Analysis and Evaluation of Unbounded VAE Gradients

Relevant to the content from Section 3.3 in the main text, it is worth acknowledging that energy functions involving infinite gradients and/or unbounded regions are already indispensable across a wide range of sparse estimation problems and structured regression [3]. This history implies that when training a VAE or other related autoencoder structure, we may borrow appropriate tools designed to mitigate the risk of converging to bad local solutions or regions of instability. In this vein, one effective strategy involves partially minimizing what amounts to a smoothed version of the original objective function. The degree of smoothness is then gradually reduced as the optimization trajectory moves towards an optimum. This procedure, which serves as a form of homotopy continuation method, is frequently used to find maximally sparse representations with minimal reconstruction error [1, 4, 6].

The VAE accomplishes something similar when we choose to iteratively estimate γ during training rather than merely setting its value to near zero as may be theoretically optimal (assuming we know that there exists sufficient network capacity to warrant this value). Initially, when the reconstruction cost is still high, γ will be relatively large and the overall VAE energy will be relatively smooth. It is only later as the data fit $\sum_{i=1}^{n} \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[\left\|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_x(\boldsymbol{z};\theta)\right\|_2^2\right]$ becomes small that γ will follow suite, and by this point it is more likely that we have already approached a basin of attraction capable of producing optimal sparse reconstructions, i.e., near zero reconstruction error using the fewest number of active latent dimensions.

In this regard, we now empirically demonstrate that learning γ, as a form of homotopy continuation method, may be better than fixing it to an arbitrarily small value. In particular, we first train a VAE model on CelebA data and learn an appropriate small value of γ denoted γ^*. We then retrain the same network from scratch but with $\gamma = \gamma^*$ fixed. The resulting models are evaluated via the reconstruction error and the maximum mean discrepancy (MMD) between the aggregated posterior $q_\phi(\boldsymbol{z}) \triangleq \sum_i q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})$ and the prior $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$. If too few latent dimensions are removed by swamping the appropriate

	CelebA	
	Rec. Err.	MMD
Learnable γ	352.8	93.3
Fix $\gamma = \gamma^*$	349.9	291.8

Table 1: Reconstruction error and MMD between $q_\phi(\boldsymbol{z})$ and $\mathcal{N}(0, \boldsymbol{I})$ on CelebA. We first train a VAE with learnable γ and obtain the optimal value γ^*. Then we fix $\gamma = \gamma^*$ and re-train the same network from scratch. Though the final reconstruction errors are almost the same, the MMDs between $q_\phi(\boldsymbol{z})$ and the standard $\mathcal{N}(0, \boldsymbol{I})$ are significantly different.

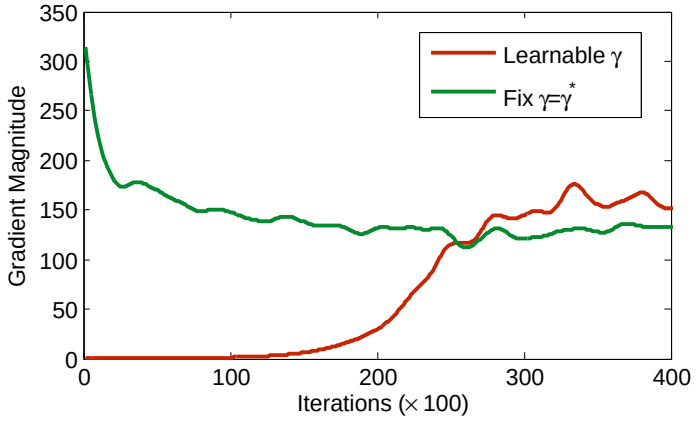


Figure 1: The Evolution of the gradient $\left\| \frac{d\mathcal{L}(\theta,\phi)}{d\boldsymbol{z}} \right\|_2$. Although both curves end up with similar final values, the large initial gradient with fixed γ is disruptive to the final solution.

channels with noise following the prior, then we would expect $q_\phi(\boldsymbol{z})$ to be confined to a low-dimensional manifold in $\mathbb{R}^\kappa$ and the MMD to be much larger.

Results are displayed in Table 1, where as expected, the reconstruction errors are nearly identical, but the learnable γ has much lower MMD values. We also plot the evolution of the gradient magnitudes $\left\| \frac{d\mathcal{L}(\theta,\phi)}{d\boldsymbol{z}} \right\|_2$ in Figure 1 (other gradients are similar). When γ is learned, the gradient increases slowly; however, with fixed $\gamma = \gamma^*$, there exists a huge gradient right from the start since γ^* is small but the reconstruction error is high. This contributes to a worse final solution per the Table 1.

3. Proof of Theorem 1

To begin, we assume that $h(u)$ is a concave, non-decreasing function defined on the domain $u \geq 0$. These are central characteristics of sparsity inducing penalty functions [2, 5] and it is not difficult to show that additional flexibility does not gain us anything in the present context. For convenience, we assume that h is differentiable everywhere, although this condition can also be relaxed. We then focus on the case where the gradient of h is bounded. Per these specifications, the largest gradient will necessarily occur at $h'(0) \equiv \lim_{u \to 0^+} h'(u)$.

Note also that this limiting gradient cannot equal zero; otherwise we trivially default to a flat penalty function such that all solutions have equal cost and the theorem guarantee is unattainable right from the start.

From here, the basic idea is to construct a counterexample that satisfies the conditions of the theorem, and yet involves a simple network structure that, if $h'(u)$ is bounded around zero, is unable to minimize the stated objective using at most r nonzero rows of $\boldsymbol{Z}$ while simultaneously achieving zero reconstruction error. To this end, consider the two-dimensional latent representation $\boldsymbol{z} = [z_1, z_2]^\top$ and a single-parameter decoder that computes

$$\boldsymbol{\mu}_x(\boldsymbol{z}; \theta) = \theta \pi(z_1) + (1-\theta) \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \tag{1}$$

where $\theta \in \Omega \triangleq [0,1]$ is a scalar parameter, $t : \mathbb{R} \rightarrow [0,1]$ truncates its argument to the interval between zero and one and $\pi : [0,1] \rightarrow \mathcal{S} \subset [0,1]^2$ is for now an arbitrary function defined on the stated interval. Per this construction, the decoder can be viewed as a tunable mixture weighted by θ, and for either $\theta = 0$ or $\theta = 1$, the range of the decoder $\boldsymbol{\mu}_x(\boldsymbol{z}; \theta)$ is contained within the unit square $[0,1]^2$.

Now suppose we have training samples $\{\boldsymbol{x}^{(i)}\}_{i=1}^n$ that were produced via the generative process

$$z_{gt}^{(i)} \sim p(z_{gt}) \quad \text{and} \quad \boldsymbol{x}^{(i)} = \pi\left(t\left[z_{gt}^{(i)}\right]\right) \tag{2}$$

for some prior $p(z_{gt})$ on the ground-truth latent variable $z_{gt} \in \mathbb{R}$. Furthermore, assume that the function π is such that for all $t[z_{gt}] \in [C,1]$ with constant $C < 1$, $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \pi(t[z_{gt}])$ satisfies $0 < |x_j| < \epsilon$ for $j = 1,2$, with $\epsilon > 0$ arbitrarily small. We also stipulate that $p(z_{gt})$ places all (or almost all) of its probability mass such that $t[z_{gt}] \in [C,1]$, which implies that the observed training points will all be arbitrarily close to zero.

Given this observed data, we can then evaluate the optimal AE for different penalties h. We allow that the encoder is sufficiently complex such that

$$\min_{\phi} \mathcal{L}_h(\theta, \phi) \quad \equiv \quad \min_{\boldsymbol{Z}} \; h\left(\tfrac{1}{dn}\sum_{i=1}^n \left\|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_x\left(\boldsymbol{z}^{(i)}; \theta\right)\right\|_2^2\right) + \tfrac{1}{d}\sum_{k=1}^{\kappa} h\left(\tfrac{1}{n}\|\boldsymbol{z}_k\|_2^2\right), \tag{3}$$

where in the present context $\kappa = d = 2$, and as mentioned in the main text, $\boldsymbol{z}_k$ represents the k-th row of $\boldsymbol{Z}$. This arrangement is equivalent to simply assuming that the encoder is capable of computing the minimizing $\boldsymbol{z}^{(i)}$ for each index (i.e., we have removed amortized inference). We adopt this assumption for simplicity of exposition, but the same conclusions can be drawn in broader conditions.

To achieve zero reconstruction under the stated conditions using only $r = 1$ nonzero rows of $\boldsymbol{Z}$, we must choose $\theta = 1$. In this restricted setting, the optimal $\boldsymbol{Z}$ will satisfy $\frac{1}{n}\|\boldsymbol{z}_1\|_2^2 \geq C^2$ and $\frac{1}{n}\|\boldsymbol{z}_2\|_2^2 = 0$ such that the overall objective value will be

$$\min_{\phi} \mathcal{L}_h(\theta = 1, \phi \in \Phi) \;=\; h(0) + \frac{1}{2}\left[h(0) + h\left(\tfrac{1}{n}\|\boldsymbol{z}_1\|_2^2\right)\right] \;\geq\; \frac{3}{2}h(0) + \frac{1}{2}h\left(C^2\right), \tag{4}$$

where Φ is the set of ϕ that lead to zero reconstruction error. In other words, within the current setup, the constraints $\theta = 1$ and $\phi \in \Phi$ are necessary conditions for any solution to achieve an optimal sparse reconstruction.

But now suppose we choose $\theta = 0$. In this revised situation, the optimal unconstrained $\boldsymbol{Z}$ will satisfy $\frac{1}{n}\|\boldsymbol{z}_1\|_2^2, \frac{1}{n}\|\boldsymbol{z}_1\|_2^2 \leq \epsilon^2$. The associated cost then becomes

$$\min_{\phi} \mathcal{L}_h(\theta = 0, \phi) = h(0) + \tfrac{1}{2}\sum_{k=1}^{2} h\left(\tfrac{1}{n}\|\boldsymbol{z}_k\|_2^2\right) \leq h(0) + h\left(\epsilon^2\right). \tag{5}$$

At this point, without loss of generality assume that $h\left(C^2\right) = 1$ and $h\left(0\right) = 0$, which can be accomplished by simply translating and rescaling the overall cost. Because $\lim_{u\to 0^+} h'(u)$ is bounded, the gap between $h(0)$ and $h\left(\epsilon^2\right)$ can be made arbitrarily small for ϵ sufficiently small. In contrast, the gap between $h\left(\epsilon^2\right)$ and $h\left(C^2\right)$ can be arbitrarily close to one. Therefore, it follows that if our data was generated with ϵ sufficiently small, then

$$\min_{\phi} \mathcal{L}_h(\theta = 1, \phi \in \Phi) \geq \tfrac{1}{2} > \min_{\phi} \mathcal{L}_h(\theta = 0, \phi) \approx 0, \tag{6}$$

and so the unique solution achieving zero construction error with a single active latent variable cannot be the global optimum. Or equivalently, any globally optimum solution will not coincide with an optimal sparse reconstruction.

Note that the situation would be completely different if $h(u) = \mathcal{I}[u > 0]$, meaning an indicator function that equals zero if $u = 0$ and one for all $u > 0$. In this case, it is obvious that $\min_{\phi} \mathcal{L}_h(\theta = 1, \phi \in \Phi) = \frac{1}{2}$ while all other solutions will be such that $\mathcal{L}_h(\theta, \phi) \geq 1$. But of course this h does not have a bounded gradient everywhere because of the discontinuity at zero.

High-level picture: While this is obviously a toy counterexample designed with a specific technical purpose in mind, it is nonetheless emblematic of situations that may naturally arise in practice. For example, it is easy to envision scenarios where data is lying on a complex r-dimensional manifold that is contained within a larger $(r+s)$-dimensional manifold (or possibly subspace) that has much simpler structure. Perfectly reconstructing such data could be accomplished using only r degrees-of-freedom or $(r+s)$ degrees-of-freedom depending on whether the low- or high-dimensional manifold was accurately modeled. But unless we have a penalty function with a strong preference for lower-dimensional structures, then the network may well favor or converge to the simpler, higher-dimensional alternative.

References

[1] Rick Chartrand and Wotao Yin. Iteratively reweighted algorithms for compressive sensing. *International Conference on Accoustics, Speech, and Signal Processing*, 2008.

[2] Yichen Chen, Dongdong Ge, Mengdi Wang, Zizhuo Wang, Yinyu Ye, and Hao Yin. Strong NP-hardness for sparse optimization with concave penalty functions. In *ICML*, pages 740–747, 2017.

[3] Irina Gorodnitsky and Bhaskar Rao. Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm. *IEEE Transactions on signal processing*, 45(3):600–616, 1997.

[4] Yue Hu, Sajan Goud Lingala, and Mathews Jacob. A fast majorize–minimize algorithm for the recovery of sparse and low-rank matrices. *IEEE Transactions on Image Processing*, 21(2):742–753, 2012.

[5] Jason Palmer, David Wipf, Kenneth Kreutz-Delgado, and Baskar Rao. Variational EM algorithms for non-Gaussian latent variable models. *Advances in Neural Information Processing Systems*, pages 1059–1066, 2006.

[6] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1107–1114, 2013.