On the Regularization Balance in Autoencoder-Based Generative Models

Anonymous Authors[1]

Abstract

Although GANs have recently displayed impressive visual results, there remains value in generative models that do not rely on adversarial training and are capable of producing posterior estimates of latent representations when conditioned on an input. Within this scope, we analyze generative models of continuous data based upon a regularized autoencoder structure, e.g., variational autoencoders (VAEs) and related architectures. First, we demonstrate that with an affine decoder there will exist a combinatorial number of distinct local minima; however, each of these will necessarily be global optima displaying balanced regularization. By this we mean that perfect reconstructions can be obtained using a minimal number of latent factors, which we contend is essential for generating realistic samples with an autoencoder structure. We then contrast the balanced case with more complex situations where over-regularized solutions may produce poor reconstructions, or under-regularized solutions may lead to unnecessarily redundant latent representations. Proceeding further, we discuss practical solutions for addressing over-regularization, while arguing that under-regularization avoidance in a certain sense requires the use of energy functions with infinite gradients (or nearly so) around optima (e.g., a Gaussian VAE model with an optimal decoder variance near zero). While potentially leading to numerical issues, such energy functions are essentially unavoidable across a wide range of traditional optimization problems designed to produce maximum parsimony, but strategies exist to at least partially mitigate instabilities. Empirical examples using real-world images corroborate our findings and demonstrate that non-adversarial training, when properly regularized, can produce reasonable samples even with full-resolution CelebA data.

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

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Suppose we have access to continuous variables $\boldsymbol{x}$ that are assumed to lie on or near an r-dimensional manifold $\boldsymbol{\chi}$ embedded in $\mathbb{R}^d$, where $r \ll d$. This construction captures the notion that our data of interest possesses low-dimensional structure relative to the high-dimensional ambient space. We denote a ground-truth probability measure on $\boldsymbol{\chi}$ as μ_{gt}, which assigns probability mass $\mu_{gt}(d\boldsymbol{x})$ to the infinitesimal $d\boldsymbol{x}$ residing on the manifold; it naturally follows that $\int_{\boldsymbol{\chi}} \mu_{gt}(d\boldsymbol{x}) = 1$. Given samples $\{\boldsymbol{x}^{(i)}\}_{i=1}^n$ drawn from this measure, our goal is to estimate the unknown μ_{gt}, or at least to provide a tractable way of drawing new samples that are indistinguishable in distribution from the training sample.

Because of the assumed low-dimensional structure, it is common to approximate the unknown ground-truth measure via a model parameterized as $p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$, where θ are parameters, $\boldsymbol{z} \in \mathbb{R}^\kappa$ serves as a low-dimensional representation, with $\kappa \approx r$ and fixed prior $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$. Samples from $p_\theta(\boldsymbol{x})$ can then be easily produced by drawing $\boldsymbol{z} \sim p(\boldsymbol{z})$ and then $\boldsymbol{x} \sim p_\theta(\boldsymbol{x}|\boldsymbol{z})$, noting that if $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ happens to be a Dirac delta function, the only source of randomness arises from sampling $\boldsymbol{z}$, which is then processed by what defaults to a deterministic transformation to produce $\boldsymbol{x}$.

Recently, generative adversarial networks (GANs) have demonstrated impressive results generating novel signals/images by learning such a deterministic $p_\theta(\boldsymbol{x}|\boldsymbol{z}) \equiv \delta\left[\boldsymbol{x} - f_\theta(\boldsymbol{z})\right]$ (Goodfellow et al., 2014). The potential downside though is that the required adversarial training can be sensitive to mode collapse (Arora & Zhang, 2017; Metz et al., 2016), hyperparameter tuning (Lucic et al., 2018), and other implementation/training issues. Moreover, standard GAN models do not provide a tractable way of computing the posterior $p_\theta(\boldsymbol{z}|\boldsymbol{x})$, which is useful for a variety of applications (Dai et al., 2018). However, there also exist modeling approaches that do not rely on adversarial training while producing direct posterior estimates. The variational autoencoder (VAE) and its many variants represent influential examples that fall within this category (Kingma & Welling, 2014; Rezende et al., 2014).

In principle we might consider minimizing the negative log-likelihood $-\log p_\theta(\boldsymbol{x})$ averaged across the training data, approximating integration across the assumed μ_{gt}, i.e., min-

imize

$$\frac{1}{n}\sum_i -\log\left[p_\theta\left(\boldsymbol{x}^{(i)}\right)\right] \;\approx\; \int -\log\left[p_\theta(\boldsymbol{x})\right]\mu_{gt}(d\boldsymbol{x}) \quad (1)$$

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 VAE instead attempts to minimize the variational upper bound $\mathcal{L}(\theta,\phi) \triangleq$

$$\begin{aligned}&\int\left\{-\log p_\theta(\boldsymbol{x}) + \mathbb{KL}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})\right]\right\}\mu_{gt}(d\boldsymbol{x})\\ &= \int\left\{-\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta\left(\boldsymbol{x}|\boldsymbol{z}\right)\right]\right.\\ &\qquad\left. + \mathbb{KL}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right]\right\}\mu_{gt}(d\boldsymbol{x}) \qquad (2)\end{aligned}$$

or a sampling-based approximation. 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 $\mathcal{L}(\theta,\phi)$, or at least its approximate version given samples from μ_{gt}, can be minimized over $\{\theta,\phi\}$ using SGD and a simple reparameterization trick (Kingma & Welling, 2014; Rezende et al., 2014).

The VAE as described above is representative of a wider class of generative models built upon an AE structure. Broadly speaking, if the overriding objective is generating realistic samples using an encoder-decoder-based architecture (VAE or otherwise), two important, well-known criteria must be satisfied:

(i) Small reconstruction error when passing through the encoder-decoder networks, and

(ii) An *aggregate posterior* distribution (Makhzani et al., 2016) over latent codes $\boldsymbol{z}$, defined by $q_\phi(\boldsymbol{z}) \triangleq \int q_\phi(\boldsymbol{z}|\boldsymbol{x})\mu_{gt}(d\boldsymbol{x})$, that is close to a known distribution like $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|0,I)$ that is easy to sample from.

The first criteria can be naturally enforced by a deterministic AE, but also for the VAE provided that sufficient randomness can be suppressed by the encoder. Of course the second criteria is equally important. Without it, we have no tractable way of generating random decoder inputs that, when passed to the decoder output, are transformed to realistic samples resembling the true data distribution. Additionally, if we assume that $\kappa = r$, then criteria (*ii*) involves comparing distributions that should have finite density across a smaller r-dimensional space. This avoids potentially problematic comparisons in $\mathbb{R}^d$ between measures that assign all or most probability mass to a lower-dimensional manifold with infinite density.

Criteria (i) can in principle be handled by a sufficiently complex aggregate AE network, while (*ii*) can be addressed by increasing the flexibility of $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ (Burda et al., 2015; Kingma et al., 2016; Rezende & Mohamed, 2015; van den Berg et al., 2018) or the prior $p(\boldsymbol{z})$ (Bauer & Mnih, 2018; Tomczak & Welling, 2018) such that there are additional pathways for pushing the intractable aggregate posterior towards a tractable prior that we can sample from. Other approaches exist as well. For example, unlike the VAE which relies on a KL regularization factor to indirectly favor $q_\phi(\boldsymbol{z}|\boldsymbol{x}) \approx p(\boldsymbol{z})$, the Wasserstein autoencdoer (WAE) (Tolstikhin et al., 2018) instead utilizes an energy function that more directly penalizes the distance between $q_\phi(\boldsymbol{z})$ and $p(\boldsymbol{z})$. The WAE can operate using either adversarial training akin to the adversarial autoencoder (Makhzani et al., 2016), or via a non-adversarial, maximum mean discrepancy (MMD) penalty as considered herein. In contrast, it has also been shown that a second-stage VAE module can be used to sample from $q_\phi(\boldsymbol{z})$, even when its distribution is far from $p(\boldsymbol{z})$ (Dai & Wipf, 2018).

Despite significant differences, all of these methods are sensitive to the regularization balance required for simultaneously addressing criteria (*i*) and (*ii*). Although alternative definitions are certainly possible, for present purposes we adopt a very specific notion of what constitutes *balanced regularization*. In particular, we will make the case that a necessary condition for achieving high-quality samples via an autoencoder structure devoid of adversarial training, is an encoder-decoder pair that produces high-quality reconstructions of the training data per criteria (*i*), using a minimal number of latent degrees-of-freedom, a prerequisite for enjoying the favorable $\kappa = r$ scenario associated with criteria (*ii*). This then leads to the following trifold distinction:

- **Balanced Regularization:** We say that a solution exhibits balanced regularization we achieve near perfect reconstruction of training samples using a minimal number of non-random/nonzero latent dimensions that convey information about $\boldsymbol{x}$.
- **Over-Regularization:** Many or all latent variables are unused or highly random such that the training data is poorly reconstructed.
- **Under-Regularization:** The reconstruction of the training data exhibits low error; however, the latent representation is redundant, i.e., an unnecessarily large number of active latent variables are used.

Achieving balanced regularization in the sense described

above does not guarantee that generated samples will be of high quality; however, as we will soon argue, over-regularized and less obviously under-regularized solutions will necessarily denigrate performance (at least when adversarial training is not used as is our focus). It therefore behooves us to pursue energy functions and/or training strategies that directly favor balanced regularization, which as will discussed in Section 2, facilitates satisfying criteria (*ii*). In doing so we will largely focus on a baseline VAE architecture while drawing parallels to alternative methods.

To this end, in Section 3 we describe sufficient conditions whereby the combinatorial set of distinct VAE local minima, with Gaussian encoder/decoder distributions and affine mean/variance networks, are all global optima exhibiting balanced regularization. To the best of our knowledge, this is the only such full characterization of the VAE local minima landscape involving an actual, non-trivial parameterization. Next, Section 4 details the over-regulation problem, while presenting practical strategies designed to avoid pruning away too many active dimensions at suboptimal local minima that may arise within more complex VAE (or related) architectures.

In contrast, Section 5 argues that under-regularization avoidance in a certain precise sense requires the use of energy functions with infinite gradients (or nearly so) around optima, e.g., a Gaussian VAE model with an optimal decoder variance near zero. While potentially leading to numerical issues, we describe how such energy functions are essentially unavoidable across a wide range of traditional optimization problems designed to produce maximum parsimony, as well as strategies to at least partially mitigate instabilities. We conclude with empirical examples using real-world images to corroborate our findings and demonstrate that non-adversarial training, when properly regularized, can produce reasonable samples even with full-resolution CelebA data. To the best of our knowledge, this is the first presentation of this type of realistic non-adversarial result on real-world images of this size and complexity.

2. The Value of Balanced Regularization

Over-Regularization may occur in two distinct circumstances leading to obviously deleterious effect. First, if we set $\kappa < r$, then it is generally impossible to reconstruct the training data and satisfy criteria (*i*).[1] Essentially all latent-variable generative modeling architectures, including GANs, will experience this straightforward form of over-regularization if κ is too small. Secondly though, if $\kappa \geq r$, then over-regularization can still potentially contribute to degraded performance. For example, the combined energy function may include solutions, or possibly bad local optima, whereby more than $\kappa - r$ latent dimensions are swamped with random noise. Within a VAE this can happen if too many dimensions of $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ become independent of $\boldsymbol{x}$ as observed in numerous prior works discussed in Section 4.

The impact of under-regularization as defined in Section 1 is decidedly less obvious and more contingent on the specific model architecture. But a key nuisance emerges with autoencoder structures: If $\kappa > r$ and the decoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is deterministic, or at least nearly deterministic along more than r latent dimensions, then the aggregate posterior $q_\phi(\boldsymbol{z})$ will occupy zero measure within the κ-dimensional space, i.e., randomness from $\boldsymbol{x} \sim \mu_{gt}$ can only "fill" r dimensions. It is therefore impossible to match with a fixed prior such as $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|0, I)$. And even if we replace such a rigid prior with a flexible, parameterized alternative, matching degenerate distributions with near infinite density on a manifold is quite challenging, and is more-or-less just another version of our original problem, namely, matching $p_\theta(\boldsymbol{x})$ to μ_{gt}. It is therefore preferable to achieve balanced regularization, whereby no more or less than $\kappa - r$ dimensions of $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ produce random noise such that $q_\phi(\boldsymbol{z})$ occupies nonzero measure across κ dimensions, allowing for a more suitable match with a non-degenerate prior. (Alternatively, if $\kappa - r$ dimensions were simply set to zero, they could presumably be detected and similarly removed as well.)

In this context, balanced regularization can be viewed as a weaker prerequisite for achieving criteria (*i*) and (*ii*): it implies that we have simultaneously achieved zero reconstruction error and a latent representation such that $q_\phi(\boldsymbol{z})$ effectively occupies nonzero measure in $\mathbb{R}^\kappa$. We next examine conditions such that this is guaranteed to happen within a Gaussian VAE architecture, at least assuming we can find a local minima of the energy function

3. Sufficient Conditions for All Local Minima Exhibiting Balanced Regularization

Because our focus is continuous data, we consider a typical VAE model with a Gaussian decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$, where the mean function $\boldsymbol{\mu}_x$ is parameterized by θ and takes $\boldsymbol{z}$ as its input. As is often assumed in practice, we set $\boldsymbol{\Sigma}_x = \gamma \boldsymbol{I}$, where $\gamma > 0$ is a scalar parameter within the parameter set θ. If the decoder mean function is restricted to be affine, i.e. $\boldsymbol{\mu}_x = \boldsymbol{W}_x \boldsymbol{z} + \boldsymbol{b}_x$ for some weight matrix $\boldsymbol{W}_x$ and bias vector $\boldsymbol{b}_x$, then the VAE energy function remains highly non-convex with potentially a combinatorial number of distinct local minima. However, we will examine conditions whereby all of these local minima are actually global optima that display balanced regularization as desired.

Although we could proceed by allowing the encoder to be arbitrarily complex, when the decoder mean function is

[1] This is true unless we have a decoder capacity that scales proportionally to the training data. The latter facilitates rote memorization even if $\kappa = 1 \ll r$ as discussed in (Dai et al., 2018).

forced to be affine and $\boldsymbol{\Sigma}_x = \gamma \boldsymbol{I}$, then a Gaussian encoder with an affine representation for $\boldsymbol{\mu}_z = \boldsymbol{W}_z \boldsymbol{x} + \boldsymbol{b}_z$ and a diagonal $\boldsymbol{\Sigma}_z = \text{diag}[\boldsymbol{s}^2]$, where $\boldsymbol{s}^2$ is an arbitrary non-negative parameter vector independent of $\boldsymbol{x}$, are sufficient to achieve the optimal VAE cost. Collectively, these specifications lead to the parameterization $\theta = \{\boldsymbol{W}_x, \boldsymbol{b}_x, \gamma\}$ and $\phi = \{\boldsymbol{W}_z, \boldsymbol{b}_z, \boldsymbol{s}^2\}$ and energy given by $\mathcal{L}(\theta, \phi) =$

$$\int \Big\{ \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \Big[\tfrac{1}{\gamma} \|\boldsymbol{x} - \boldsymbol{W}_x \boldsymbol{z} - \boldsymbol{b}_x\|_2^2 \Big] + d \log \gamma \qquad (3)$$
$$+ \sum_{k=1}^{\kappa} \left(s_k^2 - \log s_k^2 \right) + \|\boldsymbol{W}_z \boldsymbol{x} + \boldsymbol{b}_z\|_2^2 \Big\} \mu_{gt}(d\boldsymbol{x}).$$

We then have the following:

Theorem 1 *For any fixed value of γ, all local minima of (3) with respect to the parameters $\{\boldsymbol{W}_x, \boldsymbol{b}_x, \boldsymbol{W}_z, \boldsymbol{b}_z, \boldsymbol{s}^2\}$ are also global minima. Additionally, if the support of μ_{gt} is confined to an r-dimensional affine space, then these global minima will exhibit balanced regularization when $\gamma \rightarrow 0$, i.e., zero reconstruction error using only r active/non-random dimensions of $\boldsymbol{z}$.*

See the supplementary for proof details. This balanced regularization occurs because each possible local/global optima will necessarily define the principal subspace of the data using a minimum number of nonzero columns of $\boldsymbol{W}_x$. Furthermore, at the indeces of these zero-valued columns, elements of $\boldsymbol{s}^2$ will converge to zero while the corresponding elements of $\boldsymbol{\mu}_z$ will convey information about $\boldsymbol{x}$ (i.e., active, non-random dimensions). Elsewhere the variances will be set to one while the means will equal zero, indicating that no useful information about $\boldsymbol{x}$ is being transferred. And because $\|\boldsymbol{\mu}_z\|_0 = r$ and the reconstruction error is zero when $\gamma \rightarrow 0$, we have achieved balanced regularization per the stated definition.

We emphasize that, regardless of μ_{gt}, the number of potential minima can be combinatorially high. For example, if the covariance $\text{Cov}_{\mu_{gt}}[\boldsymbol{x}]$ equals $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top$, with $\text{rank}[\boldsymbol{\Psi}] = r < \kappa$, then there can exist $\frac{\kappa!}{(\kappa - r)!}$ distinct or disconnected minima of (3).[2] This follows from a much simpler analysis of affine VAE models from (Dai et al., 2018), where only the possibility of global (but not local) minima were considered. But since all are equally good per our results herein, this generally lessens the concern as there is no issue of suboptimal local minima with imbalanced regularization. Of course the model described by (3) is obviously a simplified version of the VAE; however, Theorem 1 nonetheless represents the only characterization of the full constellation of VAE local minima under non-trivial circumstances. It also complements the wide variety of recent efforts to analyze the complex loss surface of neural networks with linear layers (Choromanska et al., 2015a;b; Kawaguchi, 2016).

[2] By disconnected we mean that, to traverse from one minimum to another, we must ascend the objective function at some point along the connecting path.

4. Over-Regularization Scenarios and Practical Workarounds

As we move to more complex/deeper encoder and decoder parameterizations, it is no longer possible to characterize the complex configuration of local minima as we presented in Section 3. However, we can at least elucidate a particularly pernicious set of stationary points associated with over-regularization that may be disruptive to successful optimization trajectories in many practical settings. Note that over-regularization is well-acknowledged in the literature, with compensatory measures including heuristic annealing of the KL penalty (Bowman et al., 2015; Sønderby et al., 2016), tighter bounds on the log-likelihood (Burda et al., 2015; Rezende & Mohamed, 2015), more complex priors (Bauer & Mnih, 2018; Tomczak & Welling, 2018), or modified decoder architectures (Cai et al., 2017; Dieng et al., 2018; Yeung et al., 2017). Thus far though, published results do not indicate success generating high-resolution images, and in the majority of cases, evaluations are limited to small black-and-white and/or binary images. Additionally, unlike much existing work, in this section we analyze the problem and suggest workarounds purely from the perspective of avoiding bad VAE local minima. This increases the chances of reaching global solutions designed to achieve balanced regularization as discussed later in Section 5 and high-quality samples as shown in Section 6.

To begin, we again assume a VAE model with Gaussian encoder and decoder networks. Both encoder and decoder mean functions $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_z$, as well as the diagonal encoder covariance function $\boldsymbol{\Sigma}_z = \text{diag}[\boldsymbol{\sigma}_z^2]$, are computed by standard deep neural networks, with layers composed of linear weights followed by element-wise non-linear activations (the decoder covariance satisfies $\boldsymbol{\Sigma}_x = \gamma \boldsymbol{I}$ as before). We denote the weight matrix from the first layer of the decoder mean network as $\boldsymbol{W}^1_{\mu_x}$, while $\boldsymbol{w}^1_{\mu_x, \cdot j}$ refers to the corresponding j-th column. Assuming ρ layers, we denote $\boldsymbol{W}^\rho_{\mu_z}$ and $\boldsymbol{W}^\rho_{\sigma_z^2}$ as weights from the last layers of the encoder networks producing $\boldsymbol{\mu}_z$ and $\log \boldsymbol{\sigma}_z^2$ respectively, with j-th rows defined as $\boldsymbol{w}^\rho_{\mu_z, j\cdot}$ and $\boldsymbol{w}^\rho_{\mu_z, j\cdot}$. We then characterize the following key stationary point:

Theorem 2 *If $\boldsymbol{w}^1_{\mu_x, \cdot j} = \left(\boldsymbol{w}^\rho_{\mu_z, j\cdot}\right)^\top = \left(\boldsymbol{w}^\rho_{\sigma_z, j\cdot}\right)^\top = \boldsymbol{0}$ for any $j \in \{1, 2, \ldots, \kappa\}$, then the gradients of $\mathcal{L}(\theta, \phi)$ with respect to $\boldsymbol{w}^1_{\mu_x, \cdot j}$, $\boldsymbol{w}^\rho_{\mu_z, j\cdot}$, and $\boldsymbol{w}^\rho_{\sigma_z^2, j\cdot}$ are all equal to zero.*

Without further details, the type of stationary point described by Theorem 2 could conceivably function as a saddle point (stable or unstable), a local maximum, or a local minimum. However, in many important representative situations we would argue that it serves as the latter, or more specifically, a potentially degenerate local minimum with all or most latent dimensions suppressing information about $\boldsymbol{x}$ such that an over-regularized solution is produced with high recon-

struction error. In other words, for dimensions where (**??**) is satisfied, the decoder will simply output i.i.d. unit-Gaussian noise that will then be zeroed out by the decoder.

For example, if we consider the VAE KL term in isolation, it follows that

$$\begin{aligned} &\arg \min_{\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z \succ \mathbf{0}} \mathbb{KL}\left[q_\phi\left(\boldsymbol{z}|\boldsymbol{x}\right)||p(\boldsymbol{z})\right] \;\equiv \\ &\arg \min_{\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z \succ \mathbf{0}} \left\{\text{tr}\left[\boldsymbol{\Sigma}_z\right] + \|\boldsymbol{\mu}_z\|_2^2 - \log|\boldsymbol{\Sigma}_z|\right\} \;=\; \{\mathbf{0}, \boldsymbol{I}\}. \end{aligned} \tag{4}$$

The stationary point from Theorem 2 naturally achieves this minimal solution across all j. So based on this term alone, any movement away from $\boldsymbol{w}^{\rho}_{\mu_z,j\cdot} = \boldsymbol{w}^{\rho}_{\sigma_z^2,j\cdot} = \mathbf{0}$ to modify $\boldsymbol{\mu}_z$ and/or $\boldsymbol{\Sigma}_z = \text{diag}[\boldsymbol{\sigma}_z^2]$ will necessarily involve going against the gradient.

Of course the VAE model has a second data-dependent reconstruction factor $-\mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]$. But assuming the decoder mean network is sufficiently deep, with weights near a random initialization not in the vicinity of an optimal representation, then any perturbation from $\boldsymbol{w}^1_{\mu_x,\cdot j} = \boldsymbol{w}^{\rho}_{\mu_z,j\cdot} = \boldsymbol{w}^{\rho}_{\sigma_z,j\cdot} = \mathbf{0}$ may well involve an uphill battle. This is because the only way for any gradient to pass through the decoder is to perturb $\boldsymbol{w}^1_{\mu_x,\cdot j}$ away from zero, but the moment this happens, the large random signal coming from the encoder, which is dominated by white Gaussian noise when $\boldsymbol{\mu}_z = \mathbf{0}$ and $\boldsymbol{\Sigma}_z = \boldsymbol{I}$, will likely only serve to disrupt any attempted data fit after passing through the deep decoder, increasing the reconstruction cost.

Therefore, in the present aggregated circumstances, nudging $\boldsymbol{w}^1_{\mu_x,\cdot j} \neq \mathbf{0}$ will be against the gradient of $-\mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]$, while pushing $\boldsymbol{\mu}_z \neq \mathbf{0}$ and $\boldsymbol{\Sigma}_z \prec \boldsymbol{I}$ (through changes in $\boldsymbol{w}^{\rho}_{\mu_z,j\cdot}$ and $\boldsymbol{w}^{\rho}_{\sigma_z,j\cdot}$ for some/all j) to improve the predictive signal-to-noise ratio will necessarily be against the gradient of $\mathbb{KL}\left[q_\phi\left(\boldsymbol{z}|\boldsymbol{x}\right)||p(\boldsymbol{z})\right]$. Hence the stationary point described by Theorem 2 can easily serve as a degenerate, over-regularized local minima. It is also important to keep in mind that arriving at such an over-regularized solution only requires the movements of the isolated weight matrices $\boldsymbol{W}^1_{\mu_x}$, $\boldsymbol{W}^{\rho}_{\mu_z}$ and $\boldsymbol{W}^{\rho}_{\sigma_z^2}$ towards zero. In contrast, to actually minimize the reconstruction error requires the complex coordination of *all* the layer weights from both the deep encoder and decoder networks, which is far more circuitous than in the affine case.

Fortunately though, multiple complementary countermeasures can be introduced to lessen the impact of this type of over-regularized local minima. To see this, consider the case where $\boldsymbol{W}^1_{\mu_x} = \boldsymbol{W}^{\rho}_{\mu_z} = \boldsymbol{W}^{\rho}_{\sigma_z^2} = \mathbf{0}$, such that we are located in parameter space at a likely local minimum characterized by maximal over-regularization. Now suppose we add an additional affine skip connection between $\boldsymbol{z}$ and $\boldsymbol{x}$, by which we mean that the decoder mean function is modified to

$$\boldsymbol{\mu}_x = f_\theta(\boldsymbol{z}) + \boldsymbol{W}_x\boldsymbol{z} + \boldsymbol{b}_x, \tag{5}$$

where $f_\theta(\boldsymbol{z})$ is the original deep decoder network and $\boldsymbol{W}_x$ and $\boldsymbol{b}_x$ are additional parameters analogous to those defined in Section 3. If $\boldsymbol{W}^1_{\mu_x} = \mathbf{0}$, then $f_\theta(\boldsymbol{z})$ equals some constant $\boldsymbol{c}$ that can be absorbed into $\boldsymbol{b}_x$. Hence we are now in the regime described by Theorem 1 and the analysis that follows, from which we can conclude that, even with $f_\theta(\boldsymbol{z})$ anchored in a bad configuration, available local minima are nonetheless obtainable with $\boldsymbol{W}_x$ and $\boldsymbol{b}_x$ spanning the principal subspace of the data rather than a degenerate, maximally over-regularized solution. This provides a plausible escape route. Additionally, the larger we make $\kappa = \dim[\boldsymbol{z}]$, the better equipped this affine bridge is to provide a preliminary fit to the data, or move further away from over-regularized local minima.

This then suggest three practical prescriptions for mitigating over-regularization:

1. Include skip connections between decoder layers,
2. Increase the value of κ such that $\kappa \gg r$ or even potentially $\kappa \approx d$ if over-regularization remains an issue.
3. If it is not practically feasible to significantly increase κ, then first train a deterministic, unregularized AE with κ set to a suitable range such that a low reconstruction error is achievable using a latent code denoted as $\widetilde{\boldsymbol{x}}$. Then train a smaller VAE treating $\widetilde{\boldsymbol{x}}$ as the new data and with $\widetilde{d} \triangleq \dim[\widetilde{\boldsymbol{x}}] \approx \kappa$.

With regard to prescription 2, there are multiple reasons for potentially choosing $\kappa \gg r$. First, we generally do not know r in advance, so it is obviously easier to simply choose some κ that is significantly larger rather than trying to guess $\kappa = r$. Secondly though, and equally important, the larger we make κ, *the more likely that we can fully capitalize on skip connections and avoid over-regularized solutions*. And furthermore, redundant dimensions that exist when $\kappa \gg r$ can be pruned anyway if a balanced regularization state is eventually achieved, which is possible at certain global minima as described in Section 5. Experimental results from Section 6 demonstrate that both of the above modifications are necessary in tandem.

As for optional prescription 3, an unregularized AE provides a simple way of reducing the dimensionality (and also complexity/depth) without the risk of over-regularization. A smaller/simpler VAE can then be introduced to take maximal advantage of prescriptions 1 and 2. Note that from an implementation standpoint, we adopt the reconstruction factor $\mathbb{E}_{q_\phi(\boldsymbol{z}|\widetilde{\boldsymbol{x}})}\left[\frac{1}{2\gamma}\left\|\boldsymbol{x} - \widetilde{f}_{\widetilde{\theta}}\left[f_\theta\left(\boldsymbol{z}\right)\right]\right\|_2^2\right]$, where $\widetilde{f}_{\widetilde{\theta}}$ denotes the learned AE with parameters $\widetilde{\theta}$. This merely implies that

we measure the reconstruction cost in the original high-dimensional space. This is similar to methods that substitute various feature-space loss functions into the VAE energy (Hou et al., 2017).

Before proceeding, we should mention that skip connections have been proposed previously in an attempt to enhance VAE models (Cai et al., 2017; Dieng et al., 2018). However, the motivation was different, and the importance of pairing with a higher dimensional $\boldsymbol{z}$ for the specific purpose of escaping bad local minima has not been explored.

5. Under-Regularization and Infinite Gradients as a Necessary Risk

We now consider necessary conditions for avoiding under-regularization in generic AE architectures, which will later be used to elucidate key VAE properties. In this regards, consider the constrained objective function $\mathcal{L}_h(\theta,\phi) =$

$$h\left(\frac{1}{dn}\sum_{i=1}^{n}\left\|\boldsymbol{x}^{(i)}-f_\theta\left(\boldsymbol{z}^{(i)}\right)\right\|_2^2\right)+\frac{1}{d}\sum_{k=1}^{\kappa}h\left(\frac{1}{n}\|\boldsymbol{z}_k\|_2^2\right)$$
$$\text{s.t. } \boldsymbol{z}^{(i)}=g_\phi\left(\boldsymbol{x}^{(i)}\right)\ \forall i,\ \theta\in\Theta, \tag{6}$$

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

Given a generic AE architecture as in (6), it is natural to examine what possible functions h are such that any global minimum of $\mathcal{L}_h(\theta,\phi)$ is guaranteed to exhibit balanced regularization. This can be addressed as follows:

Theorem 3 *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 $\boldsymbol{x}^{(i)}=f_\theta\left(\boldsymbol{z}^{(i)}\right)$ for all 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.*

Note that a similar result can be obtained by replacing the reconstruction penalty with the additional constraint $\sum_{i=1}^n\left\|\boldsymbol{x}^{(i)}-f_\theta\left(\boldsymbol{z}^{(i)}\right)\right\|_2^2=0$, in which case no trade-off parameter, fixed or otherwise, need be included. We also emphasize that Theorem 3 effectively implies that, to guarantee every global optima displays balanced regularization 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 more-or-less tantamount to requiring penalty functions with an energy gap that is unbounded about zero. For example, the selection $h(u)=\mathcal{I}[u>0]$, i.e., an indicator function that equals zero if $u=0$ and one for all $u>0$, will guarantee that any global minimum of (6) displays balanced regularization under the stated conditions. However, given that $h(u)\equiv\lim_{p\rightarrow 0}u^p$ and $\lim_{p\rightarrow 0}\frac{1}{p}(u^p-1)=\log u$, 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, 2018; 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 balanced regularization. In this regard, it has been argued that as $\gamma\rightarrow 0$, the VAE can achieve 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 minimize the increase in the KL regularization factor (Dai & Wipf, 2018). This is exactly a stochastic version of balanced regularization, and ensures that $q_\phi(\boldsymbol{z})$ occupies full measure in κ-dimensional space, even if $q_\phi(\boldsymbol{z})\neq p(\boldsymbol{z})$.

Furthermore, because $q_\phi(\boldsymbol{z})$ is not a troublesome degenerative distribution, it is ripe for pairing with any standard generative modeling paradigm that facilitates tractable sampling. For example, we could apply a second-stage VAE as proposed in (Dai & Wipf, 2018), or invertible networks capable of warping a κ-dimensional Gaussian into an arbitrary κ-dimensional distribution with full measure (Dinh et al., 2016). Regardless, we contend that achieving balanced regularization represents an essential stepping stone, even if infinite gradients constitute a necessary evil.

We close this section by 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 (Gorodnitsky & Rao, 1997). 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 (Chartrand & Yin, 2008; Hu et al., 2012; Xu et al., 2013).

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. 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 latter as $\sum_{i=1}^{n} \left\| \boldsymbol{x}^{(i)} - f_\theta\left(\boldsymbol{z}^{(i)}\right) \right\|_2^2$ becomes small that γ will follow suite, and by this point it is more likely that we have already approached a basin of attraction with balanced regularization.

6. Emperimental Validation and Discussion

A Simple Over- and Under-Regularization Test: We first perform experiments to examine robustness to over-regularization using the prescriptions from Section 4 and under-regularization using a Gaussian decoder with trainable variance parameter γ per Section 5. The objective at this point is to achieve optimal reconstructions using minimal latent dimensions (i.e., balanced regularization), as opposed to the final goal of generating high-quality samples which will be addressed later. We train a VAE model on MNIST data (LeCun et al., 1998). The encoder is a 6-layer feed-forward network and the decoder is a residual network (He et al., 2016) with 3 residual blocks, each containing 2 layers, to instantiate skip connections. Note that a Gaussian decoder as we have adopted is unlike many previous works that simplify the problem by binarizing the data and then applying a Bernoulli decoder appropriate for binary images. We vary the latent dimension κ from 50 to 784, where $d = 784$ is the dimension of the data itself. See the supplementary file for training details. Three different metrics are adopted to evaluate models: the reconstruction error, the negative log-likelihood (NLL) on the test set, and the number of active latent dimensions per the metric from (Burda et al., 2015).

Table 1 displays the results, where we observe that as κ becomes sufficiently large, the reconstruction error mostly decreases while the number of active dimensions saturates around 20. Overall the NLL on the test set improves even up to $\kappa = 784$, indicating that the model has likely not under-regularized (or overfit) as well, presumably because excessive active dimensions have been pruned. Interestingly though, if we remove the skip connections, this relatively deep model completely over-regularizes, and all the active dimensions are turned off leading to completely useless reconstructions (not shown). So indeed both the skip connections and sufficiently large κ are needed.

κ	Rec. Err.	NLL	# Active
50	4.98	-315	10
100	3.42	-580	15
200	2.74	-718	20
400	2.57	-743	22
784	2.78	-806	22

Table 1. Reconstruction error (test), NLL (test), and number of active dimensions on MNIST as κ is varied.

	CelebA-64		CelebA-128	
	Rec. Err.	MMD	Rec. Err.	MMD
Learnable γ	60.5	20.24	352.8	93.3
Fix $\gamma = \gamma^*$	59.6	69.59	349.9	291.8

Table 2. Reconstruction error and MMD between $q_\phi(\boldsymbol{z})$ and $\mathcal{N}(0, \boldsymbol{I})$. 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.

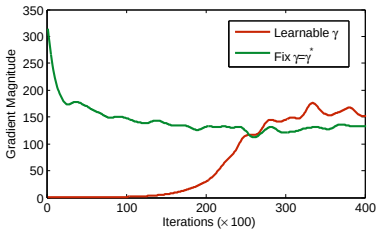

Figure 1. The Evolution of the gradient $\left\| \frac{d\mathcal{L}(\theta,\phi)}{d\boldsymbol{z}} \right\|_2$. Although both curves end up with similar final values, the large initial gradient with fixed γ is disruptive to the final solution.

Under-Regularization and the Benefits of Learning γ: Although a small γ may lead to large VAE gradients, we have argued that this may constitute a necessary risk if we want to obtain minimal reconstruction error using the fewest number of active latent dimensions, i.e. avoid under-regularization. We now demonstrate that learning γ, as a form of homotopy continuation method, is better than fixing it to an optimal small value as discussed in Section 5. In particular, we first train a VAE on the CelebA dataset (Liu et al., 2015), both 64×64 and 128×128 resolution versions, and learn an appropriate small value of γ denoted γ^* (see supplementary file for network and training details). We then retrain the same network from scratch but with $\gamma = \gamma^*$ fixed. The resulting models are evaluated via the recon-

struction error and the maximum mean discrepancy (MMD) between the aggregated posterior $q_\phi(\boldsymbol{z})$ and the prior $p(\boldsymbol{z})$. If too few latent dimensions are removed by swamping the appropriate channels with noise following the prior, then we would expect $q_\phi(\boldsymbol{z})$ to be confined to a manifold in $\mathbb{R}^\kappa$ and the MMD to be much larger.

Results are displayed in Table 2, where as expected, the reconstruction errors are nearly identical, but the learnable γ has much lower MMD values. For the 128×128 case, we also plot the evolution of the gradient magnitudes $\left\| \frac{d\mathcal{L}(\theta,\phi)}{d\boldsymbol{z}} \right\|_2$ in Figure 1. When γ is learned (red curve), 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 2 results.

From Balanced Regularization to Good Samples: Thus far we have merely investigated strategies for achieving balanced regularization. We now apply these ideas to the original goal of generating samples of high-resolution images. Note that the vast majority of AE-structured, non-adversarial approaches are only tested using NLL scores or related; however, this is largely orthogonal to the goal of improving quantitative measures of visual quality (Theis et al., 2016). Instead, metrics such as the Fréchet Inception Distance (FID) (Heusel et al., 2017) have been designed for this purpose, as well as complementary scores like the recently proposed Number of statistically-Different Bins (NDB) (Richardson & Weiss, 2018). The latter evaluates generation performance in the original images space by binning samples into K learned clusters; it therefore does not depend on subtleties of inception network like FID, and provides information about high-level sample diversity.

We quantitatively evaluate networks on FID and NDB scores using the original, full-resolution CelebA data center-cropped to 128×128. Based on the analysis herein, we first train an unregularized AE to project the original data down to a $\widetilde{d} = 64$-dimensional feature space. In this regime, it is feasible to train a small VAE with $\widetilde{d} = \kappa$ and skip connections to achieve balanced regularization (at least assuming that $\widetilde{d}$ is sufficiently large). And finally, per the suggestion of (Dai & Wipf, 2018), we train a second small VAE to sample from the aggregate posterior $q_\phi(\boldsymbol{z})$ produced by the first, which should benefit from balanced regularization. Network and training details are in the supplementary. We refer to this model as VAE+, which is simple to train since nearly all of the parameters reside in the AE structure.

Unlike other non-adversarial AE-based models, the WAE-MMD approach (Tolstikhin et al., 2018) has also been quantitatively evaluated on real-world color images like CelebA, but only on a downsampled version. For balanced comparison, we train a WAE-MMD network using the aggregate AE plus VAE structure described above. As an additional baseline, we repeat this process with a plain VAE model using the same architecture. The reconstruction and generation FID scores are shown in Table 3. The reconstruction FID, which is only sensitive to criteria (*i*), can be regarded as a lower bound of the generation FID score since it is unlikely that the generated samples could ever be better than the reconstructed ones. The fact that the VAE+ generation FID is so close to the reconstruction FID indicates that it has done well in addressing criteria (*ii*) via its sensitivity to balanced regularization. We also report results in Table 4 using the complementary NDB metric. The VAE+ again performs best by a significant margin, indicating a higher level of sample diversity. And finally, we show some of the good samples generated by VAE+ in Figure 2. More samples from VAE+ and other methods are in the supplementary.

	Recon. FID	Gen. FID
Downsample	39.2	39.2
WAE-MMD	59.8	66.9
VAE	45.9	64.8
VAE+	45.6	47.5

Table 3. Reconstruction and generation FID scores on 128×128 CelebA dataset (smaller is better). *Downsample* refers to comparing the original images with downsampled 64×64 versions. This seemingly innocuous transformation, which preserves all the major structures, can produce an FID almost as high as VAE+. This indicates that the FID may be sensitive to minor details, which could impact the performance of AE-structured models.

	$K = 100$	$K = 200$	$K = 300$
Downsample	0.05	0.06	0.04
WAE-MMD	0.64	0.62	0.58
VAE	0.88	0.85	0.81
VAE+	0.29	0.28	0.24

Table 4. NDB scores with varying numbers of bins (K). *Downsample* is defined as in Table 3; this value is near 0 as expected.

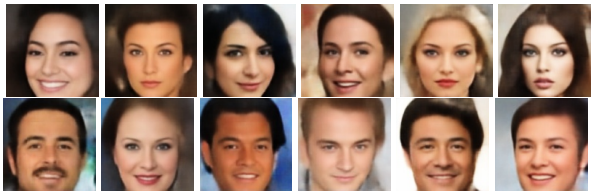

Figure 2. Select generated samples

Discussion: We have analyzed the regularization balance of AE-structured models in general and VAEs in particular, including local minima properties and necessary energy function characteristics. This leads to useful practical prescriptions and the first demonstration of high-quality, diverse generation results from AE-structured, non-adversarial training on 128×128 color images. The only other attempt at this resolution comes from (Cai et al., 2017), but the generated images are comparably blurry (see supplementary) and no quantitative evaluation has been demonstrated.

References

Arora, S. and Zhang, Y. Do gans actually learn the distribution? an empirical study. *arXiv preprint arXiv:1706.08224*, 2017.

Bauer, M. and Mnih, A. Resampled priors for variational autoencoders. *arXiv preprint arXiv:1810.11428*, 2018.

Belloni, A., Chernozhukov, V., and Wang, L. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Cai, L., Gao, H., and Ji, S. Multi-stage variational autoencoders for coarse-to-fine image generation. *arXiv preprint arXiv:1705.07202*, 2017.

Chartrand, R. and Yin, W. Iteratively reweighted algorithms for compressive sensing. *Proc. Int. Conf. Accoustics, Speech, and Signal Proc.*, 2008.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G., and LeCun, Y. The loss surfaces of multilayer networks. In *International Conference on Artificial Intelligence and Statistics*, 2015a.

Choromanska, A., LeCun, Y., and Arous, G. Open problem: The landscape of the loss surfaces of multilayer networks. In *Conference on Learning Theory*, 2015b.

Dai, B. and Wipf, D. Diagnosing and enhancing gaussian VAE models. *Advances in Neural Information Processing Systems, Bayesian Deep Learning Workshop*, 2018.

Dai, B., Wang, Y., Aston, J., Hua, G., and Wipf, D. Connections with robust PCA and the role of emergent sparsity in variational autoencoder models. *Journal of Machine Learning Research*, 2018.

Dieng, A. B., Kim, Y., Rush, A. M., and Blei, D. M. Avoiding latent variable collapse with generative skip models. *arXiv preprint arXiv:1807.04863*, 2018.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. In *arXiv preprint arXiv:1406.2661*, 2014.

Gorodnitsky, I. F. and Rao, B. D. Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *IEEE Transactions on signal processing*, 45(3):600–616, 1997.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Hou, X., Shen, L., Sun, K., and Qiu, G. Deep feature consistent variational autoencoder. In *Applications of Computer Vision, 2017 IEEE Winter Conference on*, pp. 1133–1141. IEEE, 2017.

Hu, Y., Lingala, S. G., and Jacob, M. A fast majorize–minimize algorithm for the recovery of sparse and low-rank matrices. *IEEE Transactions on Image Processing*, 21(2):742–753, 2012.

Kawaguchi, K. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, 2016.

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

Kingma, D., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pp. 4743–4751, 2016.

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

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. Are GANs created equal? A large-scale study. *Advances in Neural Information Processing Systems*, 2018.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2016.

Mattei, P.-A. and Frellsen, J. Leveraging the exact likelihood of deep latent variables models. *arXiv preprint arXiv:1802.04826*, 2018.

Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

Rezende, D., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

Richardson, E. and Weiss, Y. On gans and gmms. *arXiv preprint arXiv:1805.12462*, 2018.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. How to train deep variational autoencoders and probabilistic ladder networks. *arXiv preprint arXiv:1602.02282*, 2016.

Takahashi, H., Iwata, T., Yamanaka, Y., Yamada, M., and Yagi, S. Student-t variational autoencoder for robust density estimation. In *International Joint Conference on Artificial Intelligence*, pp. 2696–2702, 2018.

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

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

Tomczak, J. and Welling, M. VAE with a VampPrior. In *International Conference on Artificial Intelligence and Statistics*, pp. 1214–1223, 2018.

van den Berg, R., Hasenclever, L., Tomczak, J. M., and Welling, M. Sylvester normalizing flows for variational inference. In *Uncertainty in Artificial Intelligence*, 2018.

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

Xu, L., Zheng, S., and Jia, J. Unnatural l0 sparse representation for natural image deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1107–1114, 2013.

Yeung, S., Kannan, A., Dauphin, Y., and Fei-Fei, L. Tackling over-pruning in variational autoencoders. *arXiv preprint arXiv:1706.03643*, 2017.