SUPPLEMENTARY FILE: On the Regularization Balance in Autoencoder-Based Generative Models

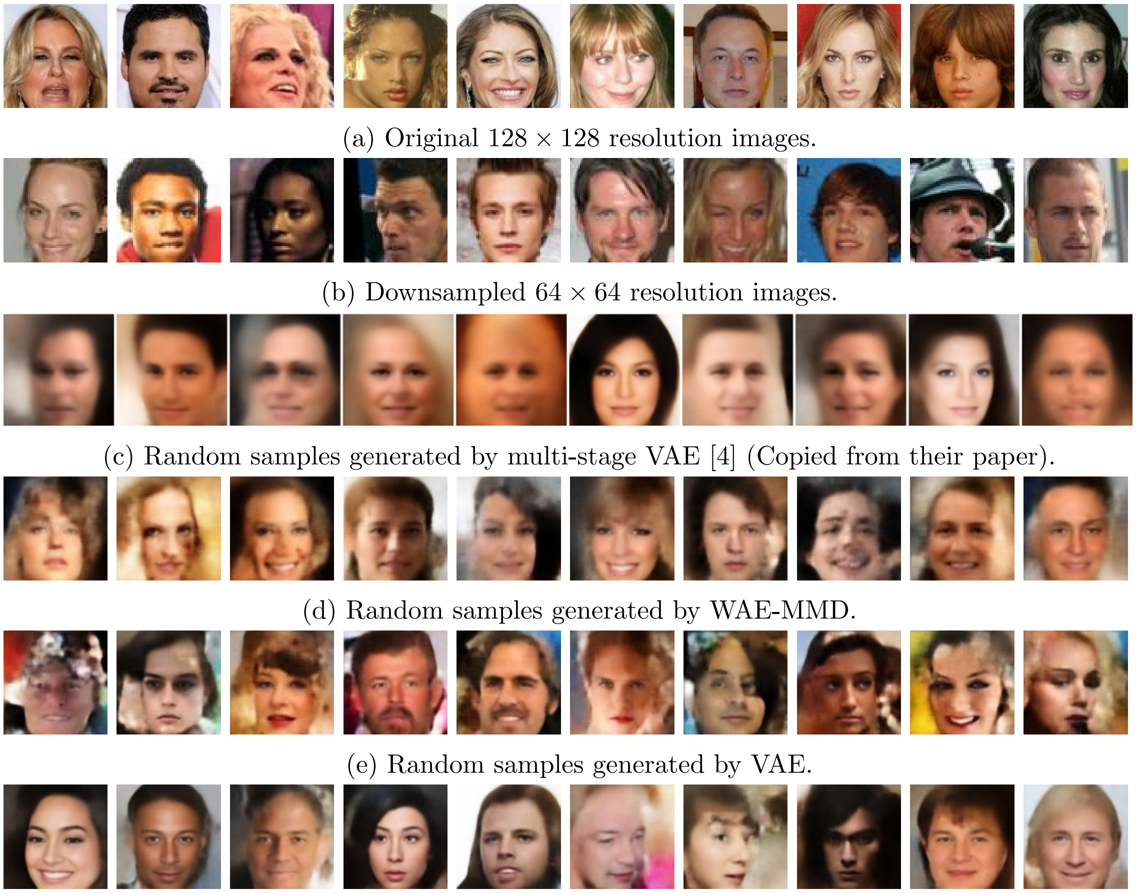
1. Extra Empirical Results

Randomly generated samples and the example training images are shown in Figure 1. The original 128×128 resolution ground-truth images and the downsampled versions at 64×64 resolution are shown in Figure 1a and 1b. The visual qualities are quite similar. If we zoom in to see details, the downsampled images are only slightly more blurry. However, the FID difference between these two versions of the training datasets are almost 40, as shown in Table 2 of the main paper (the FID of any image set with an exact copy will equal zero). This indicates that the FID score is very sensitive to tiny blurriness artifacts. So a relative high FID for a VAE model does not necessary mean that the visual quality is bad, nor that sample diversity is poor. From Figure 1c to 1f, we show randomly generated samples of different methods without cherry picking. VAE+ obviously generates samples with the best visual quality.

2. Network Structure and Experimental Settings

For the MNIST dataset, the encoder has 6 feed-forward layers while the decoder has 3 residual blocks, each containing two layers. All the hidden layers have 200 dimensions followed by a hyperbolic tangent activation. We follow the training schemes of [3]. The batch size is 20 and the learning rate is $0.001 \times 10^{(-i/7)}$ for $i \in \{0, 1, ..., 7\}$. For each i, we train the network for 3^i epochs. There is no weight decay and the optimizer is Adam with default settings of Tensorflow.

For the CelebA dataset, the network structure is shown in Figure 2. An AE is first trained on the original dataset to transform it into a low-dimensional feature space, as illustrated in Figure 2(a). The details of the encoder network are in Figure 2(c) and the decoder just flips the encoder structure. After this, we train a small VAE with skip connections as shown in Figure 2(b). The parameters of the AE are fixed in this training stage. Then we train a second-stage VAE on the latent code as proposed in [6] to warp the aggregated posterior $q_\phi(\boldsymbol{z})$ to a standard Gaussian distribution. The latent variable of the second stage VAE is denoted as $\boldsymbol{u}$. The structure of the second stage VAE is the same as the first stage VAE (yellow part in Figure 2(b) and (d)). In the generation phase, we randomly sample $\boldsymbol{u}$ from a standard Gaussian distribution and feed it through the decoder of the second stage VAE and obtain $\boldsymbol{z}$. Then we feed $\boldsymbol{z}$ thought the first stage VAE decoder and the AE decoder successively and obtain the generated samples. Only the training set of the CelebA dataset is used for training. The batch size is 128 and the network is trained 400 epochs for each stage with learning rate 0.0001 and Adam optimizer.



(a) Original 128 × 128 resolution images.

(b) Downsampled 64 × 64 resolution images.

(c) Random samples generated by multi-stage VAE [4] (Copied from their paper).

(d) Random samples generated by WAE-MMD.

(e) Random samples generated by VAE.

(f) Random samples generated by VAE+.

Figure 1: Ground-truth and generated image samples using full-resolution CelebA data center-cropped to size 128 × 128. The downsampled images in (b) are slightly more blurry than the original images in (a) but the visual quality is roughly the same. However, the FID score between the downsampled images and the original images is almost 40, as shown in the main paper. Among all the generated samples, the VAE+ produces the best visual quality in general.

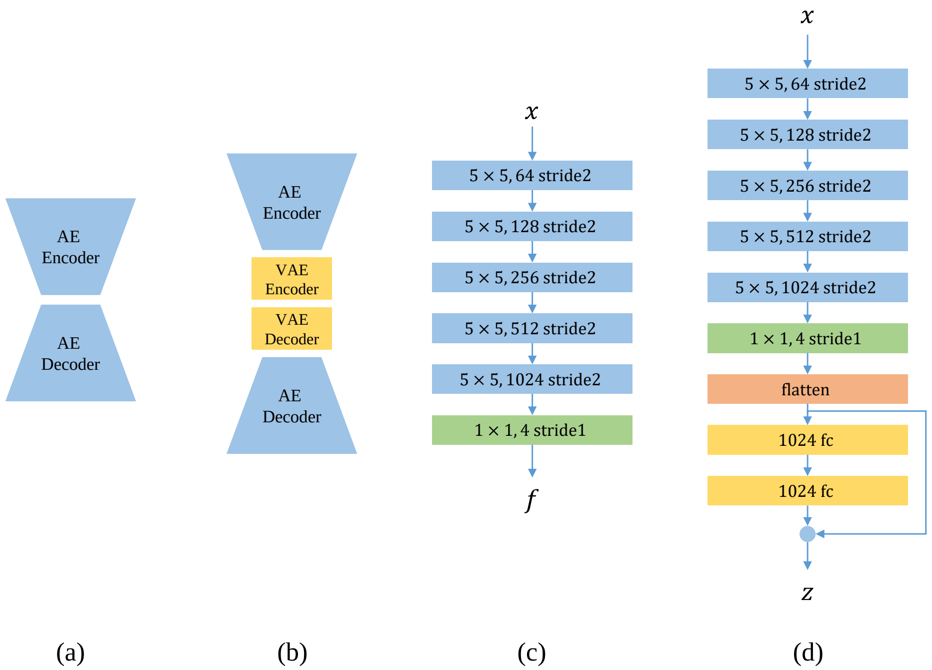


Figure 2: The encoder network for CelebA dataset. (a) We first train an unregularized AE. (b) We then fix the AE network and train a small VAE network on the features. (c) The details of the AE encoder network. The decoder network just flips this structure. The input $\boldsymbol{x}$ is a $128 \times 128 \times 3$ image. After 5 convolution layers with stride 2, we obtain a feature map with shape $4 \times 4 \times 1024$. A 1×1 convolution layer is then adopted to transform the feature map to $4 \times 4 \times 4$. For 64×64 resolution images, we remove the last stride 2 convolution layer such that the feature map is $4 \times 4 \times 512$ before the 1×1 convolution layer. (d) The whole encoder network structure when training the VAE. The blue parts are fixed when training the VAE part. Again the decoder network flips this structure.

3. Proof of Theorem 1

Proof Under the stated assumptions, the VAE cost can be further simplified as

$$\begin{aligned}
\mathcal{L}(\theta, \phi) &= \int \Big\{ \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\tfrac{1}{\gamma} \left\| \boldsymbol{x} - \boldsymbol{W}_x \boldsymbol{z} - \boldsymbol{b}_x \right\|_2^2 \right] + d \log \gamma \\
&\quad + \sum_{k=1}^{\kappa} \left(s_k^2 - \log s_k^2 \right) + \left\| \boldsymbol{W}_z \boldsymbol{x} + \boldsymbol{b}_z \right\|_2^2 \Big\} \mu_{gt}(d\boldsymbol{x}) \\
&= \int \Big\{ \tfrac{1}{\gamma} \left\| \boldsymbol{x} - \boldsymbol{W}_x \left(\boldsymbol{W}_z \boldsymbol{x} + \boldsymbol{b}_z \right) - \boldsymbol{b}_x \right\|_2^2 + d \log \gamma \\
&\quad + \sum_{k=1}^{\kappa} \left(s_k^2 - \log s_k^2 + \tfrac{1}{\gamma} s_k^2 \| \boldsymbol{w}_{x,k} \|_2^2 \right) + \left\| \boldsymbol{W}_z \boldsymbol{x} + \boldsymbol{b}_z \right\|_2^2 \Big\} \mu_{gt}(d\boldsymbol{x}), \qquad (1)
\end{aligned}$$

where $\boldsymbol{w}_{x,k}$ denotes the k-th column of $\boldsymbol{W}_x$. Although this expression is non-convex in each s_k^2, by taking derivatives and setting them equal to zero, it is easily shown that there is a single stationary point that operates as the unique minimum. Achieving the optimum requires only that $s_k^2 = \left[\frac{1}{\gamma}\|\boldsymbol{w}_{x,k}\|_2^2 + 1\right]^{-1}$ for all k. Plugging this value into (1) then leads to the revised objective

$$\begin{aligned}\mathcal{L}(\theta,\phi) \quad &\equiv \quad \int \Big\{ \tfrac{1}{\gamma}\left\|\boldsymbol{x} - \boldsymbol{W}_x\left(\boldsymbol{W}_z\boldsymbol{x} + \boldsymbol{b}_z\right) - \boldsymbol{b}_x\right\|_2^2 \qquad (2)\\ &+ \quad \sum_{k=1}^{\kappa} \log\left(\tfrac{1}{\gamma}\|\boldsymbol{w}_{x,k}\|_2^2 + 1\right) + d\log\gamma + \|\boldsymbol{W}_z\boldsymbol{x} + \boldsymbol{b}_z\|_2^2 \Big\}\, \mu_{gt}(d\boldsymbol{x})\end{aligned}$$

ignoring constant terms. Similarly we can optimize over $\boldsymbol{\mu}_z = \boldsymbol{W}_z\boldsymbol{x} + \boldsymbol{b}_z$ in terms of the other variables. This is just a convex, ridge regression problem, with the optimum uniquely satisfying

$$\boldsymbol{W}_z\boldsymbol{x} + \boldsymbol{b}_z = \boldsymbol{W}_x^\top\left(\gamma\boldsymbol{I} + \boldsymbol{W}_x\boldsymbol{W}_x^\top\right)^{-1}\left(\boldsymbol{x} - \boldsymbol{b}_x\right), \qquad (3)$$

which is naturally and affine function of $\boldsymbol{x}$ as required. Plugging (3) into (2) and applying some linear algebra manipulations, we arrive at

$$\begin{aligned}\mathcal{L}(\theta,\phi) \quad &\equiv \quad \int \left\{ \left(\boldsymbol{x} - \boldsymbol{b}_x\right)^\top \left(\boldsymbol{W}_x\boldsymbol{W}_x^\top + \gamma\boldsymbol{I}\right)^{-1}\left(\boldsymbol{x} - \boldsymbol{b}_x\right)\right\} \mu_{gt}(d\boldsymbol{x})\\ &+ \quad \sum_{k=1}^{\kappa} \log\left(\|\boldsymbol{w}_{x,k}\|_2^2 + \gamma\right) + (d-\kappa)\log\gamma. \qquad (4)\end{aligned}$$

Finally, the optimal $\boldsymbol{b}_x$ is just the convex maximum likelihood estimator given by the mean $\boldsymbol{b}_x = \int \boldsymbol{x}\mu_{gt}(d\boldsymbol{x})$, and therefore by plugging this value into (4) and applying a standard trace identity, we arrive at

$$\mathcal{L}(\theta,\phi) = \operatorname{tr}\left[\boldsymbol{\Sigma}_{gt}\left(\boldsymbol{W}_x\boldsymbol{W}_x^\top + \gamma\boldsymbol{I}\right)^{-1}\right] + \sum_j \log\left(\|\boldsymbol{w}_{x,k}\|_2^2 + \gamma\right) + (d-\kappa)\log\gamma, \qquad (5)$$

where $\boldsymbol{\Sigma}_{gt} \triangleq \mathrm{Cov}_{\mu_{gt}}\left[\boldsymbol{x}\right]$. Hence we have thus far optimized $\boldsymbol{s}^2$, $\boldsymbol{W}_z$, $\boldsymbol{b}_z$, and $\boldsymbol{b}_x$ out of the model without encountering any local minima as of yet. However, the expression (5) that remains is a highly non-convex function of $\boldsymbol{W}_x$ where non-global/suboptimal local minima may still be lurking.

We investigate this as follows. Because $\log\left(\|\boldsymbol{w}_{x,k}\|_2^2 + \gamma\right)$ is a concave non-decreasing function of $\|\boldsymbol{w}_{x,k}\|_2^2$, it can be expressed via the variational form

$$\log\left(\|\boldsymbol{w}_{x,k}\|_2^2 + \gamma\right) \quad = \quad \min_{\omega_k \geq 0}\left\{\omega_k\|\boldsymbol{w}_{x,k}\|_2^2 - h^*(\omega_k)\right\}, \qquad (6)$$

where $h^*(\omega)$ denotes the concave conjugate function [1] of $h(u) \triangleq \log(u+\gamma)$, $u \geq 0$. This formulation produces a strict upper bound once we drop the minimization, i.e.,

$$\log\left(\|\boldsymbol{w}_{x,k}\|_2^2 + \gamma\right) \quad \geq \quad \omega_k\|\boldsymbol{w}_{x,k}\|_2^2 - h^*(\omega_k), \ \forall\omega_k \geq 0. \qquad (7)$$

Now assume for the moment that $\boldsymbol{W}_x'$ is some local minimum of (5) and that, for all k, we have that

$$\log\left(\|\boldsymbol{w}_{x,k}'\|_2^2 + \gamma\right) \quad = \quad \omega_k'\|\boldsymbol{w}_{x,k}'\|_2^2 - h^*(\omega_k') \qquad (8)$$

for some corresponding non-negative ω'_k values. By the above variational construction we are guaranteed that such ω'_k will always exist. If $\boldsymbol{W}'_x$ is truly a local minimum, then it must also be a local minimum of the upper bound

$$\bar{\mathcal{L}}(\theta,\phi) \triangleq \tag{9}$$
$$\operatorname{tr}\left[\boldsymbol{\Sigma}_{gt}\left(\boldsymbol{W}_x\boldsymbol{W}_x^\top + \gamma\boldsymbol{I}\right)^{-1}\right] + \sum_{k=1}^{\kappa}\left[\omega'_k\|\boldsymbol{w}_{x,k}\|_2^2 - h^*(\omega'_k)\right] + (d-\kappa)\log\gamma.$$

Note that after optimizing away all the other variables, at the point $\boldsymbol{W}_x = \boldsymbol{W}'_x$, $\bar{\mathcal{L}}(\theta,\phi) = \mathcal{L}(\theta,\phi)$, and yet by construction $\bar{\mathcal{L}}(\theta,\phi) \geq \mathcal{L}(\theta,\phi)$ everywhere. Therefore if the bound is not minimized at this point, then $\boldsymbol{W}'_x$ cannot be a minimum to $\mathcal{L}(\theta,\phi)$.

Using the standard singular value decomposition, let $\boldsymbol{W}_x = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are square orthonormal matrices and $\boldsymbol{\Lambda}$ is a positive semi-definite diagonal matrix. Any minimum of $\bar{\mathcal{L}}(\theta,\phi)$ or $\mathcal{L}(\theta,\phi)$ with respect to $\boldsymbol{W}_x$ must also be a minimum with respect to $\boldsymbol{U}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{V}$, otherwise we could smoothly alter one of these components to smoothly change $\boldsymbol{W}_x$ and reduce the cost. Now denote $\boldsymbol{\Omega}$ as a zero-valued matrix with each ω'_k placed in the k-th diagonal position. Then we have that

$$\sum_{k=1}^{\kappa}\omega'_k\|\boldsymbol{w}_{x,k}\|_2^2 = \operatorname{tr}\left[\boldsymbol{W}_x^\top\boldsymbol{W}_x\boldsymbol{\Omega}\right] = \operatorname{tr}\left[\boldsymbol{V}\boldsymbol{\Lambda}^2\boldsymbol{V}^\top\boldsymbol{\Omega}\right]. \tag{10}$$

Because $\boldsymbol{W}_x\boldsymbol{W}_x^\top = \boldsymbol{U}\boldsymbol{\Lambda}^2\boldsymbol{U}^\top$ is independent of $\boldsymbol{V}$, minimization of $\bar{\mathcal{L}}(\theta,\phi)$ need only involve the term $\operatorname{tr}\left[\boldsymbol{V}\boldsymbol{\Lambda}^2\boldsymbol{V}^\top\boldsymbol{\Omega}\right]$. Therefore, a necessary condition for any true local minimum is that it must occur at a stationary point of the reduced problem

$$\min_{\boldsymbol{V}} \operatorname{tr}\left[\boldsymbol{V}\boldsymbol{\Lambda}^2\boldsymbol{V}^\top\boldsymbol{\Omega}^2\right], \quad \text{s.t. } \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}. \tag{11}$$

Using results from [2], it can be shown that the stationary points of (11) must occur when $\boldsymbol{V}$ is a permutation matrix, at least assuming diagonal elements of $\boldsymbol{\Lambda}^2$ and $\boldsymbol{\Omega}$ are distinct; however, a simple continuity argument can be used to extend to the general case.

Proceeding further, if $\boldsymbol{V}$ must be a permutation matrix, then at any local minimum with respect to $\boldsymbol{V}$, it must be that $\|\boldsymbol{w}_{x,k}\|_2^2 = \lambda_k^2$, where λ_k is the k-th diagonal element of $\boldsymbol{\Lambda}$. From this observation we may infer that

$$\begin{aligned}\min_{\omega_k\geq 0}\left\{\omega_k\|\boldsymbol{w}_{x,k}\|_2^2 - h^*(\omega_k)\right\} &= \min_{\omega_k\geq 0}\left\{\omega_k\lambda_k^2 - h^*(\omega_k)\right\} \\ &= \log\left|\boldsymbol{\Lambda}^2 + \gamma\boldsymbol{I}\right| \equiv \log\left|\boldsymbol{U}\boldsymbol{\Lambda}^2\boldsymbol{U}^\top + \gamma\boldsymbol{I}\right|. \end{aligned} \tag{12}$$

Since $\boldsymbol{W}_x\boldsymbol{W}_x^\top = \boldsymbol{U}\boldsymbol{\Lambda}^2\boldsymbol{U}^\top$, it then follows that any local minimum with respect to $\boldsymbol{U}$ and $\boldsymbol{\Lambda}^2$ must also be a local minimum of the revised cost

$$\widetilde{\mathcal{L}}(\theta,\phi) \triangleq \operatorname{tr}\left[\boldsymbol{\Sigma}_{gt}\left(\boldsymbol{U}\boldsymbol{\Lambda}^2\boldsymbol{U}^\top + \gamma\boldsymbol{I}\right)^{-1}\right] + \log\left|\boldsymbol{U}\boldsymbol{\Lambda}^2\boldsymbol{U}^\top + \gamma\boldsymbol{I}\right|. \tag{13}$$

It has been demonstrated in [9] that all local minima of this simplified objective are global, with $\boldsymbol{U}\boldsymbol{\Lambda}$ spanning the principal subspace of $\boldsymbol{\Sigma}_{gt}$ associated with eigenvalues larger than γ.

Additionally, by Hadamard's inequality [7] we have

$$\begin{aligned}\log\left|\boldsymbol{U}\boldsymbol{\Lambda}^2\boldsymbol{U}^\top+\gamma\boldsymbol{I}\right| &= \log\left|\tfrac{1}{\gamma}\boldsymbol{W}_x^\top\boldsymbol{W}_x+\boldsymbol{I}\right|+d\log\gamma \\ &\le \sum_k\left(1+\tfrac{1}{\gamma}\|\boldsymbol{w}_{x,k}\|_2^2\right)+d\log\gamma=\sum_k\log\left(\|\boldsymbol{w}_{x,k}\|_2^2+\gamma\right)+(d-\kappa)\log\gamma.\end{aligned} \tag{14}$$

Therefore it must also be true that $\widetilde{\mathcal{L}}(\theta,\phi)\le\mathcal{L}(\theta,\phi)$, and so these global minima must also be global minima to our original objective.

To recap, we have shown that any minimizing decoder weight matrix (local or global) must satisfy $\boldsymbol{W}_x=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{P}$, where $\boldsymbol{U}$ is orthonormal, $\boldsymbol{\Lambda}$ is diagonal PSD matrix, and $\boldsymbol{P}$ is a permutation matrix. Moreover, this $\boldsymbol{W}_x$ must span the principle subspace of $\boldsymbol{\Sigma}_{gt}$ associated with eigenvalues larger than γ. If μ_{gt} is confined to an r-dimensional affine space and $\gamma\to 0$ becomes arbitrarily small, then

$$\operatorname{rank}[\boldsymbol{W}_x]=\operatorname{rank}[\boldsymbol{\Lambda}]=\operatorname{rank}[\boldsymbol{\Sigma}_{gt}]=r, \tag{15}$$

which also implies that $\boldsymbol{W}_x$ can only have r nonzero columns. The elements of $\boldsymbol{z}$ associated with these columns will therefore not contribute to the VAE data fit term, i.e., the first term in (1). Therefore, only the remaining convex KL factors are left to dictate the the values of $\boldsymbol{\mu}_z=\boldsymbol{W}_z\boldsymbol{x}+\boldsymbol{b}_z$ and $\boldsymbol{s}^2$ along these indices. This will necessarily force the corresponding elements of $\boldsymbol{s}^2$ to one and $\boldsymbol{\mu}_z$ to zero at the optimum (rows of $\boldsymbol{W}_z$ and elements of $\boldsymbol{b}_z$ will be set to zero to make this happen for all $\boldsymbol{x}$). Consequently, $\kappa-r$ superfluous dimensions of $\boldsymbol{z}$ are effectively shut off, transmitting no information about $\boldsymbol{x}$ to the decoder.

The remaining elements of $\boldsymbol{z}$ behave quite differently and allow the reconstruction error to converge to zero as $\gamma\to 0$. Per previous arguments, $s_k^2=\left[\frac{1}{\gamma}\|\boldsymbol{w}_{x,k}\|_2^2+1\right]^{-1}$ for all k. Therefore, if $\|\boldsymbol{w}_{x,k}\|_2^2>0$ for some k, as $\gamma\to 0$ it follows that $s_k^2\to 0.$, collapsing $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ to a Dirac delta function along such a dimensions which ultimately facilitates zero reconstruction error. More specifically, let $\bar{\boldsymbol{W}}_x$ denote the nonzero columns of $\boldsymbol{W}_x$, and $\bar{\boldsymbol{W}}_z$ and $\bar{\boldsymbol{b}}_z$ the corresponding nonzero rows/elements of $\boldsymbol{W}_z$ and $\boldsymbol{b}_z$ respectively, all positioned at some minimizing solution. Then in aggregate, the above analysis implies that at any minimum,

$$\begin{aligned}&\lim_{\gamma\to 0}\int\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\|\boldsymbol{x}-\boldsymbol{W}_x\boldsymbol{z}-\boldsymbol{b}_x\|_2^2\right]\mu_{gt}(d\boldsymbol{x}) \\ &\quad=\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\left\|\boldsymbol{x}-\bar{\boldsymbol{W}}_x\left(\bar{\boldsymbol{W}}_z+\bar{\boldsymbol{b}}_z\right)-\boldsymbol{b}_x\right\|_2^2\right]\mu_{gt}(d\boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[0]\,\mu_{gt}(d\boldsymbol{x}) = 0,\end{aligned} \tag{16}$$

meaning zero reconstruction error. Finally, this ultimately guarantees that we have achieved balanced regularization at any minimum. ■

4. Proof of Theorem 2

Proof The variational upper bound is defined in (2) of the main paper. We define $\mathcal{L}(\boldsymbol{x};\theta,\phi)$ as the loss at a data point $\boldsymbol{x}$, *i.e.*

$$\mathcal{L}(\boldsymbol{x};\theta,\phi)=-\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]+\mathbb{KL}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right]. \tag{17}$$

The total loss is the integration of $\mathcal{L}(\boldsymbol{x};\theta,\phi)$ over $\boldsymbol{x}$. Further more, we denote $\mathcal{L}_{kl}(\boldsymbol{x};\theta)$ and $\mathcal{L}_{gen}(\boldsymbol{x};\theta,\phi)$ as the KL loss and the generation loss at $\boldsymbol{x}$ respectively, *i.e.*

$$\begin{aligned} \mathcal{L}_{kl}(\boldsymbol{x};\phi) &= \mathbb{KL}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right] = \sum_{i=1}^{\kappa} \mathbb{KL}\left[q_\phi(z_j|\boldsymbol{x})||p(z_j)\right], \\ &= \frac{1}{2}\sum_{j=1}^{\kappa}\left(\mu_{z,j}^2 + \sigma_{z,j}^2 - \log\sigma_{z,j}^2 - 1\right) \qquad (18) \\ \mathcal{L}_{gen}(\boldsymbol{x};\phi,\theta) &= -\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]. \qquad (19) \end{aligned}$$

The second equality in (18) holds because the covariance of $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and $p(\boldsymbol{z})$ are both diagonal. The last encoder layer and the first decoder layer are denoted as $\boldsymbol{h}_e^\rho$ and $\boldsymbol{h}_d^1$. If $\boldsymbol{w}_{\mu_z,j\cdot}^\rho = 0, \boldsymbol{w}_{\sigma_z^2,j\cdot}^\rho = 0$, then we have

$$\mu_{z,j} = \boldsymbol{w}_{\mu_z,j\cdot}^\rho \boldsymbol{h}_e^\rho = 0, \quad \sigma_{z,j}^2 = \exp\left(\boldsymbol{w}_{\sigma_z^2,j\cdot}\right) = 1, \quad q(z_j|\boldsymbol{x}) = \mathcal{N}(0,1). \qquad (20)$$

The gradient of $\mu_{z,j}$ and $\sigma_{z,j}$ from $\mathcal{L}_{kl}(\boldsymbol{x};\phi)$ becomes

$$\frac{\partial\mathcal{L}_{kl}(\boldsymbol{x};\phi)}{\partial\mu_{z,j}} = \mu_{z,j} = 0, \quad \frac{\partial\mathcal{L}_{kl}(\boldsymbol{x};\phi)}{\partial\sigma_{z,j}} = 1 - \sigma_{z,j}^{-1} = 0. \qquad (21)$$

So the gradient of $\boldsymbol{w}_{\mu_z,j\cdot}^\rho$ and $\boldsymbol{w}_{\sigma_z^2,j\cdot}^\rho$ from $\mathcal{L}_{kl}$ is

$$\frac{\partial\mathcal{L}_{kl}(\boldsymbol{x};\phi)}{\partial\boldsymbol{w}_{\mu_z,j\cdot}^\rho} = \frac{\partial\mathcal{L}_{kl}(\boldsymbol{x};\phi)}{\partial\mu_{z,j}}\boldsymbol{h}_e^{\rho\top} = 0, \qquad (22)$$

$$\frac{\partial\mathcal{L}_{kl}(\boldsymbol{x};\phi)}{\partial\boldsymbol{w}_{\sigma_z^2,j\cdot}^\rho} = \frac{\partial\mathcal{L}_{kl}(\boldsymbol{x};\phi)}{2\sigma_{z,j}\cdot\partial\sigma_{z,j}}\boldsymbol{h}_e^{\rho\top} = 0. \qquad (23)$$

Now we consider the gradient from $\mathcal{L}_{gen}(\boldsymbol{x};\theta,\phi)$. We have

$$\frac{-\partial\log p_\theta(\boldsymbol{x}|\boldsymbol{z})}{\partial z_j} = \frac{-\partial\log p_\theta(\boldsymbol{x}|\boldsymbol{z})}{\partial\boldsymbol{h}_d^1}\frac{\partial\boldsymbol{h}_d^1}{\partial z_j}. \qquad (24)$$

Since

$$\boldsymbol{h}_d^1 = \text{act}\left(\sum_{j=1}^{\kappa}\boldsymbol{w}_{\mu_x,\cdot j}^1 z_j\right), \qquad (25)$$

where act$(\cdot)$ is the activation function, we can obtain

$$\frac{\partial\boldsymbol{h}_d^1}{\partial z_j} = \text{act}'\left(\sum_{j=1}^{\kappa}\boldsymbol{w}_{\mu_x,\cdot j}^1 z_j\right)\boldsymbol{w}_{\mu_x,\cdot j}^1 = 0. \qquad (26)$$

Plugging this back into (24) gives

$$\frac{-\partial\log p_\theta(\boldsymbol{x}|\boldsymbol{z})}{\partial z_j} = 0. \qquad (27)$$

According to the chain rule, we have

$$\frac{\partial \mathcal{L}_{gen}(\boldsymbol{x};\theta,\phi)}{\partial \boldsymbol{w}^{\rho}_{\mu_z,j\cdot}} = \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{-\partial \log p_\theta(\boldsymbol{x}|\boldsymbol{z})}{\partial z_j}\frac{\partial z_j}{\partial \boldsymbol{w}^{\rho}_{\mu_z,j\cdot}}\right] = 0, \tag{28}$$

$$\frac{\partial \mathcal{L}_{gen}(\boldsymbol{x};\theta,\phi)}{\partial \boldsymbol{w}^{\rho}_{\sigma_z^2,j\cdot}} = \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{-\partial \log p_\theta(\boldsymbol{x}|\boldsymbol{z})}{\partial z_j}\frac{\partial z_j}{\partial \boldsymbol{w}^{\rho}_{\sigma_z^2,j\cdot}}\right] = 0. \tag{29}$$

After combining these two equations with (22) and (23) and then integrating over $\boldsymbol{x}$, we have

$$\frac{\partial \mathcal{L}(\theta,\phi)}{\partial \boldsymbol{w}^{\rho}_{\mu_z,j\cdot}} = 0, \tag{30}$$

$$\frac{\partial \mathcal{L}(\theta,\phi)}{\partial \boldsymbol{w}^{\rho}_{\sigma_z^2,j\cdot}} = 0. \tag{31}$$

Then we consider the gradient with respect to $\boldsymbol{w}^{1}_{\mu_x,\cdot j}$. Since $\boldsymbol{w}_{\mu_x,\cdot j}$ is part of θ, it only receives gradient from $\mathcal{L}_{gen}(\boldsymbol{x};\theta,\phi)$. So we do not need to consider the KL loss. If $\boldsymbol{w}^{1}_{\mu_x,\cdot j} = 0$, $\boldsymbol{h}^1_d = \sum_{j=1}^{\kappa} \boldsymbol{w}^{1}_{\mu_x,\cdot j} z_j$ is not related to $\boldsymbol{z}_j$. So $p_\theta(\boldsymbol{x}|\boldsymbol{z}) = p_\theta(\boldsymbol{x}|\boldsymbol{z}_{\neg j})$, where $\boldsymbol{z}_{\neg j}$ represents $\boldsymbol{z}$ without the j-th dimension. The gradient of $\boldsymbol{w}^{1}_{\mu_x,\cdot j}$ is

$$\begin{aligned}\frac{\partial \mathcal{L}_{gen}(\boldsymbol{x};\theta,\phi)}{\partial \boldsymbol{w}^{1}_{\mu_x,\cdot j}} &= \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{-\partial \log p_\theta(\boldsymbol{x}|\boldsymbol{z})}{\partial \boldsymbol{w}^{1}_{\mu_x,\cdot j}}\right] = \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{-\partial \log p_\theta(\boldsymbol{x}|\boldsymbol{z})}{\partial \boldsymbol{h}^1_d} z_j\right] \\ &= \mathbb{E}_{\boldsymbol{z}_{\neg j}\sim q(\boldsymbol{z}_{\neg j}|\boldsymbol{x})}\left[\mathbb{E}_{z_j\sim\mathcal{N}(0,1)}\left[\frac{-\partial \log p_\theta(\boldsymbol{x}|\boldsymbol{z}_{\neg j})}{\partial \boldsymbol{h}^1_d} z_j\right]\right] \\ &= \mathbb{E}_{\boldsymbol{z}_{\neg i}\sim q(\boldsymbol{z}_{\neg i}|\boldsymbol{x})}\left[\frac{-\partial \log p_\theta(\boldsymbol{x}|\boldsymbol{z}_{\neg j})}{\partial \boldsymbol{h}^1_d}\mathbb{E}_{z_j\sim\mathcal{N}(0,1)}[\boldsymbol{z}_j]\right] = 0.\end{aligned} \tag{32}$$

The integration over $\boldsymbol{x}$ should also be 0. So we obtain

$$\frac{\partial \mathcal{L}(\theta;\phi)}{\partial \boldsymbol{w}^{1}_{\mu_x,\cdot j}} = 0. \tag{33}$$

■

5. Proof of Theorem 3

Proof 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 [5, 8] 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

$$f_\theta(\boldsymbol{z}) = \theta \pi\left(z_1\right) + (1-\theta) \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \tag{34}$$

where $\theta \in \Omega \triangleq [0,1]$ is a scalar parameter, $t : \mathbb{R} \to [0,1]$ truncates its argument to the interval between zero and one and $\pi : [0,1] \to \mathcal{S} \subset [0,1]^2$ 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 $f_\theta(\boldsymbol{z})$ 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

$$z_{gt}^{(i)} \sim p\left(z_{gt}\right) \quad \text{and} \quad \boldsymbol{x}^{(i)} = \pi\left(t\left[z_{gt}^{(i)}\right]\right) \tag{35}$$

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

$$\min_\phi \mathcal{L}_h(\theta, \phi) \quad = \quad \min_{\boldsymbol{Z}} \; h\left(\tfrac{1}{dn} \sum_{i=1}^n \left\| \boldsymbol{x}^{(i)} - f_\theta\left(\boldsymbol{z}^{(i)}\right)\right\|_2^2\right) + \tfrac{1}{d} \sum_{k=1}^{\kappa} h\left(\tfrac{1}{n}\left\|\boldsymbol{z}_k\right\|_2^2\right), \tag{36}$$

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

$$\min_\phi \mathcal{L}_h(\theta = 1, \phi \in \Phi) \; = \; h(0) + \frac{1}{2}\left[h(0) + h\left(\tfrac{1}{n}\|\boldsymbol{z}_1\|_2^2\right)\right] \; \geq \; \frac{3}{2}h(0) + \frac{1}{2}h\left(C^2\right), \tag{37}$$

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 balanced regularization.

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

$$\min_{\phi} \mathcal{L}_h(\theta = 0, \phi) \ = \ h(0) + \tfrac{1}{2}\sum_{k=1}^{2} h\left(\tfrac{1}{n}\|\boldsymbol{z}_k\|_2^2\right) \ \leq \ h(0) + h\left(\epsilon^2\right). \tag{38}$$

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

$$\min_{\phi} \mathcal{L}_h(\theta = 1, \phi \in \Phi) \ \geq \ \tfrac{1}{2} \ > \ \min_{\phi} \mathcal{L}_h(\theta = 0, \phi) \ \approx \ 0, \tag{39}$$

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 exhibit balanced regularization.

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] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[2] R.W. Brockett. Least squares matching problems. *Linear Algebra and Its Applications*, pages 761–777, 1989.

[3] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

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

[5] 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.

[6] Bin Dai and David Wipf. Diagnosing and enhancing gaussian VAE models. *Advances in Neural Information Processing Systems, Bayesian Deep Learning Workshop*, 2018.

[7] D.J.H. Garling. *A Journey into Linear Analysis.* Cambridge University Press, 2007.

[8] 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.

[9] Michael Tipping and Christopher Bishop. Probabilistic principal component analysis. *J. Royal Statistical Society, Series B*, 61(3):611–622, 1999.