Marginalization is not Marginal: No Bad VAE Local Minima when Learning Optimal Sparse Representations

David Wipf [1]

Abstract

Although the variational autoencoder (VAE) represents a widely-used deep generative model, the underlying energy function when applied to continuous data remains poorly understood. In fact, most prior theoretical analysis has assumed a simplified affine decoder such that the model collapses to probabilistic PCA, a restricted regime whereby existing classical algorithms can also be trivially applied to guarantee globally optimal solutions. To push our understanding into more complex, practically-relevant settings, this paper instead adopts a deceptively sophisticated single-layer decoder that nonetheless allows the VAE to address the fundamental challenge of learning optimally sparse representations of continuous data originating from popular multiple-response regression models. In doing so, we can then examine VAE properties within the non-trivial context of solving difficult, NP-hard inverse problems. More specifically, we prove rigorous conditions which guarantee that any minimum of the VAE energy (local or global) will produce the optimally sparse latent representation, meaning zero reconstruction error using a minimal number of active latent dimensions. This is ultimately possible because VAE *marginalization* over the latent posterior selectively smooths away bad local minima as has been conjectured but not actually proven in prior work. We then discuss how equivalent-capacity deterministic autoencoders, even with appropriate sparsity-promoting regularization of the latent space, maintain bad local minima that do not correspond with such parsimonious representations. Overall, these results serve to elucidate key properties of the VAE loss surface relative to finding low-dimensional structure in data.

[1]Amazon Web Services (a portion of this work was initially conducted while the author was with Microsoft Research). Correspondence to: David Wipf <davidwipf@gmail.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1. Introduction

The variational autoencoder (VAE) (Kingma & Welling, 2014; Rezende et al., 2014) is a commonly-used latent variable model targeting, among other things, data $\boldsymbol{x} \in \mathbb{R}^d$ assumed to possess some unknown low-dimensional structure. This is reflected in the trainable marginalized distribution $p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$ that underpins canonical VAE models, where $\theta \in \Theta$ are flexible parameters and $\boldsymbol{z} \in \mathbb{R}^\kappa$ represents unobservable latent factors. Low-dimensional structure is either explicitly enforced when $\kappa < d$, or implicitly whenever the dimensions of $\boldsymbol{z}$ that significantly influence $\boldsymbol{x}$ are limited in number. For example, the maximum likelihood value of θ could be such that only a handful of entries in $\boldsymbol{z}$ actually impact the distribution $p_\theta(\boldsymbol{x}|\boldsymbol{z})$.

In this vein, given a set of n training points $\boldsymbol{X} = [\boldsymbol{x}_{:1}, \ldots, \boldsymbol{x}_{:n}] \in \mathbb{R}^{d\times n}$, we may equivalently choose to minimize the negative log-likelihood expression $\frac{1}{n}\sum_i -\log\left[p_\theta\left(\boldsymbol{x}_{:i}\right)\right]$. However, because the marginalization over $\boldsymbol{z}$ required to produce $p_\theta(\boldsymbol{x}_{:i})$ is often intractable, we may instead minimize the variational upper bound on the negative log-likelihood given by $\mathcal{L}(\theta,\phi) \triangleq$

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

with equality iff $q_\phi(\boldsymbol{z}|\boldsymbol{x}_{:i}) = p_\theta(\boldsymbol{z}|\boldsymbol{x}_{:i})$ for each datapoint i. The trainable parameters ϕ define the variational distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x}_{:i})$ that is designed to approximate the true (but generally intractable) $p_\theta(\boldsymbol{z}|\boldsymbol{x})$. Excluding the KL-divergence-based regularization factor, the first term in (1) amounts to a form of stochastic reconstruction loss that mirrors the basic structure of an autoencoder (AE) but modified to include marginalization over the latent space. In fact, if $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ collapses to a Dirac delta function this term directly defaults to a deterministic AE (more on this later).

Given our present focus on continuous data, we adopt the convention that the so-called *decoder* and *encoder* distributions satisfy

$$p_\theta\left(\boldsymbol{x}|\boldsymbol{z}\right) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_x, \gamma\boldsymbol{I}) \;\text{ and }\; q_\phi\left(\boldsymbol{z}|\boldsymbol{x}\right) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) \tag{2}$$

along with prior $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$, where $\gamma > 0$ is a scalar variance that may be trained or held fixed. The Gaussian moment functions $\boldsymbol{\mu}_x \equiv \boldsymbol{\mu}_x\left(\boldsymbol{z};\theta\right)$, $\boldsymbol{\mu}_z \equiv \boldsymbol{\mu}_z\left(\boldsymbol{x};\phi\right)$, and

$\boldsymbol{\Sigma}_z \equiv \boldsymbol{\Sigma}_z(\boldsymbol{x};\phi)$ are generally instantiated via neural network layers, with input $\boldsymbol{x}$ for the encoder functions, and $\boldsymbol{z}$ for the decoder. The VAE objective (1) can then be optimized over $\{\theta,\phi\}$ using SGD and a reparameterization trick (Kingma & Welling, 2014; Rezende et al., 2014). Once trained, the VAE model can be applied to generating new samples via the prior and decoder distributions, or to producing compact latent representations of an arbitrary training (or test) data point via $\boldsymbol{z} = \boldsymbol{\mu}_z(\boldsymbol{x};\phi)$. As discussed below, the latter can often be useful for downstream tasks not directly related to generative modeling.

While frequently deployed in practical settings, either for generating samples or representation learning, relatively little analysis of the VAE energy function is currently available. As will detailed further in Section 2, essentially all performance guarantees relate to simplified scenarios where the VAE decoder is affine and the model collapses to simple probabilistic PCA (Tipping & Bishop, 1999). Moreover, these results do not actually differentiate the capabilities of a VAE versus the corresponding deterministic AE with the same affine decoder; both are equally capable of learning the principal subspace of the training data (Dai et al., 2018; Kunin et al., 2019; Lucas et al., 2019), and care must be taken in extrapolating from such results.

To push past these limitations, this paper instead considers how VAE minima (local or global) can align with the fundamental challenge of learning optimally sparse, low-dimensional representations of data as formalized in Section 3, and later specialized to a widely-used multiple-response regression setting in Section 4.1 using a deceptively sophisticated single-layer decoder. From this vantage point, we can then examine VAE properties within the non-trivial context of solving difficult, NP-hard inverse problems that existing conventional algorithms typically fail to solve, i.e., there is no analogue to PCA for addressing such problems.

More specifically, in Section 4.2 we prove rigorous conditions which guarantee that any minimum of the VAE energy (local or global) will produce a representation with asymptotically negligible reconstruction error using a minimal number of active latent dimensions. This is possible because VAE marginalization over the posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ during training effectively smooths away all bad local minima that rely on an excessive number of latent dimensions to reduce the reconstruction error. Section 4.3 then discusses how any possible deterministic AE (within a broadly defined class), even with appropriate sparsity-promoting regularization of the latent space and equivalent representational capacity, maintains a (possibly combinatorial) number of bad local and/or global minima that do *not* correspond with optimally parsimonious representations. We also analyze the impact of diagonalizing $\boldsymbol{\Sigma}_z$ on the possibility of spurious minima in Sections 4.4 and 4.5. And finally, we provide corroborating simulation results in Section 5.

Overall, the pursuit of optimal sparse representations will serve as a useful lens with which to quantitatively evaluate the local minima profile of VAEs and highlight critical advantages over deterministic AE models that are not necessarily directly related to generating samples. Given these considerations, our primary contributions distill as:

- We provide the first formal proof of a VAE model whereby there are no bad local minima and all global minima optimize a non-trivial inverse problem for which classical alternative approaches do not exist with an equivalent guarantee. This problem involves producing an optimal sparse representation of training data, which in general is NP-hard but becomes uniquely aligned with VAE optima in certain conditions. This is ultimately possible because VAE *marginalization* over the latent posterior fills in suboptimal minima while preserving optimally sparse global solutions. Such a selective VAE smoothing effect has been conjectured as a possibility (Dai et al., 2018); however, no rigorous proof has thus far been provided.
- We prove that no possible analogous AE with equivalent capacity can achieve something similar. In doing so, we elucidate the first clear-cut differentiation between the performance of VAEs (with marginalization) and AE models (without) in the specific context of learning optimal low-dimensional representations.

2. Related Work

Analysis of VAE Local Minima: Of particular interest herein is the analysis of situations where all local minima can be explicitly characterized in term of some optimality criteria. There are primarily two extremes that have been previously considered for this purpose, differentiated by the complexity of the VAE decoder. First, if the VAE decoder mean function is sufficiently simple and unstructured, specifically a basic affine transformation for $\boldsymbol{\mu}_x$, then it has been shown, e.g. (Dai et al., 2019; Lucas et al., 2019), that all minima of the VAE objective produce principal components of the data across a broad class of encoder functions (see Section 3 below for a more formal treatment). Such results may also loosely extend to some broader contexts (Rolinek et al., 2019). In contrast, at the other extreme of an arbitrarily complex decoder/encoder pair, where both $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_z$ are treated as infinite capacity functions with no constraints, the calculus of variations can be applied to easily obtain explicit expressions for any critical point of the VAE loss (Rezende & Viola, 2018). The downside here though is that the optimal solution essentially involves the model memorizing the training data, with all probability mass consumed by delta functions placed at each training point, and optimal sparse representations are not meaningful

in this context. And with unconstrained decoder complexity, it has been shown that just a single active latent dimension is sufficient for perfect reconstructions of any finite training set (Dai et al., 2018).

Analysis of VAE Global Minima: In addition to the work mentioned above, some effort has been made to more narrowly explore characteristics of VAE global minima, under conditions whereby local minimizers are known to likely exist, but whose cardinality and properties are not easily quantifiable or of central interest. We may further subdivide research in this category between demonstrations of both desirable and undesirable properties of global optima. With respect to the latter, several pathologies have recently been noted. For example, as addressed in (Mattei & Frellsen, 2018), if instead of $\mathbf{\Sigma}_x = \gamma \boldsymbol{I}$ as assumed herein the decoder covariance is a flexible, data-dependent function, then even relatively simple VAE models can have an energy that is unbounded from below by simply memorizing a single training data point, assigning infinite density to this point and small but nonzero density elsewhere. Other problematic VAE global solutions may prioritize learning a good inference model at the expense of learning a good generative model (Yacoby et al., 2020; 2022). For example, this situation may occur if the training data and VAE parameterization are such that the true data likelihood is within the decoder capacity but the latent posterior is poorly modeled by the encoder.

In contrast, desirable conditions have been derived (Dai et al., 2018) whereby VAE models with flexible, data-specific decoder variances are capable of reproducing NP-hard robust PCA (Candès et al., 2011; Chandrasekaran et al., 2011) solutions. This work also hypothesizes that VAE marginalization has the ability to selectively smooth away bad local minima while preserving desirable global optima; however, no rigorous proof is provided whereby this phenomena is guaranteed to occur. Additionally, (Dai & Wipf, 2019) derives technical conditions for perfect recovery of ground-truth distributions in the limit of infinite training data while complementary VAE optimization trajectory analysis is contained in (Koehler et al., 2022). Other notable VAE analysis work pointed out by reviewers includes (Damm et al., 2023; Shekhovtsov et al., 2022; Zietlow et al., 2021).

Bayesian Precursors to the VAE: As already alluded to above, probabilistic PCA (Tipping & Bishop, 1999) can be viewed as a simplified precursor to more flexible VAE architectures. Related probabilistic frameworks for structured regression, such as sparse Bayesian learning (SBL) (Bishop & Tipping, 2000; Tipping, 2001), also share commonalities with the VAE, at least in the more narrow context of finding optimal sparse representations (as opposed to explicitly generating new samples from a target distribution of interest). And in certain cases the global and/or local minima from such models may have optimality guarantees (Aravkin et al., 2014; Prasad & Murthy, 2012; Wipf et al., 2011; 2015) of the general sort we would ideally like to establish for the VAE under relevant/analogous settings.

Loss Surface of (Deep) Linear Networks: The exploration of VAE models with restrictive assumptions placed on the decoder and/or encoder structure closely follows the established tradition of analyzing the complex loss surface of deep networks with linear layers and/or i.i.d. random activation patterns (Choromanska et al., 2015a;b; Goodfellow et al., 2016; Kawaguchi, 2016; Saxe et al., 2014). Likewise for deterministic AE models with linear encoder/decoder pairs, whereby it has been shown that all critical points are associated with PCA directions (Kunin et al., 2019).

3. Optimal Sparse (Lossless) Representations

To begin we must precisely define and motivate what type of low-dimensional or sparse representations will be considered optimal. At a high level, we require a means of quantifying the most parsimonious latent representation of the training data that nonetheless allows us to obtain high-quality reconstructions when passed through a given class of decoder networks. As a representative example, for data lying on a low-dimensional linear subspace, the corresponding optimal sparse representation obtainable via a linear decoder could plausibly be defined by the smallest subspace containing all or most of the data variance, i.e., the standard PCA solution. With this conception in mind, we borrow the following definition from (Dai et al., 2021):

Definition 3.1. An autoencoder-based architecture (VAE or otherwise) with decoder $\boldsymbol{\mu}_x(\cdot;\theta)$, constraint $\theta \in \Theta$, and arbitrary encoder $\boldsymbol{\mu}_z$ component produces an *optimal sparse representation* of a training set $\boldsymbol{X}$ w.r.t. Θ if the following two conditions simultaneously hold:

(i) The reconstruction error is zero, meaning $\frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_{:i} - \boldsymbol{\mu}_x[\boldsymbol{\mu}_z(\boldsymbol{x}_{:i};\phi);\theta]\|_2^2 = 0$.

(ii) Conditioned on achieving perfect reconstructions per criteria (i) above, the number of latent dimensions such that $\mu_z(\boldsymbol{x}_{:i};\phi)_j = 0$ for all i is maximal across any $\theta \in \Theta$ and any encoder function $\boldsymbol{\mu}_z$. A j-th latent dimension so-defined provides no benefit in reducing the reconstruction error and could be removed.

In its reliance on lossless reconstructions, Definition 3.1 may at first glance seem to involve an overly restrictive assumption. And yet, as pointed out in (Dai et al., 2021), many celebrated under-determined inverse problems have been formulated as the search for lossless reconstructions of observed measurements subject to some optimal measure of parsimony (Candès & Recht, 2009; Donoho & Elad, 2003; Sun et al., 2018). And natural images, a popular input to VAE models, have a very low intrinsic dimension relative to

the high-dimensional pixel space, which facilitates nearly exact reconstruction using low-dimensional representations (Pope et al., 2021). Please see (Dai et al., 2021) for further details regarding the relevance of Definition 3.1 to typical VAE and VAE-adjacent application domains including feature extraction (Bengio et al., 2013; Ng, 2011), compression (Ballé et al., 2018; Donoho, 2006; Minnen et al., 2018), manifold learning (Silva et al., 2006), corruption removal (Dai et al., 2018), or the generation of realistic samples (Dai & Wipf, 2019).

Producing Optimal Representations Using VAEs: Even while invoking a stochastic encoder, it has been shown in (Dai & Wipf, 2019) that as $\gamma \to 0$, VAE global minima can produce representations that asymptotically align with the conditions of Definition 3.1. This is possible because, along a superfluous latent dimension j, the KL regularization within the VAE energy favors $q_\phi\left(z_j|\boldsymbol{x}\right) \to \mathcal{N}(0,1)$, i.e., the posterior is converted to (zero-mean) white noise that can be subsequently blocked/ignored by the decoder. And along informative/active latent dimensions we have $\sigma_z\left(\boldsymbol{x}_{:i};\phi\right)_j \to 0$ as $\gamma \to 0$, $\forall i$. These diverging behaviors allow the VAE global minima to produce reconstructions satisfying $\sum_{i=1}^n \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}_{:i})}\left[\|\boldsymbol{x}_{:i} - \boldsymbol{\mu}_x\left[\boldsymbol{z};\theta\right]\|_2^2\right] \to$

$$\sum_{i=1}^n \|\boldsymbol{x}_{:i} - \boldsymbol{\mu}_x\left[\boldsymbol{\mu}_z\left(\boldsymbol{x}_{:i};\phi\right);\theta\right]\|_2^2 \;\to\; 0 \qquad (3)$$

to asymptotically satisfy criteria (i) of Definition 3.1, all while utilizing the fewest number of informative latent dimensions to achieve criteria (ii). Of course *actually finding such a global optima* while avoiding a potentially large constellation of bad local minima may still be challenging.

VAE/PCA Equivalence w.r.t. Optimal Sparsity: For later context in presenting our main results, it is useful to interpret existing VAE analysis of minima, akin to (Dai et al., 2019; Lucas et al., 2019), within the framework of optimal sparse representations as follows:

Lemma 3.2. *Assume a Gaussian VAE model of continuous data defined by (2), where $\boldsymbol{\mu}_x = \boldsymbol{W}_x\boldsymbol{z} + \boldsymbol{b}_x$ for some weight matrix $\boldsymbol{W}_x$ and bias vector $\boldsymbol{b}_x$; similarly $\boldsymbol{\mu}_z = \boldsymbol{W}_z\boldsymbol{x} + \boldsymbol{b}_z$ while $\boldsymbol{\Sigma}_z = diag[\boldsymbol{s}]^2$, where $\boldsymbol{s}$ is an arbitrary parameter vector independent of $\boldsymbol{x}$. Then for any fixed value of γ, all local minima of the resulting VAE objective with respect to the remaining parameters $\{\boldsymbol{W}_x, \boldsymbol{b}_x, \boldsymbol{W}_z, \boldsymbol{b}_z, \boldsymbol{s}\}$ are also global minima. Additionally, these global minima will produce optimally sparse representations per Definition 3.1 in the limit $\gamma \to 0$.*

All proofs are deferred to the appendices. Intuitively, an optimal sparse representation occurs because each possible local/global optima defines 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 $\boldsymbol{s}$ tend to zero as $\gamma \to 0$, while the corresponding elements of $\boldsymbol{\mu}_z$ convey the information about $\boldsymbol{x}$ (i.e., active, non-random dimensions) needed for losslessly reconstructing the data. Elsewhere the latent variances will be set to one while the means will equal zero, indicating that no useful information about $\boldsymbol{x}$ is being transferred. Of course the model described by Lemma 3.2 is obviously a simplified version of the VAE; however, it nonetheless represents the only realistic scenario (e.g., excluding infinite capacity decoders that memorize the training data) whereby thus far the full constellation of VAE local minima has been characterized as alluded to in Section 2.

4. New VAE Optimal Sparsity Guarantees

As we have seen, Lemma 3.2 provides sufficient conditions for equating all VAE local minima with global, optimally sparse representations akin to the principal subspace containing the training data. But clearly this same task can be accomplished using various PCA instantiations instead of an AE or VAE model. Therefore, the underlying value of this type of analysis is in elucidating properties of the respective energy functions, *not in actually solving an important practical problem per se*.

In contrast, we now consider a more challenging regime where, without further assumptions, the underlying recovery of an optimally sparse representation is actually NP-hard, i.e., no straightforward alternative procedure exists. Within this more practically-relevant setting, we derive results whereby any VAE minima (local or global) is uniquely associated with the optimally sparse representation. We then discuss how analogous AEs within a broad class, even with appropriate sparsity-promoting regularization, maintain a (potentially combinatorial) number of bad local/global minima that do not correspond with optimal sparse representations. Later, we relax the requirement of lossless reconstructions and consider diagonalized VAE posterior covariances.

4.1. The Simultaneous Sparse Approximation Problem

To begin, we consider n training points aggregated as $\boldsymbol{X} = [\boldsymbol{x}_{:1}, \ldots, \boldsymbol{x}_{:n}] \in \mathbb{R}^{d\times n}$ that we assume were generated via $\boldsymbol{X} = \boldsymbol{\Phi}\boldsymbol{U}_0$. In this expression, $\boldsymbol{\Phi} \in \mathbb{R}^{d\times\kappa}$ represents a known dictionary of κ basis/feature vectors, $\boldsymbol{U}_0 \in \mathbb{R}^{\kappa\times n}$ denotes a row-sparse matrix of ground-truth latent factors, and we allow for $\kappa > d$. More precisely, we assume $\boldsymbol{U}_0 \in$

$$\arg\min_{\boldsymbol{U}} \rho(\boldsymbol{U}),\ \text{s.t. } \boldsymbol{X} = \boldsymbol{\Phi}\boldsymbol{U},\ \rho(\boldsymbol{U}) \triangleq \sum_{j=1}^{\kappa} \mathbb{1}\left[\|\boldsymbol{u}_{j:}\| \neq 0\right], \qquad (4)$$

where $\mathbb{1}\left[\|\boldsymbol{u}_{j:}\| \neq 0\right]$ denotes an indicator function on the j-th row norm. This formulation implies that $\boldsymbol{U}_0$ is maximally row-sparse, or equivalently, that $\boldsymbol{X}$ is formed using an expansion involving the fewest number of columns/features

from $\boldsymbol{\Phi}$. This observation naturally aligns $\boldsymbol{U}_0$ with our definition of an optimal sparse representation. Note that any feasible solution to (4) exhibits zero reconstruction error consistent with criteria (i) of Definition 3.1, and minimizing $\rho(\boldsymbol{U})$ is tantamount to satisfying criteria (ii).

Solving this type of combinatorial problem, commonly referred to as simultaneous sparse approximation or multiple-response sparse regression (Cotter et al., 2005; Tropp, 2006), is foundational to many diverse application domains including multi-task learning (Ji et al., 2009; Ling et al., 2013; Wakin et al., 2006; Zeng et al., 2011), manifold learning (Silva et al., 2006), array processing (Choi et al., 2017; Malioutov et al., 2005; Thoota & Murthy, 2022; Wipf et al., 2015), and functional brain imaging (Bannier et al., 2021; Bhutada et al., 2022; Cai et al., 2018). Unfortunately though, the underlying objective (4) is NP-hard, with a combinatorial number of suboptimal local minima. Convex relaxations of the indicator function in $\rho(\boldsymbol{U})$ have been proposed for practical feasibility, but the resulting modified objective will often fail to recover $\boldsymbol{U}_0$ when columns of $\boldsymbol{\Phi}$ display significant correlation structure, i.e., off-diagonal elements of $\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ are relatively large (Tropp, 2006). Hence unlike probabilistic PCA, there is no readily-available classical algorithm for guaranteeing that the correct solution can always be found (beyond infeasible combinatorial search).

4.2. VAEs and Simultaneous Sparse Approximation

While perhaps not obvious at first glance, we will demonstrate that the VAE energy function under the appropriate encoder/decoder parameterizations is particularly well-suited to solving (4). As in (2), we adopt a Gaussian decoder given that our data is continuous, and select $\boldsymbol{\Sigma}_x = \gamma\boldsymbol{I}$ as before, with $\gamma > 0$. However, for the decoder mean, we choose $\boldsymbol{\mu}_x = \boldsymbol{\Phi}\text{diag}\left[\boldsymbol{w}_x\right]\boldsymbol{z}$, where $\boldsymbol{w}_x \in \mathbb{R}^\kappa$ denotes a parameter vector to learn. For the encoder we adopt $\boldsymbol{\mu}_z = \boldsymbol{W}_z\boldsymbol{x}$ (no bias term), but a full covariance $\boldsymbol{\Sigma}_z = \boldsymbol{S}\boldsymbol{S}^\top$, where $\boldsymbol{S} \in \mathbb{R}^{\kappa\times\kappa}$ is an arbitrary matrix independent of $\boldsymbol{x}$. Perhaps counterintuitively, this parameterization is sufficiently flexible for addressing the difficulty in solving (4). Stated differently, more complex, nonlinear encoder structures (or a bias term) do not provide any advantage in optimizing the overall VAE loss given the specified decoder assumptions.

With these stipulations in place, the VAE energy from (1) as applied to the training set $\boldsymbol{X}$ reduces to

$$\mathcal{L}(\theta,\phi) = \sum_{i=1}^{n}\Big(\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}_{:i})}\left[\tfrac{1}{\gamma}\left\|\boldsymbol{x}_{:i} - \boldsymbol{\Phi}\text{diag}\left[\boldsymbol{w}_x\right]\boldsymbol{z}\right\|_2^2\right] \quad (5)$$
$$+ d\log\gamma + \text{tr}\left[\boldsymbol{S}\boldsymbol{S}^\top\right] - \log\left|\boldsymbol{S}\boldsymbol{S}^\top\right| + \|\boldsymbol{W}_z\boldsymbol{x}_{:i}\|_2^2\Big),$$

with $\theta = \{\boldsymbol{w}_x, \gamma\}$ and $\phi = \{\boldsymbol{W}_z, \boldsymbol{S}\}$.

To proceed with the analysis of (5) as applied to solving (4), for subtle technical reasons mentioned in (Wipf et al., 2015) we assume that

$$\boldsymbol{\Phi} = \boldsymbol{\Xi} + \epsilon\boldsymbol{\Delta}, \quad (6)$$

where $\boldsymbol{\Xi}$ is any arbitrary matrix, $\epsilon > 0$ is an arbitrarily small constant, and $\boldsymbol{\Delta}$ is matrix formed with entries drawn i.i.d. from any distribution with a properly-defined density function. By constructing $\boldsymbol{\Phi}$ in this way, we ensure that each $d \times d$ sub-matrix of $\boldsymbol{\Phi}$ is almost surely full rank. The latter is equivalent to the requirement that spark$[\boldsymbol{\Phi}] = d+1$, where spark$[\boldsymbol{\Phi}]$ is defined as the smallest number of linearly dependent columns in $\boldsymbol{\Phi}$ (Donoho & Elad, 2003).

We are now prepared to describe conditions whereby the objective (5) will be such that *any* minimum, global or local, will produce an optimal sparse representation capable of recovering the ground-truth $\boldsymbol{U}_0$. For convenience, we let $\pi_{(j)}\left[\boldsymbol{U}\right]$ denote the value of the j-th largest ℓ_2 row-norm of a matrix $\boldsymbol{U}$. We then have the following:

Theorem 4.1. *Let $\{\theta^*, \phi^*\} \equiv \{\boldsymbol{w}_x^*, \boldsymbol{W}_z^*, \boldsymbol{S}^*\}$ denote any local minimum of (5) in the limit as $\gamma \to 0$. Then there exists a set of $d-2$ constants $\nu_j \in (0, 1]$ such that for any $\boldsymbol{X} = \boldsymbol{\Phi}\widehat{\boldsymbol{U}}$ generated with $\boldsymbol{\Phi}$ satisfying (6), $\rho(\widehat{\boldsymbol{U}}) < d$, and $\pi_{(j+1)}[\widehat{\boldsymbol{U}}] \leq \nu_j\pi_{(j)}[\widehat{\boldsymbol{U}}]$ for all $j = 1, \ldots, d-2$, we have with probability one that*

- *(i) $\{\boldsymbol{w}_x^*, \boldsymbol{W}_z^*, \boldsymbol{S}^*\}$ is also a global minimum of (5),*
- *(ii) $\sum_{i=1}^n \left\|\boldsymbol{x}_{:i} - \boldsymbol{\mu}_x\left[\boldsymbol{\mu}_z\left(\boldsymbol{x}_{:i};\phi^*\right);\theta^*\right]\right\|_2^2 = 0$, i.e., perfect reconstructions,*
- *(iii) $\text{diag}\left[\boldsymbol{w}_x^*\right]\boldsymbol{\mu}_z(\boldsymbol{x}_{:i};\phi^*) = \widehat{\boldsymbol{u}}_{:i}$ for all i, i.e., the VAE parameters allow us to analytically compute the unknown $\widehat{\boldsymbol{U}}$, and*
- *(iv) the problem (4) has a unique optimally sparse solution $\boldsymbol{U}_0 = \widehat{\boldsymbol{U}}$.*

Note that the reason we consider the limit $\gamma \to 0$, rather than simply $\gamma = 0$, is for technical reasons related the ill-defined nature of Gaussian distributions with zero variance. Moreover, as detailed in the proof, for any fixed γ, the optimal solution w.r.t. the remaining parameters satisfies

$$\boldsymbol{\mu}_z\left(\boldsymbol{x}_{:i};\phi^*\right) = \text{diag}\left[\boldsymbol{w}_x^*\right]\boldsymbol{\Phi}^\top\left(\boldsymbol{\Phi}\text{diag}\left[\boldsymbol{w}_x^*\right]^2\boldsymbol{\Phi}^\top + \gamma\boldsymbol{I}\right)^{-1}\boldsymbol{x}_{:i}. \quad (7)$$

And for data lying on a $r < d$ dimensional subspace, the optimal diag$[\boldsymbol{w}_x^*]$ will be rank r such that the required inverse is not actually defined when $\gamma = 0$. However, when we instead take the limit $\gamma \to 0$, we obtain the well-defined limiting encoder mean $\boldsymbol{\mu}_z(\boldsymbol{x}_{:i};\phi) \to \left(\boldsymbol{\Phi}\text{diag}\left[\boldsymbol{w}_x^*\right]\right)^\dagger\boldsymbol{x}_{:i}$ for all i, where $(\cdot)^\dagger$ denotes the Moore-Penrose pseudo-inverse. In contrast, the optimal decoder mean satisfies $\boldsymbol{\mu}_x(\boldsymbol{z};\phi^*) = \boldsymbol{\Phi}\text{diag}\left[\boldsymbol{w}_x^*\right]\boldsymbol{z}$ for all $\boldsymbol{z} \in \mathbb{R}^\kappa$, and hence is always well-defined.

But conceptually speaking, this result can be loosely viewed as implying that there exists a $\gamma' > 0$ sufficiently small such

that for any nonzero $\gamma < \gamma'$ the basic idea holds. Additionally, achieving the global VAE optimum actually requires $\gamma \to 0$ assuming sufficient capacity to reconstruct the training data (Dai & Wipf, 2019), so this is not a restrictive stipulation per the present context.

Interpretation of Theorem 4.1: Beyond these minor technical considerations, Theorem 4.1 defines a scenario whereby any local minimum of the VAE loss surface is also a global minimum (part i) that computes an optimal sparse representation, meaning a solution that perfectly reconstructs $\boldsymbol{X}$ (part ii) using the fewest number of informative latent factors (parts iii and iv). In particular, if the row norms of the ground-truth coefficient matrix are of sufficiently different scales, then any VAE minima will be optimal, which entails recovering all nonzero ground-truth latent factors no matter how small some of them might be. This is remarkable given that the original loss from (4), which defines the canonical simultaneous sparse approximation problem, will *necessarily* have a combinatorial number of suboptimal local minima, including under the stated conditions of Theorem 4.1.

We also emphasize that Theorem 4.1 only involves *sufficient* conditions for obtaining a loss surface devoid of suboptimal minima, but these conditions are not necessary. Even if appropriate row-norm scaling as mitigated by the constants $\{\nu_j\}_{j=1}^{d-2}$ is not exactly present, the VAE is still likely to produce good results in broader regimes whereby many, even if not all, suboptimal minima have been smoothed away via marginalization. But the overall impact of Theorem 4.1 is best appreciated when contrasted with an analogous deterministic AE as will be explored next.

4.3. Comparisons with an Equivalent-Capacity AE

Consider the AE objective $\mathcal{L}_{AE}(\boldsymbol{w}_x, \boldsymbol{W}_z) =$

$$\sum_{i=1}^{n} \tfrac{1}{\gamma} \left\| \boldsymbol{x}_{:i} - \boldsymbol{\Phi}\text{diag}\left[\boldsymbol{w}_x\right] \boldsymbol{z}_{:i} \right\|_2^2 + \|\boldsymbol{w}_x\|_2^2 + \sum_{j=1}^{\kappa} h\left(\tfrac{1}{n}\|\boldsymbol{z}_{j:}\|_2^2\right), \tag{8}$$

s.t. $\boldsymbol{z}_{:i} = \boldsymbol{W}_z \boldsymbol{x}_{:i}$, where h represents a concave, non-decreasing function of the squared row norms. Inspired by (Ng, 2011), this expression is nothing more than an AE model with encoder and decoder matching the VAE mean functions from (5), plus a quadratic penalty on the decoder weights (akin to weight decay as is frequently used in practice) and a sparsity-promoting penalty applied to the latent codes (Chen et al., 2017; Fan & Li, 2001; Palmer et al., 2006); both terms are needed to avoid problematic scaling degeneracies.[1] Hence, (8) has an equivalent modeling capacity as the VAE, but with a commonsense, deterministic regularization scheme. Additionally, in the limit $\gamma \to 0$, the data-dependent term is effectively converted to an extra constraint $\boldsymbol{x}_{:i} = \boldsymbol{\Phi}\text{diag}\left[\boldsymbol{w}_x\right] \boldsymbol{z}_{:i}$ while the remaining factors stay the same. This implies that as $\gamma \to 0$, minimizers of (8) are equivalent to minimizers of $\mathcal{L}_{AE}(\boldsymbol{w}_x, \boldsymbol{W}_z) \equiv$

$$\|\boldsymbol{w}_x\|_2^2 + \sum_{j=1}^{\kappa} h\left(\tfrac{1}{n}\|\boldsymbol{z}_{j:}\|_2^2\right), \;\; \text{s.t.} \;\; \begin{array}{c} \boldsymbol{z}_{:i} = \boldsymbol{W}_z \boldsymbol{x}_{:i}, \;\; \forall i \\ \boldsymbol{x}_{:i} = \boldsymbol{\Phi}\text{diag}\left[\boldsymbol{w}_x\right] \boldsymbol{z}_{:i} \end{array}, \tag{9}$$

in which any feasible solution produces perfect reconstructions and the corresponding latent codes are penalized to favor sparsity.

Interestingly though, solving (8) in the stated limit does not enjoy the same theoretical guarantees as the VAE.

Theorem 4.2. *With probability one there will exist dictionaries $\boldsymbol{\Phi}$ formed via (6) that satisfy the following: For any set of $d-2$ scaling constants $\nu_j \in (0, 1]$, there will always be coefficients $\widehat{\boldsymbol{U}}$ consistent with the requirements of Theorem 4.1, such that any AE model in the form of (8) will have minima (either local and/or global) that do not produce an optimal sparse representation in the limit $\gamma \to 0$.*

In brief, this result indicates that even with the stated data generation restrictions in place that ensure optimality of the VAE loss surface vis-à-vis Theorem 4.1, the equivalent-capacity AE model can still exhibit poor minima that fail to produce optimal sparse representations or recover ground-truth $\boldsymbol{U}_0$. Moreover, this negative result is independent of the specific functional form for the latent-space penalty h applied within (8). So indeed the VAE maintains a distinct advantage in this regard. Note also that, although we assume a quadratic penalty on the decoder weights $\boldsymbol{w}_x$, it is straightforward to extend Theorem 4.2 to arbitrary penalties of the form $\sum_{j=1}^{\kappa} f[(w_x)_j]$. So indeed this is a quite general result that is not a nuanced consequence of this particular choice.

4.4. Impact of Diagonalizing $\boldsymbol{\Sigma}_z$

In the previous sections we quantified an explicit, non-trivial situation whereby the VAE model maintains a distinct advantage over the analogous deterministic AE. However, this analysis did not restrict $\boldsymbol{\Sigma}_z$ to be a diagonal form as is sometimes assumed in practical VAE implementations. We now investigate the impact of this diagonalization via the following straightforward result:

Lemma 4.3. *Suppose that we replace $\boldsymbol{\Sigma}_z = \boldsymbol{S}\boldsymbol{S}^\top$ with $\boldsymbol{\Sigma}_z = diag[\boldsymbol{s}]^2$ in (5), where $\boldsymbol{s}^2 \in \mathbb{R}_+^{\kappa}$ defines the encoder variances. Additionally, assume that $h(\cdot) = \log(\cdot)$ in (8). Then in the limit as $\gamma \to 0$, the objectives (5) and (8) are equivalent (excluding constant terms).*

Lemma 4.3 indicates that once we have diagonalized $\boldsymbol{\Sigma}_z$, the VAE no longer really has any advantage within the specified

[1] For example, without the penalty on $\boldsymbol{w}_x$, we could simply push $\boldsymbol{z}_{:i}$ to zero for all i to reduce the h term, while proportionally increasing $\boldsymbol{w}_x$ towards infinity to maintain a perfect data fit. However, w.l.o.g. we do not require an explicit trade-off parameter applied to $\|\boldsymbol{w}_x\|_2^2$ given the $1/\gamma$ factor applied to the data term and the fact that h can implicitly absorb any desirable scaling.

context over its AE counterpart as defined by (8). This is because the restriction $\boldsymbol{\Sigma}_z = \text{diag}[\boldsymbol{s}]^2$ will inexorably introduce a combinatorial constellation of suboptimal local minima in all but the most trivial of situations, and there is in fact no scenario in which the diagonalized VAE satisfies Theorem 4.1. This is quite unlike the probabilistic PCA model analyzed in (Dai et al., 2019; Kunin et al., 2019; Lucas et al., 2019) (which we referenced in Sections 2 and 3) whereby a diagonal covariance fragments the loss surface into a number of distinct minima, but each separate basin retains global optimality.

4.5. Extension to Lossy Reconstructions

For equivalent capacity models and an appropriate choice for h, our results above have suggested that the relative ability to smooth away or avoid bad local minima follows

$$\text{full covariance VAE} \;>\; \text{diagonal covariance VAE} \;=\; \text{AE}$$

in the limit $\gamma \to 0$. However, there exists additional nuance when we relax the strict requirement of zero reconstruction error and instead compare VAE and AE models with γ set to an arbitrary but fixed value greater than zero. While it is not possible to exhaustively characterize the degree to which local minima can be smoothed away via VAE marginalization for any $\gamma > 0$, we can at least consider an admittedly simplified scenario that is emblematic of behavior we have empirically observed in a broader context (the experiments in Section 5 provide a representative example). In doing so we apply the notation $\boldsymbol{w}_{x,\setminus i}$ to describe $\boldsymbol{w}_x$ with the i-th element removed. Similarly, $\boldsymbol{W}_{z,\setminus i}$ refers to $\boldsymbol{W}_z$ with the i-th row removed.

Theorem 4.4. *If we fix γ and $\boldsymbol{w}_{x,\setminus i}$ for any $i \in \{1,\dots,\kappa\}$, then all local minima are global when we optimize (5) over the remaining parameters $\{w_{x,i}, \boldsymbol{W}_z, \boldsymbol{S}\}$.*

Corollary 4.5. *Theorem 4.4 does not hold if we replace $\boldsymbol{S}$ with $\boldsymbol{s}$, i.e., diagonalizing the encoder covariance can potentially introduce bad (i.e., non-global) local minima within the specified context. However, if we instead fix γ, $\boldsymbol{w}_{x,\setminus i}$ and $\boldsymbol{W}_{z,\setminus i}$ for any $i \in \{1,\dots,\kappa\}$, then all local minima are now global when we optimize (5) over the remaining parameters $\{w_{x,i}, \boldsymbol{w}_{z,i:}, \boldsymbol{S}\}$ or $\{w_{x,i}, \boldsymbol{w}_{z,i:}, \boldsymbol{s}\}$, i.e., diagonalization can no longer introduce any bad local minima.*

And yet while the full covariance VAE still holds some advantage per Theorem 4.4 and Corollary 4.5, the inferior diagonalized counterpart is nonetheless still superior to the AE in the following sense:

Corollary 4.6. *A result analogous to Corollary 4.5 does not hold for the AE model from (8) when we choose $h(\cdot) = \log(\cdot)$; likewise for any h such that the composite $h\left([\cdot]^2\right)$ is a concave, non-decreasing, nonlinear function.*

Note that if our goal is obtaining optimally sparse latent representations in general conditions (exact/lossless or otherwise), then it can be shown that a concave nonlinear function (i.e., not simultaneously convex as would be the case if $h\left([\cdot]^2\right)$ were linear) on the latent row norms is generally required (Cotter et al., 2005).[2] Hence Corollary 4.6 basically ensures that whenever we apply an optimal penalty for maximal sparsity, we cannot rule out bad AE local minima even within the restricted context described above. This implies that the diagonalized VAE may indeed still maintain some advantage over the analogous AE. To summarize then, our results suggest that the revised extent of local minima smoothing in inexact/lossy situations with $\gamma > 0$ is more accurately characterized as

$$\text{full covariance VAE} \;>\; \text{diagonal covariance VAE} \;>\; \text{AE},$$

meaning some degree of VAE local minima smoothing is preserved despite the potentially deleterious effects of diagonalizing $\boldsymbol{\Sigma}_z$. We will investigate this and other theoretically-motivated observations via experiments described next.

4.6. Additional Perspectives

We close this section by providing additional context with respect to the uniqueness of VAE solutions, prior analysis of VAE global minima, and the broader positioning of VAEs between alternative convex and non-convex approaches to finding optimally sparse representations of data.

Uniqueness of VAE Solutions: While we demonstrated conditions in Section 4.2 whereby any VAE minima of (5) (local or global) will necessarily produce a unique, maximally sparse representation, we did not specify that the VAE solution itself is unique, leaving open the possibility of multiple VAE optima with equivalent recovery guarantees. And in fact, a simple observation reveals that indeed VAE minima need not be unique because of an intrinsic invariance to sign permutations in the following sense: The value of (5) is invariant to the transformations $\boldsymbol{w}_x \to \boldsymbol{D}\boldsymbol{w}_x$, $\boldsymbol{W}_z \to \boldsymbol{D}\boldsymbol{W}_z$, and $\boldsymbol{S} \to \boldsymbol{D}\boldsymbol{S}$, where $\boldsymbol{D}$ is a diagonal matrix with diagonal elements given by 1 or -1. (This sign ambiguity also naturally extends to more general VAE models as well, since we can always multiply the encoder network by a diagonal sign matrix, and then compensate with another diagonal sign matrix on the decoder side, while the KL term remains unchanged.) Practically speaking though, this sign ambiguity is inconsequential since the learned VAE predictor for the latent ground-truth sparse $\boldsymbol{U}_0$ is given by $\text{diag}[\boldsymbol{w}_x]\boldsymbol{W}_z\boldsymbol{X} = \text{diag}[\boldsymbol{D}\boldsymbol{w}_x]\boldsymbol{D}\boldsymbol{W}_z\boldsymbol{X}$. Therefore if all bad local minima have been smoothed away by VAE marginalization (as we have shown is provably possible), the global minima that remain, while not strictly unique, are nonetheless all effectively equivalent in terms of finding

[2] A strictly concave function (Rockafellar, 1970) can be viewed as a special case of a concave nonlinear function; similarly for the selection $h\left([x]^2\right) = \mathbb{1}\left[x \neq 0\right]$, which leads to an ℓ_0-norm-based regularization factor.

ground-truth optimal sparse representations.

Global Minima Analysis from (Zheng et al., 2023): Prior work from (Zheng et al., 2023) has derived quite general conditions whereby VAE global optima are guaranteed to produce optimal sparse representations, a conjecture originally proposed in (Dai & Wipf, 2019). Critically though, none of these prior results address the challenging issue of ruling out bad local optima. Indeed even standard AE models can be readily designed to have global minima that align with optimal sparse latent codes, akin to Theorem 1 from (Zheng et al., 2023). But of course the unresolved difficulty for both VAE and AE models alike remains *avoiding bad local optima to actually find these desirable global solutions*. And it is with respect to the latter that our key contributions in this section lie. This notion is further contextualized when we consider the following.

Convex vs Non-Convex Methods: More broadly, prior work on finding sparse representations (or related low-dimensional structure) generally categorize as follows:

(i) A *convex* energy function is adopted such that the global minimum is relatively easy to find, but this global minima may not align with the optimal sparse representation. For example, classical group-lasso-based approaches (Malioutov et al., 2005; Yuan & Lin, 2006) to approximating (4) fall into this category.

(ii) A *non-convex* loss is chosen that, unlike the convex case, may more generally maintain this desired alignment (meaning if we manage to find the global optimum we obtain maximal sparsity); however, this comes at the cost of introducing a combinatorial explosion of local minima (so actually finding the global optimum becomes difficult). Examples include AE models with appropriate non-convex latent-space regularization as described in Section 4.3, as well as simpler non-convex group-lasso-like models involving ℓ_p pseudo-norms with $p < 1$ (Cotter et al., 2005).

Our message is that the VAE can accomplish something remarkably different: *It can preserve a global minima anchored to the maximally sparse solution (like prior non-convex methods) while provably smoothing away bad local optima (akin to convex relaxations)*. We believe this to be a fundamental insight into the capabilities of VAE models that has not been previously acknowledged in the literature.

5. Empirical Validation

Although this work is primarily a theoretical contribution, the analysis from Section 4 can nonetheless be strengthened by an empirical demonstration of the natural ability of VAE marginalization to selectively smooth away bad local minima in securing optimally sparse representations.

5.1. Experiments with Verifiable Ground-Truth

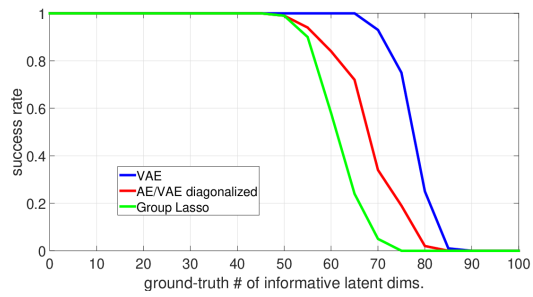
We explore the simultaneous sparse approximation problem in such a way that we can have access to ground-truth representations to facilitate optimality comparisons. To this end we generate data via $\boldsymbol{X} = \boldsymbol{\Phi}\boldsymbol{U}_0$, where $\boldsymbol{\Phi}$ and $\boldsymbol{U}_0$ are drawn randomly and stored as ground-truth. Please see the appendices for details regarding how these data were created for each experiment, as well as additional results solving a neuromagnetic inverse problem involving $\boldsymbol{\Phi}$ chosen as a MEG leadfield matrix (Sarvas, 1987).

Results are displayed in Figure 2, where each subplot includes the success percentage recovering $\boldsymbol{U}_0$ (y-axis) as a function of the ground-truth number of nonzero/informative dimensions of $\boldsymbol{U}_0$ (x-axis). We show performance curves for the VAE from (5) along with the analogous AE from (8); for the latter h is chosen to be the $\log$ function for the most direct head-to-head comparison, i.e., with this choice, the diagonalized VAE and AE energy functions converge to one another as $\gamma \to 0$ per Lemma 4.3. We also compare against a group-lasso-based solution, which represents a popular convex alternative to solving (4). This is tantamount to replacing the ideal non-convex penalty ρ with the convex mixed norm $\|\boldsymbol{U}\|_{1,2} \triangleq \sum_{j=1}^{\kappa} \|\boldsymbol{u}_{j:}\|_2$. In all cases results are averaged over 100 independent trials.

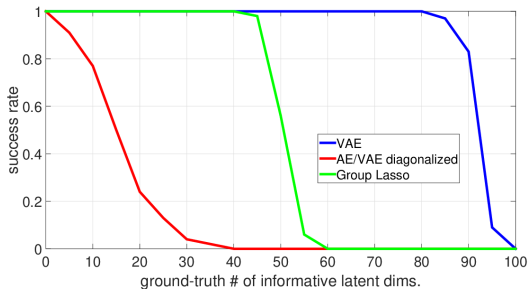
First, in Figure 1(a) we display results where nonzero rows of $\boldsymbol{U}_0$ have been scaled to unit ℓ_2-norm. This scenario deviates from the assumptions of Theorem 4.1 and provides a baseline for the minimal expected amount of VAE local optima smoothing. And yet even in this regime the full/non-diagonal VAE outperforms the AE. Meanwhile, the diagonalized VAE performs equivalently to the AE since we have chosen $\gamma = 10^{-10}$ for all models (and hence per Lemma 4.3 they will essentially be equivalent). Additionally, both VAE and AE models outperform the convex group lasso.

In contrast, for Figure 1(b) we conduct the same experiment only now rows of $\boldsymbol{U}_0$ have been rescaled to loosely approximate the favorable conditions predicted by Theorem 4.1. Consistent with expectations, we observe that the VAE performance improves considerably, with nearly perfect performance all the way up to the theoretical limit of any possible algorithm; as detailed in the appendices, this occurs when $\rho(\boldsymbol{U}_0) = 100$. Meanwhile, the other methods do not actually benefit from this rescaling, with AE performance actually degrading significantly.

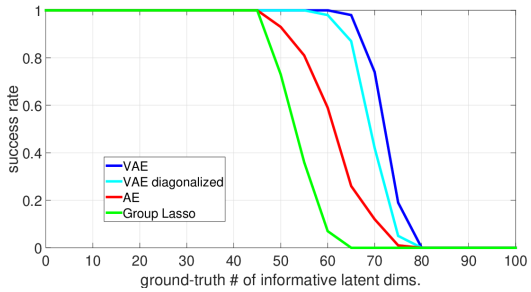
And finally, we repeat the basic experiment from Figure 1(a) with the inclusion of 20dB additive Gaussian noise. Results are presented in Figure 1(c), where λ is set to the true noise variance for all VAE/AE models; for subtle technical reasons (and to improve performance) λ is set to the square-root of the noise variance for the group lasso estimator. Of particular note, we observe that the diagonalized VAE curve



(a)



(b)



(c)

Figure 1. Success rates producing optimal sparse representations. (*a*) Unit-norm nonzero rows; (*b*) Highly scaled nonzero row norms that loosely approximate the conditions of Theorem 4.1; (*c*) Repeat of plot *a* with 20 dB additive noise.

is now superior to the AE, which suggests that even the more modest degree of local minima smoothing described in Section 4.5 may provide some advantage when the full posterior covariance is not used.

5.2. Extension to More Complex Decoder Models

As suggested by an ICML reviewer during the initial evaluation of our paper, it could be informative to examine the degree to which the local minima smoothing that provably exists in certain single-layer decoder settings informs behavior in more general multi-layer regimes. Motivated by this possibility, we consider additional tests involving VAEs applied to MNIST data with standard ResNet blocks forming the encoder and decoder architectures. We then explore two equivalent-capacity VAE variants that differ only in the last layer of the respective encoder networks: (i) a standard VAE with diagonal encoder variance, and (ii) an analogous full covariance alternative. Per the arguments presented in Section 4 and empirically verified in Section 5.1, we expect that the VAE with full covariance is more likely to avoid bad local minimizers that do not produce optimal sparsity.

Full details are deferred to a revised version of (Zheng et al., 2023), where an experimental setup is described that naturally accommodates head-to-head MNIST comparisons between VAE models seeking to recover optimally sparse representations. However, we nonetheless summarize the main conclusions here. Simply put, these new experiments inspired by our analysis unequivocally demonstrate that the full-covariance VAE produces a *lower* reconstruction error while simultaneously relying on a *fewer* number of nonzero/active latent dimensions, i.e., dimensions that are not set to the prior in accordance with Definition 3.1. This outcome closely aligns with the predictions of our theory.

6. Conclusion

While the VAE remains a celebrated deep generative model capable of producing high-quality samples when outfitted with appropriate decoder/encoder architectures, it also exhibits close ties with classical approaches for finding low-dimensional structure in high-dimensional data, as well as solving challenging underdetermined inverse problems. With respect to the latter, we have demonstrated non-trivial conditions whereby marginalization over the latent posterior allows the VAE to selectively smooth away bad local minima while retaining global optima anchored at optimal sparse representations of the training data. Equivalent capacity AE models, which lack such marginalization, enjoy no such optimality guarantees, and when sharing an identical decoder and encoder, and analogous regularization factors, will generally possess a combinatorial number of bad local minima. Additionally, we have also demonstrated that diagonalizing the VAE encoder covariance, which mutes the impact of marginalization, can in fact introduce bad local minima, although to a lesser extent than deterministic AEs.

Interestingly, these results are in contradistinction to prior analysis of simpler, structure-free affine decoder models whereby all local minima exactly align with probabilistic PCA solutions regardless of whether or not $\boldsymbol{\Sigma}_z$ is diagonal. As typical VAE use-cases often involve highly-structured decoders, our results therefore suggest that the consequences of diagonalized covariances may be worth reconsidering if the computational budget actually allows for handling full covariances (or approximations thereof). Overall though, we believe that the insights provided herein extend our knowledge of the VAE loss surface and complement prior analyses, build bridges with more traditional dimensionality reduction methodologies, and suggest broader usage regimes for VAE models, e.g., beyond generating samples from $p_\theta(\boldsymbol{x})$.

References

Aravkin, A., Burke, J. V., Chiuso, A., and Pillonetto, G. Convex vs non-convex estimators for regression and sparse estimation: The mean squared error properties of ARD and GLasso. *Journal of Machine Learning Research*, 15: 217–252, 2014.

Ballé, J., Minnen, D., Singh, S., Hwang, S. J., and Johnston, N. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018.

Bannier, P.-A., Bertrand, Q., Salmon, J., and Gramfort, A. Electromagnetic neural source imaging under sparsity constraints with sure-based hyperparameter tuning. *arXiv preprint arXiv:2112.12178*, 2021.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8):1798–1828, 2013.

Bhutada, A. S., Cai, C., Mizuiri, D., Findlay, A., Chen, J., Tay, A., Kirsch, H. E., and Nagarajan, S. S. Clinical validation of the champagne algorithm for evoked response source localization in magnetoencephalography. *Brain topography*, 35:96–107, 2022.

Bishop, C. and Tipping, M. Variational relevance vector machines. *Uncertainty in Artificial Intelligence*, 2000.

Cai, C., Sekihara, K., and Nagarajan, S. Hierarchical multiscale Bayesian algorithm for robust MEG/EEG source reconstruction. *NeuroImage*, 183:698–715, 2018.

Candès, E. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Candès, E., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM*, 58(2), May 2011.

Chandrasekaran, V., Sanghavi, S., Parrilo, P., and Willsky, A. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal of Optimization*, 21(3), June 2011.

Chen, Y., Ge, D., Wang, M., Wang, Z., Ye, Y., and Yin, H. Strong NP-hardness for sparse optimization with concave penalty functions. *International Conference on Machine Learning*, 2017.

Choi, J. W., Shim, B., Ding, Y., Rao, B., and Kim, D. I. Compressed sensing for wireless communications: Useful tips and tricks. *IEEE Communications Surveys & Tutorials*, 19(3):1527–1550, 2017.

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

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

Cotter, S., Rao, B., Engan, K., and Kreutz-Delgado, K. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Transactions on Signal Processing*, 53(7):2477–2488, April 2005.

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

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*, 19:1–42, 2018.

Dai, B., Wang, Y., Aston, J., Hua, G., and Wipf, D. Hidden talents of the variational autoencoder. *arXiv preprint arXiv:1706.05148*, 2019.

Dai, B., Wenliang, L., and Wipf, D. On the value of infinite gradients in variational autoencoder models. *Advances in Neural Information Processing Systems*, 2021.

Damm, S., Forster, D., Velychko, D., Dai, Z., Fischer, A., and Lücke, J. The ELBO of variational autoencoders converges to a sum of entropies. *International Conference on Artificial Intelligence and Statistics*, 2023.

Donoho, D. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

Donoho, D. and Elad, M. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100 (5):2197–2202, 2003.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

Ji, S., Dunson, D., and Carin, L. Multi-task compressive sensing. *IEEE Transactions on Signal Processing*, 57(1): 92–106, Jan 2009.

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

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

Koehler, F., Mehta, V., Zhou, C., and Risteski, A. Variational autoencoders in the presence of low-dimensional data: Landscape and implicit bias. *International Conference on Learning Representations*, 2022.

Kunin, D., Bloom, J., Goeva, A., and Seed, C. Loss landscapes of regularized linear autoencoders. *International Conference on Machine Learning*, 2019.

Kuznetsova, A., Nurislamova, Y., and Ossadtchi, A. Modified covariance beamformer for solving MEG inverse problem in the environment with correlated sources. *NeuroImage*, 228:117677, 2021.

Ling, Q., Wen, Z., and Yin, W. Decentralized jointly sparse optimization by reweighted ℓ_q minimization. *IEEE Transactions on Signal Processing*, 61(5):1165–1170, 2013.

Lucas, J., Tucker, G., Grosse, R., and Norouzi, M. DonFLt blame the ELBO! A linear VAE perspective on posterior collapse. *Advances in Neural Information Processing Systems*, 2019.

Malioutov, D., Çetin, M., and Willsky, A. Sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE Transactions Signal Processing*, 53 (8):3010–3022, 2005.

Mattei, P.-A. and Frellsen, J. Leveraging the exact likelihood of deep latent variables models. *Advances in Neural Information Processing Systems*, 2018.

Minnen, D., Ballé, J., and Toderici, G. Joint autoregressive and hierarchical priors for learned image compression. *Advances in Neural Information Processing Systems*, 2018.

Ng, A. Sparse autoencoder. *CS294A Lecture notes*, 72 (2011):1–19, 2011.

Palmer, J., Wipf, D., Kreutz-Delgado, K., and Rao, B. Variational EM algorithms for non-Gaussian latent variable models. *Advances in Neural Information Processing Systems*, 2006.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. *International Conference on Learning Representations*, 2021.

Prasad, R. and Murthy, C. Cramér-Rao-type bounds for sparse Bayesian learning. *IEEE Transactions on Signal Processing*, 61(3):622–632, 2012.

Rezende, D. J. and Viola, F. Taming VAEs. *arXiv preprint arXiv:1810.00597*, 2018.

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

Rockafellar, R. *Convex Analysis*. Princeton University Press, 1970.

Rolinek, M., Zietlow, D., and Martius, G. Variational autoencoders pursue PCA directions (by accident). *Computer Vision and Pattern Recognition*, 2019.

Sarvas, J. Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. *Physics in Medicine & Biology*, 32:11–22, 1987.

Saxe, A., McClelland, J., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *International Conference on Learning Representations*, 2014.

Shekhovtsov, A., Schlesinger, D., and Flach, B. VAE approximation error: ELBO and exponential families. *International Conference on Learning Representations*, 2022.

Silva, J., Marques, J., and Lemos, J. Selecting landmark points for sparse manifold learning. *Advances in Neural Information Processing Systems*, 2006.

Sun, J., Qu, Q., and Wright, J. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18 (5):1131–1198, 2018.

Thoota, S. S. and Murthy, C. R. Massive MIMO-OFDM systems with low resolution ADCs: Cramér-Rao bound, sparse channel estimation, and soft symbol decoding. *IEEE Transactions on Signal Processing*, March 2022.

Tipping, M. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

Tipping, M. and Bishop, C. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61(3):611–622, 1999.

Tropp, J. Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. *Signal Processing*, 86:589–602, April 2006.

Tropp, J., Gilbert, A., and Strauss, M. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Processing*, 86:572–588, April 2006.

Wakin, M., Duarte, M., Sarvotham, S., Baron, D., and Baraniuk, R. Recovery of jointly sparse signals from a few random projections. *Advances in Neural Information Processing Systems*, 2006.

Wipf, D. and Rao, B. An empirical Bayesian strategy for solving the simultaneous sparse approximation problem. *IEEE Transactions on Signal Processing*, 55(7), July 2007.

Wipf, D. and Zhang, H. Revisiting Bayesian blind deconvolution. *Journal of Machine Learning Research*, 15: 3775–3814, 2014.

Wipf, D., Rao, B., and Nagarajan, S. Latent variable Bayesian models for promoting sparsity. *IEEE Transactions on Information Theory*, 57(9), Sept 2011.

Wipf, D., Yun, J.-M., and Ling, Q. Augmented Bayesian compressive sensing. *Data Compression Conference*, 2015.

Yacoby, Y., Pan, W., and Doshi-Velez, F. Characterizing and avoiding problematic global optima of variational autoencoders. *Symposium on Advances in Approximate Bayesian Inference*, 2020.

Yacoby, Y., Pan, W., and Doshi-Velez, F. Failure modes of variational autoencoders and their effects on downstream tasks. *arXiv preprint arXiv:2007.07124*, 2022.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

Yuille, A. and Rangarajan, A. The concave-convex procedure (CCCP). *Advances in Neural Information Processing Systems*, 2001.

Zeng, F., Li, C., and Tian, Z. Distributed compressive spectrum sensing in cooperative multihop cognitive networks. *IEEE Journal of Selected Topics in Signal Processing*, 5 (2):37–48, 2011.

Zheng, Y., He, T., Qiu, Y., and Wipf, D. Learning manifold dimensions with conditional variational autoencoders. *arXiv preprint arXiv:2302.11756*, 2023.

Zietlow, D., Rolinek, M., and Martius, G. Demystifying inductive biases for (β-)VAE based architectures. *International Conference on Machine Learning*, 2021.

A. Details of Empirical Validation and an Additional Neuroimaging Example

In this section we provide more details regarding the setup of our experiments in Section 5 of the main paper. We also include an additional neuroimaging example that demonstrates additional VAE advantages.

Basic Experimental Design: We first create $\boldsymbol{\Phi}$ by drawing elements i.i.d. from a standardized Gaussian distribution. We then scale each column to have unit ℓ_2 norm as is customary in many/most applications. A ground-truth latent code matrix $\boldsymbol{U}_0$ is created with r nonzero rows placed at indeces drawn from a uniform distribution. The nonzero submatrix of $\boldsymbol{U}_0$ formed by concatenating these rows is assigned the value $\sum_{t=1}^{\tau} \boldsymbol{a}\boldsymbol{b}^\top$. Here $\boldsymbol{a} \in \mathbb{R}^r$ and $\boldsymbol{b} \in \mathbb{R}^n$ are drawn i.i.d. from a standardized Gaussian, and if we choose $\tau < r < n$, then $\boldsymbol{U}_0$ will exhibit correlation structure along the nonzero rows consistent with many application domains (Kuznetsova et al., 2021). We also scale the nonzero rows of $\boldsymbol{U}_0$ to have unit ℓ_2 norm. This is to contrast with later experiments described below where we rescale the rows such that the norms are approximately consistent with the conditions of Theorem 4.1 in the main paper. And finally, we compute that actual observed data via $\boldsymbol{X} = \boldsymbol{\Phi}\boldsymbol{U}_0$. It can be shown that per this construction, $\boldsymbol{U}_0$ will almost surely provide the optimal sparse representation of the data, i.e., with probability one there will not exist another feasible $\boldsymbol{U}'$ such that $\rho(\boldsymbol{U}') \leq \rho(\boldsymbol{U}_0)$.

Given access to the optimal ground-truth representation, we can execute various algorithms provided with $\boldsymbol{X}$ and $\boldsymbol{\Phi}$ and compare success rates in recovering $\boldsymbol{U}_0$. We test the VAE from (M.5) along with the analogous AE from (M.8); for the latter h is chosen to be the $\log$ function for the most direct head-to-head comparison, i.e., with this choice, the diagonalized VAE and AE energy functions converge to one another as $\gamma \to 0$ per Lemma 4.3. Although we could implement either algorithm using SGD, for faster convergence we instead apply a simple majorization-minimization approach to exploit the problem-specific structure (Cotter et al., 2005; Wipf & Rao, 2007) and allow for stable training with arbitrarily small values of γ; unless otherwise specified, we chose $\gamma = 10^{-10}$ for both VAE and AE models. We also compare against a Group-Lasso-based solution, which represents a popular convex alternative to solving (M.4). This is tantamount to replacing the ℓ_0 row-norm with the convex mixed norm $\|\boldsymbol{U}\|_{1,2} \triangleq \sum_{j=1}^{\kappa} \|\boldsymbol{u}_{j:}\|_2$.

We iterate this experimental procedure as r, the number of nonzero rows, varies from 0 to 100. In all cases we use $d = 100$ and $\kappa = 200$, although the basic results and subsequent conclusions are similar across different problem sizes. For each value of r, we conduct 100 independent trials and display the average success rate recovering the ground-truth $\boldsymbol{U}_0$ in Figure 2(a). The respective curves indicate that the full/non-diagonal VAE is superior to the AE. Given that $\gamma \approx 0$, the latter is effectively equivalent to the diagonalized version of the VAE as described in Section 4.4. Additionally, both VAE and AE models outperform the convex Group Lasso.

Modifications that Approximate the Conditions of Theorem 4.1: We have thus far tested under conditions which are explicitly *counter* to the stipulations of Theorem 4.1 given that the ground-truth nonzero rows are all of unit norm. We now repeat the experiment described above, but with $\boldsymbol{U}_0 \to \text{diag}[\boldsymbol{\nu}]\boldsymbol{U}_0$, were $\boldsymbol{\nu} \in \mathbb{R}^{\kappa}$ has elements ν_i drawn from a heavily skewed distribution favoring values with significantly different scales. Specifically, each ν_i is sampled i.i.d. from the approximate Jeffreys prior $p(\nu_i) = -1/[2\log(a)\nu_i]$ for $\nu_i \in [a, 1/a]$ with range parameter $a \in (0, 1)$, and $p(\nu_i) = 0$ otherwise. Note that this distribution behaves like the scale-invariant Jeffreys prior within the range $[a, 1/a]$; for example, if $a = 0.01$, the probability mass assigned to the range $[0.1, 1]$ would equal the mass between $[1, 10]$, etc. We choose $a = 0.01$ for all experiments to loosely approximate the row-scaling assumption required by Theorem 4.1. Consistent with expectations, Figure 2(b) demonstrates that the VAE performance improves considerably, with nearly perfect performance all the way up to the theoretical limit of any possible algorithm when $r = d = 100$. Meanwhile, the other methods do not actually benefit from this rescaling, with AE performance actually degrading significantly.

Inclusion of Dictionary formed from Real-World MEG Leadfield Matrix: The performance of most classical approaches to solving the simultaneous sparse approximation problem is heavily dependent on the correlation structure among the columns of $\boldsymbol{\Phi}$. Generally speaking, it is now well-established that the more correlated these columns (meaning $\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ has significant off-diagonal energy), the more difficult it is to recover $\boldsymbol{U}_0$ with existing methods (Tropp et al., 2006; Tropp, 2006). However, our analysis suggests that the local-minima smoothing capabilities of the VAE may still persist even when strong dictionary correlations are present. To examine this possibility, we consider the MEG source localization problem that involves estimating sparse neural currents within the brain using sensors placed near the surface of the scalp. The effective dictionary or forward model is referred to as the MEG leadfield, which at a high level can be viewed as a mapping from the electromagnetic (EM) activity within κ brain voxels to d sensors placed near the scalp surface. Computed using Maxwell's equations and a spherical shell head model (Sarvas, 1987), the resulting Φ is characterized by highly correlated columns, as brain voxels in a local cortical patch tend to project similar EM signals to the scalp sensors. We repeat the basic recovery experiment using such an MEG leadfield $\boldsymbol{\Phi}$, with results displayed in Figure 2(c). As expected, the Group Lasso

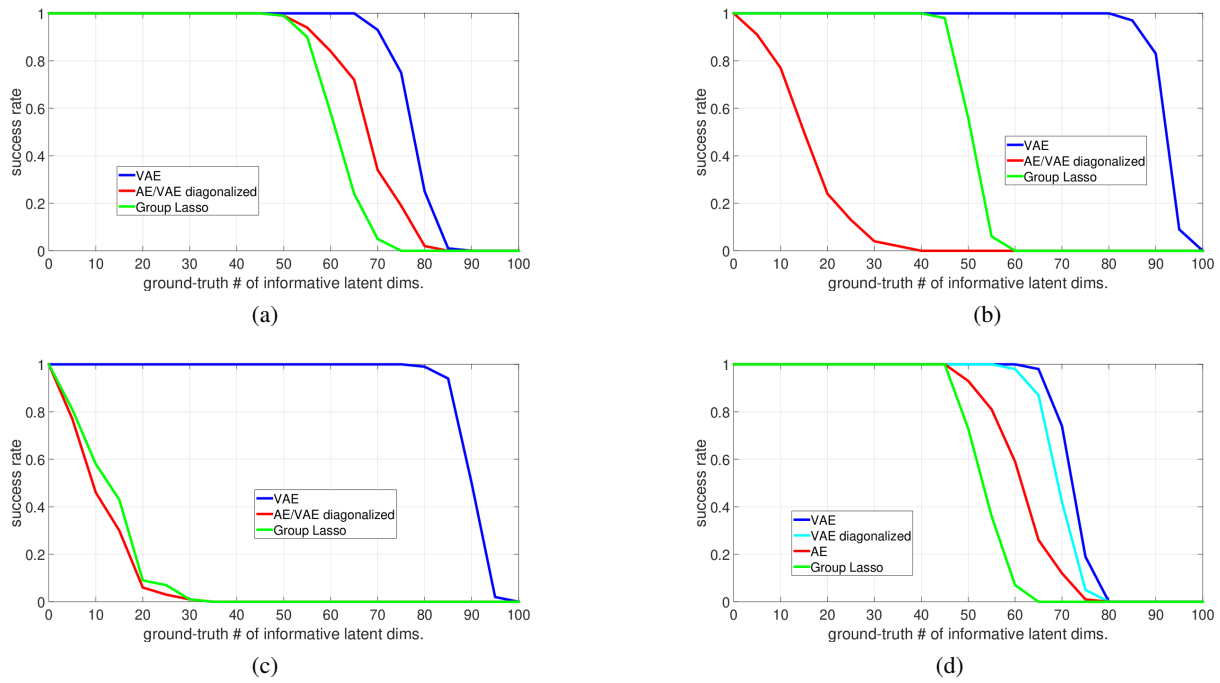


Figure 2. Success rates producing optimal sparse representations. (*a*) Unit-norm nonzero rows; (*b*) Highly scaled nonzero row norms; (*c*) Results using MEG leadfield matrix with correlated columns; (*d*) Repeat of plot *a* with 20 dB additive noise.

and AE performance drops off precipitously (the diagonalized VAE as well since it is equivalent to the AE per Lemma 4.3). In contrast though, the full VAE maintains a high success rate owing to fewer bad local minima as suggested by the theory.

Noisy Experiment to Illustrate Theorem 4.4 and Corollaries 4.5 and 4.6: Finally, we repeat the basic experiment from Figure 2(a) with the inclusion of 20dB additive Gaussian noise. Results are presented in Figure 2(d), where λ is set to the true noise variance for all VAE/AE models; for subtle technical reasons (and to improve performance) λ is set to the square-root of the noise variance for the Group Lasso estimator. Of particular note, we observe that the diagonalized VAE curve is now superior to the AE, which suggests that even the more modest degree of local minima smoothing described in Section 4.5 may provide some advantage when the full posterior covariance is not used.

B. Proof of Lemma 3.2

For simplicity we assume that the column mean of $\boldsymbol{X}$ is zero; if this were not the case then it is straightforward to show the optimal $\boldsymbol{b}_x$ will equal the column mean such that its influence can be subsequently ignored. An optimal sparse representation under the stated assumptions of Lemma 3.2 and Definition 3.1 will then be such that

$$\begin{aligned} \boldsymbol{\mu}_z\left(\boldsymbol{x};\phi^*\right) &= \boldsymbol{W}_x^*\boldsymbol{x} \\ \boldsymbol{\mu}_x\left(\boldsymbol{z};\theta^*\right) &= \boldsymbol{W}_z^*\boldsymbol{z} \\ \boldsymbol{x}_{:i} &= \boldsymbol{\mu}_x\left[\boldsymbol{\mu}_z\left(\boldsymbol{x}_{:i};\phi^*\right);\theta^*\right] = \boldsymbol{W}_x^*\boldsymbol{W}_z^*\boldsymbol{x}_{:i} \;\; \forall i \\ \rho\left(\left[\boldsymbol{W}_x^*\right]^\top\right) &= \rho\left(\boldsymbol{W}_z^*\right) = \text{rank}[\boldsymbol{X}]. \end{aligned} \tag{10}$$

Furthermore, from (Dai et al., 2019), it follows that any local minimum of the VAE loss will satisfy

$$\boldsymbol{\mu}_z\left(\boldsymbol{x}_{:i};\phi^*\right) = \boldsymbol{W}_z^*\boldsymbol{x}_{:i} = \left(\boldsymbol{W}_x^*\right)^\top\left[\boldsymbol{W}_x^*(\boldsymbol{W}_x^*)^\top + \gamma\boldsymbol{I}\right]^{-1}\boldsymbol{x}_{:i}, \tag{11}$$

where span$[\boldsymbol{W}_x^*]$ and $\rho\left(\left[\boldsymbol{W}_x^*\right]^\top\right)$ are equal to the span and cardinality respectively of the singular vectors of $\boldsymbol{X}$ associated with singular values greater than $\sqrt{\gamma}$. Additionally, we have that

$$\lim_{\gamma\to 0}\left(\boldsymbol{W}_x^*\right)^\top\left[\boldsymbol{W}_x^*(\boldsymbol{W}_x^*)^\top + \gamma\boldsymbol{I}\right]^{-1}\boldsymbol{x}_{:i} = \left(\boldsymbol{W}_x^*\right)^\dagger\boldsymbol{x}_{:i}, \tag{12}$$

where $(\cdot)^\dagger$ denotes the Moore-Penrose pseudoinverse. From this expression it then follows that $\rho\left([\boldsymbol{W}_x^*]^\top\right) = \rho\left(\boldsymbol{W}_z^*\right) = \text{rank}[\boldsymbol{X}]$ and $\boldsymbol{X} = \boldsymbol{W}_x^*(\boldsymbol{W}_x^*)^\dagger \boldsymbol{X}$ such that all the conditions of (10) are satisfied.

C. Proof of Theorem 4.1

We first optimize over the encoder parameters to arrive at a condensed loss, which is subsequently analyzed w.r.t. the decoder parameters, the latter occupying the bulk of the proof. We then combine pieces to show the four conclusions of Theorem 4.1.

C.1. Optimizing Away Encoder Parameters

The objective (5) immediately reduces to

$$\begin{aligned}\mathcal{L}(\theta,\phi) \;\; = \;\; & \sum_{i=1}^n \left(\left[\tfrac{1}{\gamma} \left\| \boldsymbol{x}_{:i} - \boldsymbol{\Phi}\text{diag}\left[\boldsymbol{w}_x\right] \boldsymbol{W}_z \boldsymbol{x}_{:i} \right\|_2^2 \right] + \tfrac{1}{\gamma}\text{tr}\left[\text{diag}\left[\boldsymbol{w}_x\right] \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \text{diag}\left[\boldsymbol{w}_x\right] \boldsymbol{S}\boldsymbol{S}^\top \right] \right. \\ & \left. + d\log\gamma + \text{tr}\left[\boldsymbol{S}\boldsymbol{S}^\top\right] - \log\left|\boldsymbol{S}\boldsymbol{S}^\top\right| + \left\|\boldsymbol{W}_z \boldsymbol{x}_{:i}\right\|_2^2 \right) \end{aligned} \tag{13}$$

with $\theta = \{\boldsymbol{w}_x, \gamma\}$ and $\phi = \{\boldsymbol{W}_z, \boldsymbol{S}\}$. Although a nonconvex loss, we can nonetheless take derivatives with respect to $\boldsymbol{S}\boldsymbol{S}^\top$ to demonstrate the existence of a single stationary point. In doing so, we find that $\boldsymbol{S}\boldsymbol{S}^\top = \left(\frac{1}{\gamma}\text{diag}\left[\boldsymbol{w}_x\right] \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \text{diag}\left[\boldsymbol{w}_x\right] + \boldsymbol{I}\right)^{-1}$ is the unique minimum which, when plugged into (13) leads to the revised cost

$$\begin{aligned}\mathcal{L}(\theta,\phi) \;\; \equiv \;\; & \sum_{i=1}^n \left(\left[\tfrac{1}{\gamma} \left\| \boldsymbol{x}_{:i} - \boldsymbol{\Phi}\text{diag}\left[\boldsymbol{w}_x\right] \boldsymbol{W}_z \boldsymbol{x}_{:i} \right\|_2^2 \right] + d\log\gamma \right. \\ & \left. + \log\left| \tfrac{1}{\gamma}\text{diag}\left[\boldsymbol{w}_x\right] \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \text{diag}\left[\boldsymbol{w}_x\right] + \boldsymbol{I} \right| + \left\|\boldsymbol{W}_z \boldsymbol{x}_{:i}\right\|_2^2 \right). \end{aligned} \tag{14}$$

As (14) is convex in $\boldsymbol{W}_z$, we may optimize away these parameters as well without encountering any bad local minima, noting that the optimal value satisfies

$$\boldsymbol{W}_z \boldsymbol{X} = \text{diag}\left[\boldsymbol{w}_x\right] \boldsymbol{\Phi}^\top \left(\boldsymbol{\Phi}\text{diag}\left[\boldsymbol{w}_x\right]^2 \boldsymbol{\Phi}^\top + \gamma \boldsymbol{I} \right)^{-1} \boldsymbol{X}. \tag{15}$$

Note that column-wise this expression is tantamount to the requirement that

$$\boldsymbol{\mu}_z(\boldsymbol{x}_{:i};\phi) = \text{diag}\left[\boldsymbol{w}_x\right] \boldsymbol{\Phi}^\top \left(\boldsymbol{\Phi}\text{diag}\left[\boldsymbol{w}_x\right]^2 \boldsymbol{\Phi}^\top + \gamma \boldsymbol{I} \right)^{-1} \boldsymbol{x}_{:i}, \;\; \forall i, \tag{16}$$

which satisfies

$$\lim_{\gamma \to 0} \boldsymbol{\mu}_z(\boldsymbol{x}_{:i};\phi) = \left(\boldsymbol{\Phi}\text{diag}\left[\boldsymbol{w}_x\right]\right)^\dagger \boldsymbol{x}_{:i}, \;\; \forall i, \tag{17}$$

as will be useful later on below.

To simplify notation moving forward, we define $\boldsymbol{w} \triangleq \boldsymbol{w}_x^2$, where the squaring operator is understood to apply element-wise. We also define $\boldsymbol{W} \triangleq \text{diag}[\boldsymbol{w}]$. Given these definitions, we can plug (15) into (14) to further reduce the remaining parameters. After applying standard determinant identities, the effective VAE cost can then be equivalently expressed as

$$\mathcal{L}(\boldsymbol{w},\gamma) = \text{tr}\left[\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\Sigma}_x^{-1}\right] + n\log\left|\boldsymbol{\Sigma}_x\right|, \text{ with } \boldsymbol{\Sigma}_x \triangleq \boldsymbol{\Phi}\boldsymbol{W}\boldsymbol{\Phi}^\top + \gamma \boldsymbol{I}. \tag{18}$$

C.2. Handling Decoder Parameters

The remainder of the proof primarily involves demonstrating that any local minimum of (18) w.r.t. the decoder parameters $\theta = \{\boldsymbol{w}_x, \gamma\}$ must satisfy the stated conditions of Theorem 4.1. This is in fact a considerable undertaking because while the first term in convex in $\boldsymbol{w}$, the second is concave and as such, could potentially introduce a huge constellation of bad local minima. Note that almost any arbitrary function can be expressed as the sum of a convex and concave component (Yuille

& Rangarajan, 2001), so we must explicitly rely on the specific form of (18) in framing our arguments. In this regard, it can be shown that in the special case where $n = 1$, all local minimizers are necessarily constrained to a fixed set of $\binom{\kappa}{d}$ possible candidates, each with exploitable decoupling properties that dramatically simplify the analysis per results from (Wipf et al., 2011). Unfortunately though, with the more general VAE model involving $n > 1$ it is not possible to follow an analogous route and exclude upfront local minimizers occurring beyond a fixed, quantifiable set. Consequently, we must adopt a completely different strategy as detailed herein.

For later convenience, and without loss of generality, we will assume that $\boldsymbol{\Phi} = [\boldsymbol{I}\ \boldsymbol{H}] \in \mathbb{R}^{d\times\kappa}$, where $\boldsymbol{I}$ is an identity matrix and $\boldsymbol{H} \in \mathbb{R}^{d\times(\kappa-d)}$ can be arbitrary. Such a dictionary can always be obtained from a general $\boldsymbol{\Phi}$ by left-multiplication, and as long as we left-multiply $\boldsymbol{X}$ by the same factor, the constraint set $\boldsymbol{X} = \boldsymbol{\Phi}\boldsymbol{U}$, as well as the overall VAE objective from (18) is unchanged (at least when $\gamma \to 0$ as stipulated). Likewise, the resulting $\boldsymbol{H}$ can still be implicitly decomposed as in (6) with probability one.

We will also further assume that the rows of $\widehat{\boldsymbol{U}}$ which putatively generated the data are arranged in decreasing order of row norm, again without loss of generality since a permutation matrix and its inverse can always be inserted between $\boldsymbol{\Phi}$ and $\boldsymbol{X}$. The rationale for this restructuring will become apparent as the proof progresses. Finally, we will require that the rows of $\widehat{\boldsymbol{U}}$ satisfy $\|\widehat{\boldsymbol{u}}_{(i+1)\cdot}\|_2 \leq \epsilon\|\widehat{\boldsymbol{u}}_{i\cdot}\|_2$. Note however that if we prove this result for some $\epsilon \in (0, 1]$, there will always exist at set of ν_i such that the proof also holds under the looser conditions of the theorem statement. We choose to work with a single ϵ simply because it leads to less cluttered notation. Finally, we also define $r \triangleq \rho(\widehat{\boldsymbol{U}})$ and $\widehat{\boldsymbol{w}} \triangleq [\widehat{w}_1, \ldots, \widehat{w}_\kappa]^\top$, where $\widehat{w}_i \triangleq \frac{1}{n}\|\widehat{\boldsymbol{u}}_{i\cdot}\|_2^2$.

The overall strategy of the proof relies on induction. First we will demonstrate that, under the appropriate conditions, at any candidate local minimum as $\gamma \to 0$, it must be that $w_1 = \widehat{w}_1 + O(\epsilon^2)$ while $w_i = O(\epsilon^2)$ for all $i > 1$. Later we will consider for any $K < r$ the set of solutions satisfying $w_i = \widehat{w}_i + O(\epsilon^{2i})$ for all $i = 1, \ldots, K$ and $w_i = O(\epsilon^{2K})$ for all $i > K$. If such a solution is to represent a local minimum, we show that additionally, it must be that $w_{(K+1)} = \widehat{w}_{(K+1)} + O(\epsilon^{2(K+1)})$ and $w_i = O(\epsilon^{2(K+1)}$ for all $i > K + 1$. Finally, we will establish that if $K = r$, there is a unique minimizer $\hat{\boldsymbol{w}}$ that satisfies these bounds as well as conditions of the theorem statement.

We begin by considering feasible solutions to $\boldsymbol{X} = \boldsymbol{\Phi}\boldsymbol{U}$. Given the conditions described above, the i-th row of $\boldsymbol{X}$ is given by $\boldsymbol{x}_{i\cdot} = \widehat{\boldsymbol{u}}_{i\cdot}$ for all $i = 1, \ldots, r$ and zero otherwise. Because the overall problem scaling is irrelevant, we may also assume that $\|\boldsymbol{x}_{1\cdot}\|_2 = \|\widehat{\boldsymbol{u}}_{1\cdot}\|_2 = 1$, and therefore, by the theorem statement we have that $\|\boldsymbol{x}_{i\cdot}\|_2 = \|\widehat{\boldsymbol{u}}_{i\cdot}\|_2 = O\left(\epsilon^{(i-1)}\right)$ for $i = 1, \ldots, r$ and zero otherwise.

Obviously there exist an infinite number of feasible solutions; however, the following lemma demonstrates that our assumptions constrain the space of possibilities.

Lemma C.1. *Let $\bar{\boldsymbol{U}}$ denote any feasible solution to $\boldsymbol{X} = \boldsymbol{\Phi}\boldsymbol{U}$. For ϵ sufficiently small, the row norms of $\bar{\boldsymbol{U}}$ must fall into one of the following three categories:*

1. *At least d rows satisfy $\|\bar{\boldsymbol{u}}_{i\cdot}\|_2 = \Omega(1)$.*
2. *Let f be some non-negative function such that $f(\epsilon) \to 0$ and $f(\epsilon)/\epsilon \to \infty$ as $\epsilon \to 0$. Then $\|\bar{\boldsymbol{u}}_{1\cdot}\|_2 = \Theta(1)$, and for at least $d - 1$ other rows $\|\bar{\boldsymbol{u}}_{i\cdot}\|_2 = \Theta\left(f(\epsilon)\right)$. The remaining $\kappa - d$ rows are $O\left(f(\epsilon)\right)$.*
3. *$\|\bar{\boldsymbol{u}}_{1\cdot}\|_2 = \|\widehat{\boldsymbol{u}}_{1\cdot}\|_2 + O(\epsilon)$ and $\|\bar{\boldsymbol{u}}_{i\cdot}\|_2 = O(\epsilon)$ for all $i > 1$.*

Proof: Note that $\boldsymbol{X} = \boldsymbol{\Theta}(1)$ per the above assumptions. Additionally, given that $\boldsymbol{\Phi}$ must satisfy (6), it follows that spark$[\boldsymbol{\Phi}] = d + 1$. Hence any set of d columns of $\boldsymbol{\Phi}$ excluding $\boldsymbol{\phi}_{\cdot 1}$, denoted $\tilde{\boldsymbol{\Phi}}$, will be invertible and the corresponding d nonzero rows of $\boldsymbol{U}$ given by $\tilde{\boldsymbol{U}} = \tilde{\boldsymbol{\Phi}}^{-1}\boldsymbol{X} = \boldsymbol{\Theta}(1)$ will also form a feasible solution. Note that *all* rows of any such $\tilde{\boldsymbol{U}}$ must be of order $\boldsymbol{\Theta}(1)$. Otherwise when ϵ becomes small, columns of $\boldsymbol{X}$ approach a scaled version of $\boldsymbol{\phi}_{\cdot 1}$, and we violate the spark assumption if any row norm of $\tilde{\boldsymbol{U}}$ tends towards zero (this would imply that d or fewer columns of $\boldsymbol{\Phi}$ are linearly dependent). Of course we can always have additional feasible solutions with $\rho(\boldsymbol{U}) > d$ that are of order $\Omega(1)$, i.e., some rows could become arbitrarily large. Collectively then, we have described the first category of solutions specified in the lemma.

Now we consider the second category. Let $\boldsymbol{\Phi}'$ denote $\boldsymbol{\Phi}$ with the first row and column removed, and let $\boldsymbol{X}'$ and $\boldsymbol{U}'$ be the corresponding $\boldsymbol{X}$ and $\boldsymbol{U}$ with the first row removed. Any feasible solution to $\boldsymbol{X} = \boldsymbol{\Phi}\boldsymbol{U}$ must also be feasible to

$\boldsymbol{X}' = \boldsymbol{\Phi}'\boldsymbol{U}'$ by construction given that $\boldsymbol{\Phi} = [\boldsymbol{I}\boldsymbol{H}]$. Now suppose that at most $P < d-1$ rows of $\boldsymbol{U}'$ are $\Theta\left(f(\epsilon)\right)$ and the rest are $O(\epsilon)$.[3] Let $\boldsymbol{U}'_p$ denote these rows and $\boldsymbol{\Phi}'_p$ the associated columns of $\boldsymbol{\Phi}'$. Then any feasible solution must be such that $\boldsymbol{X}' = \boldsymbol{\Phi}'_p\boldsymbol{U}'_p + O(\epsilon)$.

However, because $\boldsymbol{\Phi}'_p$ must be both full rank and overdetermined by assumption, $\boldsymbol{\Phi}'_p\boldsymbol{U}'_p = \Theta\left(f(\epsilon)\right)$. This is because $\text{spark}[\boldsymbol{\Phi}'] = \text{spark}[\boldsymbol{\Phi}] - 1 = d$ by design via (6). Therefore, since $\boldsymbol{X}'$ is $O(\epsilon)$, feasibility requires that $\Theta\left(f(\epsilon)\right) = O(\epsilon)$, which is a contradiction. Hence at least $d-1$ rows of $\boldsymbol{U}'$ must be $\Theta\left(f(\epsilon)\right)$. The remaining rows must be $O\left(f(\epsilon)\right)$, otherwise we encounter the same contradiction. Additionally, if all rows of $\boldsymbol{U}'$ are at most $\Theta\left(f(\epsilon)\right)$, then $\boldsymbol{u}_{1:}$ must be $\Theta(1)$ to maintain feasibility. This completes the second category.

Finally, the only remaining possibility is the third category, which includes the generative solution $\widehat{U}$. ■

Lemma C.1 leads to useful information regarding $\boldsymbol{\Sigma}_x$ at any candidate local minimum.

Lemma C.2. *Let $\bar{\boldsymbol{w}}$ denote an arbitrary candidate local minimum, with $\bar{\boldsymbol{W}} = diag[\bar{\boldsymbol{w}}]$. Then the singular values $\bar{\boldsymbol{\lambda}} = [\bar{\lambda}_1, \ldots, \bar{\lambda}_d]^\top$ of $\bar{\boldsymbol{\Sigma}}_x = \boldsymbol{\Phi}\bar{\boldsymbol{W}}\boldsymbol{\Phi}^\top + \gamma\boldsymbol{I}$ fall into one of the following two categories:*

1. *$\bar{\lambda}_1 = \Omega(1)$ and $\bar{\lambda}_i = \Omega\left(f(\epsilon)^2\right)$ for all $i > 1$.*
2. *$\bar{\lambda}_1 = \widehat{\lambda}_1 + O(\epsilon^2)$, $\lambda_i = O(\epsilon^2)$ for all $i > 1$, where $\widehat{\lambda}_1$ is the first singular value of $\boldsymbol{\Phi}\widehat{\boldsymbol{W}}\boldsymbol{\Phi}^\top$.*

Proof: If $\bar{\boldsymbol{w}}$ truly represents a local minimum, then it must satisfy the fixed point equation

$$\bar{w}_i = \tfrac{1}{n}\|\bar{\boldsymbol{u}}_{i:}\|_2^2 + \bar{\sigma}_{ii}^2, \quad \forall i, \tag{19}$$

where $\bar{\sigma}_{ii}^2$ denotes the i-th diagonal element of $\bar{\boldsymbol{\Sigma}} \triangleq \bar{\boldsymbol{W}} - \bar{\boldsymbol{W}}\boldsymbol{\Phi}^\top\bar{\boldsymbol{\Sigma}}_x^{-1}\boldsymbol{\Phi}\bar{\boldsymbol{W}}$. This follows from the upper bounds that were used to arrive at (18). Because $\bar{\boldsymbol{\Sigma}}$ is a symmetric, positive semi-definite matrix, it follows that $\bar{\sigma}_{ii}^2 \geq 0\ \forall i$. We may then conclude that $\bar{w}_i \geq \frac{1}{n}\|\bar{\boldsymbol{u}}_{i:}\|^2$. Consequently, if feasible solution $\bar{\boldsymbol{U}}$ falls into the first or second category in Lemma C.1, then it must be that $\bar{\lambda}_i$ is at least $\Omega\left(f(\epsilon)^2\right)$ for all i given that $\text{spark}[\boldsymbol{\Phi}] = d+1$ with probability one. Additionally, $\bar{\lambda}_1$ must be $\Omega(1)$. This is because for all categories from Lemma C.1, at least one row of $\bar{\boldsymbol{U}}$ is always at least $\Theta(1)$ for feasibility.

In contrast, if $\bar{\boldsymbol{U}}$ falls into the third category of Lemma C.1, we can consider the fact that any local minimizer $\bar{\boldsymbol{w}}$ must also be a local minimizer to the upper bound on (18), evaluated at $\bar{\boldsymbol{w}}$, given by

$$\bar{\mathcal{L}}(\boldsymbol{w}, \gamma) \;\triangleq\; \text{tr}\left[\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^\top\boldsymbol{W}^{-1}\right] + n\log|\boldsymbol{\Sigma}_x| + C \;\geq\; \mathcal{L}(\bar{\boldsymbol{w}}, \gamma), \tag{20}$$

where C is a constant independent of $\boldsymbol{w}$. Although not convex, functions of this form have been shown to have a single minimum, global or local. At this minimum, it must be that $\partial\bar{\mathcal{L}}(\boldsymbol{w}, \gamma)/\partial w_i = 0$, or after a few standard manipulations, that

$$\tfrac{1}{n}\|\bar{\boldsymbol{u}}_{i:}\|_2^2 \;=\; \boldsymbol{\phi}_{:i}^\top\left(\bar{\boldsymbol{\Sigma}}_x\right)^{-1}\boldsymbol{\phi}_{:i}. \tag{21}$$

When $\bar{\boldsymbol{U}}$ falls into the third category of Lemma C.1, this fixed-point equation can always be satisfied by $\bar{w}_1 = \frac{1}{n}\|\bar{\boldsymbol{u}}_{1:}\|_2^2 + O(\epsilon^2)$ and $\bar{w}_i = O(\epsilon^2)$ for all $i > 1$ when $\gamma \to 0$. Based on these observations, Lemma C.2 directly follows. ■

The first category in Lemma C.2 only provides a *lower* bound on the singular values because it does not take into account the effect of $\bar{\boldsymbol{\Sigma}}$ from (19) at each candidate local minimum, which as a non-negative additive term can only increase the value of each w_i and therefore each λ_i. However, we can also establish a simple but useful *upper* bound based on the following:

[3] Actually any $O\left(g(\epsilon)\right)$, where g is a nonnegative function such that $f(\epsilon)/g(\epsilon) \to \infty$ as ϵ becomes small, would do here. However, the simpler choice $O(\epsilon)$ can be assumed without loss of generality.

Corollary C.3. *At any local minimum, the singular values of $\bar{\boldsymbol{\Sigma}}_x$ also satisfy the upper bound $\bar{\lambda}_i = O(1)$ for all i.*

Proof: The second category of Lemma C.2 already achieves this upper bound. So here we focus on the first category, where all singular values are lower bounded by $\Omega\left(f(\epsilon)^2\right)$. Now suppose that some collection of P singular values is greater than $\Theta(1)$, meaning they are of order $\Omega\left(h(\epsilon)\right)$, where $h(\epsilon) \rightarrow \infty$ as $\epsilon \rightarrow 0$ for some non-negative function h. This implies that at least P elements of $\bar{\boldsymbol{w}}$ are also of order $\Omega\left(h(\epsilon)\right)$. Let $\mathcal{I}$ denote the indeces of these elements, and define $\boldsymbol{\Psi}$ such that $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top = \sum_{i \in \mathcal{I}} \bar{w}_i \boldsymbol{\phi}_{:i} \boldsymbol{\phi}_{:i}^\top$, which represents the contribution to $\bar{\boldsymbol{\Sigma}}_x$ of the associated basis vectors.

If $\bar{\boldsymbol{W}}$ is a local minimum, then it naturally follows that $\beta = 0$ must be a local minimum of

$$\mathcal{L}(\beta) = n \log \left| \bar{\boldsymbol{\Sigma}}_x + \beta \boldsymbol{\Psi}\boldsymbol{\Psi}^\top \right| + \text{trace}\left[\boldsymbol{X}\boldsymbol{X}^\top \left(\bar{\boldsymbol{\Sigma}}_x + \beta \boldsymbol{\Psi}\boldsymbol{\Psi}^\top \right)^{-1} \right], \tag{22}$$

otherwise we could just add or subtract a contribution from $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top$ to reduce the cost function. Given (22), a necessary condition for a local minimum is therefore

$$\left. \frac{\partial \mathcal{L}(\beta)}{\partial \beta} \right|_{\beta=0} = 0. \tag{23}$$

Taking derivatives, we have

$$\left. \frac{\partial \mathcal{L}(\beta)}{\partial \beta} \right|_{\beta=0} = n\text{tr}\left[\boldsymbol{\Psi}^\top \bar{\boldsymbol{\Sigma}}_x^{-1} \boldsymbol{\Psi} \right] - \text{tr}\left[\boldsymbol{\Psi}^\top \bar{\boldsymbol{\Sigma}}_x^{-1} \boldsymbol{X}\boldsymbol{X}^\top \bar{\boldsymbol{\Sigma}}_x^{-1} \boldsymbol{\Psi} \right]. \tag{24}$$

The first term in (24) is $\Theta(1)$, and with ϵ small converges to nP. For the second term, we note that $\left(\bar{\boldsymbol{\Sigma}}_x\right)^{-1} \boldsymbol{\Psi}$ is of order $O\left(h(\epsilon)^{-1/2}\right)$. Since $\boldsymbol{X}\boldsymbol{X}^\top = \Theta(1)$, the second term is therefore $O\left(h(\epsilon)^{-1}\right)$ which converges to zero. Hence we arrive at a contradiction, and $\bar{\lambda}_i$ cannot be more than $\Theta(1)$. However, since it can sometimes be smaller, we arrive at the upper bound $O(1)$. ■

The categorization and bounding of singular values provided by Lemma C.2 and Corollary C.3 ultimately allows us to establish that local minimum must be highly constrained as follows:

Lemma C.4. *If $\bar{\boldsymbol{w}}$ is a local minimum to the VAE objective under the previously stated conditions, then $\bar{w}_1 = \widehat{w}_1 + O(\epsilon^2)$ and $\bar{w}_i = O(\epsilon^2)$ for all $i > 1$.*

Proof: If $\bar{\boldsymbol{w}}$ is a local minimum, then it naturally follows that $\alpha = 1$, $\beta = 0$ must be a local minimum of

$$\mathcal{L}(\alpha, \beta) = n \log \left| \alpha \boldsymbol{\Phi} \bar{\boldsymbol{W}} \boldsymbol{\Phi}^\top + \beta \boldsymbol{e}_1 \boldsymbol{e}_1^\top \right| + \text{trace}\left[\boldsymbol{X}\boldsymbol{X}^\top \left(\alpha \boldsymbol{\Phi} \bar{\boldsymbol{W}} \boldsymbol{\Phi}^\top + \beta \boldsymbol{e}_1 \boldsymbol{e}_1^\top \right)^{-1} \right], \tag{25}$$

where $\boldsymbol{e}_1$ represents a unit vector with all zeroes and a one in the first position (similarly for $\boldsymbol{e}_i$ as will be applied in presenting later results). This occurs because at a true local minimum we can never rescale $\boldsymbol{w}$ by some constant α to reduce the cost, nor can we add a contribution from any column of $\boldsymbol{\Phi}$, for example, $\boldsymbol{\phi}_{:1} = \boldsymbol{e}_1$ without increasing the cost.

Given (34), necessary conditions for $\bar{\boldsymbol{w}}$ to be a local minimum are

$$\left. \frac{\partial \mathcal{L}(\alpha, \beta)}{\partial \alpha} \right|_{\alpha=1,\beta=0} = 0, \quad \left. \frac{\partial \mathcal{L}(\alpha, \beta)}{\partial \beta} \right|_{\alpha=1,\beta=0} \geq 0, \tag{26}$$

where we recognize that a positive gradient with respect to β can still be consistent with a local minimum if $\bar{w}_1 = 0$ (in contrast, if $\bar{w}_1 > 0$, then $\beta < 0$ is within the allowable constraint set and the gradient would have to exactly equal zero). Taking derivatives, this gives

$$\begin{aligned} \left. \frac{\partial \mathcal{L}(\alpha, \beta)}{\partial \alpha} \right|_{\alpha=1,\beta=0} &= n\text{tr}\left[\boldsymbol{I}\right] - \text{tr}\left[\boldsymbol{X}^\top (\bar{\boldsymbol{\Sigma}}_x)^{-1} \boldsymbol{X} \right], \\ \left. \frac{\partial \mathcal{L}(\alpha, \beta)}{\partial \beta} \right|_{\alpha=1,\beta=0} &= n\text{tr}\left[\boldsymbol{e}_1^\top \bar{\boldsymbol{\Sigma}}_x^{-1} \boldsymbol{e}_1 \right] - \text{tr}\left[\boldsymbol{e}_1^\top \bar{\boldsymbol{\Sigma}}_x^{-1} \boldsymbol{X}\boldsymbol{X}^\top \bar{\boldsymbol{\Sigma}}_x^{-1} \boldsymbol{e}_1 \right]. \end{aligned} \tag{27}$$

where $\bar{\boldsymbol{\Sigma}}_x = \boldsymbol{\Phi}\bar{\boldsymbol{W}}\boldsymbol{\Phi}^\top$ as before.

First consider the special case where all singular values achieve the upper bound from Corollary C.3, meaning $\bar{\lambda}_i = \Theta(1)$ for all i and so $\bar{\boldsymbol{\Sigma}}_x^{-1}$ is $\Theta(1)$ as well. Equating (27) to zero then requires that $\text{tr}\left[\boldsymbol{X}^\top \bar{\boldsymbol{\Sigma}}_x^{-1}\boldsymbol{X}\right] = \boldsymbol{e}_1^\top \bar{\boldsymbol{\Sigma}}_x^{-1}\boldsymbol{e}_1 + O(\epsilon) = nd$ since $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{e}_1\boldsymbol{e}_1^\top + O(\epsilon)$. Similarly, we note that

$$\boldsymbol{e}_1^\top \bar{\boldsymbol{\Sigma}}_x^{-1}\boldsymbol{X}\boldsymbol{X}^\top \bar{\boldsymbol{\Sigma}}_x^{-1}\boldsymbol{e}_1 = \boldsymbol{e}_1^\top \bar{\boldsymbol{\Sigma}}_x^{-1}\boldsymbol{e}_1\boldsymbol{e}_1^\top \bar{\boldsymbol{\Sigma}}_x^{-1}\boldsymbol{e}_1 + O(\epsilon). \tag{28}$$

Plugging these results into (27), a necessary condition for a local minimum is that

$$\boldsymbol{e}_1^\top \bar{\boldsymbol{\Sigma}}_x^{-1}\boldsymbol{e}_1 - \left(\boldsymbol{e}_1^\top \bar{\boldsymbol{\Sigma}}_x^{-1}\boldsymbol{e}_1\right)^2 + O(\epsilon) \geq 0. \tag{29}$$

However, with $\boldsymbol{e}_1^\top \bar{\boldsymbol{\Sigma}}_x^{-1}\boldsymbol{e}_1 = nd + O(\epsilon)$ this is not possible. Hence we cannot have a local minimum with $\bar{\lambda}_i = \Theta(1)$ for all i.

Now consider the more general case from Lemma C.2 where $\bar{\lambda}_1 = \Omega(1)$ and $\bar{\lambda}_i = \Omega\left(f(\epsilon)^2\right)$ for all $i > 1$. For the moment we will also assume that $\bar{\sigma}_{ii}^2 = 0$ for all i, which implies that $\bar{\lambda}_1 = \Theta(1)$ and $\bar{\lambda}_i = \Theta\left(f(\epsilon)^2\right)$ for all $i > 1$. This follows from Lemma C.1) and the fact that we have already ruled out the first category where all the row norms are $\Omega(1)$, in which case all the singular values would achieve the upper bound. Additionally, based on Lemmas C.1 and C.2, the only way such a singular value decomposition is possible is if $\bar{w}_1 = \Theta(1)$ and $\bar{w}_i = O\left(f(\epsilon)^2\right)$ for all $i > 1$.

We then partition $\bar{\boldsymbol{\Sigma}}_x^{-1}$ as $\boldsymbol{A} = \left[a_{11}\ \boldsymbol{a}_{21}^\top;\ \boldsymbol{a}_{21}\ \boldsymbol{A}_{22}\right]$, where $a_{11} = \left(\bar{\boldsymbol{\Sigma}}_x^{-1}\right)_{11}$, $\boldsymbol{A}_{22}$ represents $\bar{\boldsymbol{\Sigma}}_x^{-1}$ with the first row and column removed, and the vectors $\boldsymbol{a}_{21}^\top$ and $\boldsymbol{a}_{21}$ represent the remaining elements of the first row and column respectively. Proceeding further, based on the expression for the inverse of a partitioned matrix, it follows that a_{11} and $\boldsymbol{a}_{21}$ are $\Theta(1)$, given that $\bar{w}_1\boldsymbol{\phi}_{:1}\boldsymbol{\phi}_{:1}^\top = \bar{w}_1\boldsymbol{e}_1\boldsymbol{e}_1^\top$ with $\bar{w}_1 = \Theta(1)$ and all other $\bar{w}_i$ are of a smaller order. In contrast, $\boldsymbol{A}_{22} = \Theta\left(f(\epsilon)^{-2}\right)$. Finally, in a similar manner we write $\boldsymbol{X}\boldsymbol{X}^\top = \left[b_{11}\ \boldsymbol{b}_{21}^\top;\ \boldsymbol{b}_{21}\ \boldsymbol{B}_{22}\right]$ such that by construction $b_{11} = 1$, $\boldsymbol{b}_{21}$ is $O(\epsilon)$, and $\boldsymbol{B}_{22} = O(\epsilon^2)$.

Returning to (27), we find that

$$\text{tr}\left[\boldsymbol{X}\boldsymbol{X}^\top(\bar{\boldsymbol{\Sigma}}_x)^{-1}\right] = a_{11} + O\left([\epsilon f(\epsilon)^{-1}]^2\right) + O(\epsilon). \tag{30}$$

where $\epsilon f(\epsilon)^{-1} \to 0$ by definition. This implies that $a_{11} = nd + O\left([\epsilon f(\epsilon)^{-1}]^2\right) + O(\epsilon)$ at a local minimum. Again, similar to above we have that

$$\text{tr}\left[\boldsymbol{e}_1^\top \bar{\boldsymbol{\Sigma}}_x^{-1}\boldsymbol{X}\boldsymbol{X}^\top \bar{\boldsymbol{\Sigma}}_x^{-1}\boldsymbol{e}_1\right] = a_{11}^2 + O(\epsilon). \tag{31}$$

Consequently, based on (27) at any local minimum we require that

$$a_{11} - a_{11}^2 + O(\epsilon) = nd - (nd)^2 + O\left([\epsilon f(\epsilon)^{-1}]^2\right) + O(\epsilon) \geq 0. \tag{32}$$

However this can never be the case when ϵ is small enough.

Finally, we must consider the more general situation where $\bar{\sigma}_{ii}^2$ may be greater than zero for some or all i. Based on the upper bound on the singular values of $\bar{\boldsymbol{\Sigma}}_x$ from Corollary C.3, it follows that $\bar{\sigma}_{ii}^2 = O(1)$ for all i; if any were larger then one or more singular values would necessarily violate the bound. Given these conclusions we may proceed through the same analysis as above, only now with

$$\tilde{\boldsymbol{A}} \triangleq \bar{\boldsymbol{\Sigma}}_x^{-1} = \left[\boldsymbol{A}^{-1} + \sum_i \bar{\sigma}_{ii}^2\boldsymbol{\phi}_{:i}\boldsymbol{\phi}_{:i}^\top\right]^{-1} = \left[\boldsymbol{A}^{-1} + O(1)\right]^{-1}, \tag{33}$$

where $\boldsymbol{A}$ is exactly same as before.

In this revised situation, the results turn out to be essentially equivalent. $\tilde{a}_{11}$ and $\tilde{\boldsymbol{a}}_{21}$ are still of order $\Theta(1)$, while now $\tilde{\boldsymbol{A}}_{22} = O\left(f(\epsilon)^{-2}\right)$ instead of $\Theta\left(f(\epsilon)^{-2}\right)$. In other words $\tilde{\boldsymbol{A}}_{22}$ can potentially be smaller than before, which ultimately contributes an even greater violation to the local minimum condition. The rest of the analysis carries through with $\tilde{a}_{11}$ replacing a_{11}.

In conclusion, the only remaining candidate for a local minimum is when $\bar{\lambda}_1 = \widehat{\lambda}_1 + O(\epsilon^2)$, $\bar{\lambda}_i = O(\epsilon^2)$ for all $i > 0$. Additionally, based on Lemmas C.1 and C.2, the only way such a singular value decomposition is possible is if $\bar{w}_1 = \widehat{w}_1 + O(\epsilon^2)$ and $\bar{w}_i = O(\epsilon^2)$ for all $i > 1$. ■

The final remaining piece of the overall proof of Theorem 3 is the following inductive result.

Lemma C.5. *Assume we are given the candidate solution $\bar{w}_i = \widehat{w}_i + O(\epsilon^{2i})$ for all $i = 1, \ldots, K$ and $\bar{w}_i = O(\epsilon^{2K})$ for all $i > K$, with $K < r$. Now consider further optimization of the VAE cost function over only the $\bar{w}_i$ with $i > K$ (i.e., the first K elements remain fixed). If this $\bar{\boldsymbol{w}}$ truly represents a local minimum, then $\bar{w}_{K+1} = \widehat{w}_{K+1} + O(\epsilon^{2(K+1)})$ and $\bar{w}_i = O(\epsilon^{2(K+1)}$ for all $i > K+1$.*

Proof: Per the stipulation of this lemma, we assume that $\bar{w}_i = \widehat{w}_i + O(\epsilon^{2k})$ for all $i = 1, \ldots, K$ and $\bar{w}_i = O(\epsilon^{2K})$ for all $i > K$. Because the overall problem scaling is irrelevant, without loss of generality we may instead assume that $\|\widehat{\boldsymbol{u}}_{:K+1}\| = 1$. This then leads to the equivalent assumption that $\bar{w}_i = \widehat{w}_i + O(\epsilon^{2(k-K)})$ for all $k = 1, \ldots, K$ and $\bar{w}_i = O(1)$ for all $i > K$. Additionally, $\|\widehat{\boldsymbol{x}}_{(K+1):}\| = \|\widehat{\boldsymbol{u}}_{(K+1):}\| = 1$. We will now consider these first K parameters to be frozen and investigate optimization with respect to the remaining $\kappa - K$ $\bar{w}_i$ values.

Let $\boldsymbol{\Phi}_{\backslash K}$ denote $\boldsymbol{\Phi}$ with the first K columns and rows removed and let $\boldsymbol{X}_{\backslash K}$ and $\boldsymbol{U}_{\backslash K}$ denote the corresponding $\boldsymbol{X}$ and $\boldsymbol{U}$ with first K rows removed such that $\boldsymbol{X}_{\backslash K} = \boldsymbol{\Phi}_{\backslash K}\boldsymbol{U}_{\backslash K}$. We may apply Lemmas C.1 and C.2 to this reduced system to obtain a lower bound on the singular values of the bottom $d - K$ singular values of $\bar{\boldsymbol{\Sigma}}_x$ (the first K are of course fixed by $\bar{w}_1, \ldots, \bar{w}_K$).

We may also obtain an upper bound on the bottom $d - K$ singular values using a slight modification of Corollary C.3. Specifically, the derivative in (24) is nearly the same except that the first K columns and rows of $\bar{\boldsymbol{\Sigma}}_x^{-1}$ are now of order $O(\epsilon)$ because of the effect of $\sum_{i=1}^K \bar{w}_i \boldsymbol{\phi}_{:i}\boldsymbol{\phi}_{:i}^\top = \sum_{i=1}^K \bar{w}_i \boldsymbol{e}_i\boldsymbol{e}_i^\top$ (note that because of the rescaling, $\bar{w}_i = \Omega(\epsilon\frac{1}{n})$ for all $i \leq K$). This is easily shown using the expression for the inverse of a partitioned matrix. But the essential conclusion still holds, namely, the bottom $d - K$ singular values are bounded by $O(1)$.

From here we are positioned to show the desired result by adapting Lemma C.4. In particular, if $\bar{\boldsymbol{w}}$ is a local minimum, then $\alpha = 1$, $\beta = 0$ must be a local minimum of

$$\mathcal{L}(\alpha, \beta) = n \log \left|\bar{\boldsymbol{\Sigma}}_x\right| + \text{trace}\left[\boldsymbol{X}\boldsymbol{X}^\top \left(\bar{\boldsymbol{\Sigma}}_x\right)^{-1}\right]. \tag{34}$$

where now

$$\bar{\boldsymbol{\Sigma}}_x = \sum_{i=1}^K \bar{w}_i \boldsymbol{e}_i\boldsymbol{e}_i^\top + \alpha \sum_{i=K+1}^{\kappa} \bar{w}_i \boldsymbol{\phi}_{:i}\boldsymbol{\phi}_{:i}^\top + \beta \boldsymbol{e}_{(K+1)}\boldsymbol{e}_{(K+1)}^\top. \tag{35}$$

Again, using the formula for the inverse of a partitioned matrix and the lower bounds on singular values, the derivatives from (27) and (27) become

$$\begin{aligned}
\left.\frac{\partial \mathcal{L}(\alpha,\beta)}{\partial \alpha}\right|_{\alpha=1,\beta=0} &= n(d-K) - \text{tr}\left[\boldsymbol{X}_{\backslash K}^\top \left[(\bar{\boldsymbol{\Sigma}}_x)_{\backslash K}\right]^{-1} \boldsymbol{X}_{\backslash K}\right] + O(\epsilon), \\
\left.\frac{\partial \mathcal{L}(\alpha,\beta)}{\partial \beta}\right|_{\alpha=1,\beta=0} &= \\
& n\text{tr}\left[\boldsymbol{e}_1^\top \left[(\bar{\boldsymbol{\Sigma}}_x)_{\backslash K}\right]^{-1} \boldsymbol{e}_1\right] - \text{tr}\left[\boldsymbol{e}_1^\top \left[(\bar{\boldsymbol{\Sigma}}_x)_{\backslash K}\right]^{-1} \boldsymbol{X}_{\backslash K}\boldsymbol{X}_{\backslash K}^\top \left[(\bar{\boldsymbol{\Sigma}}_x)_{\backslash K}\right]^{-1} \boldsymbol{e}_1\right] + O(\epsilon),
\end{aligned} \tag{36}$$

where $(\bar{\boldsymbol{\Sigma}}_x)_{\backslash K}$ denotes $\bar{\boldsymbol{\Sigma}}_x$ with the first K columns and rows removed and $\boldsymbol{e}_1$ has replaced $\boldsymbol{e}_{(K+1)}$ to retain the proper alignment. These derivatives behave exactly like in Lemma C.4 only now in the reduced $(d-K)$-dimensional subspace. Hence we may borrow the previous arguments and conclude that $\bar{w}_{(K+1)} = \widehat{w}_{(K+1)} + O(\epsilon^2)$ and $\bar{w}_i = O(\epsilon^2)$ for all $i > K+1$. This is can be viewed as the generalized version of Lemma C.4. Finally, after rescaling the problem back to its original form, and combining with the first K fixed values of $\boldsymbol{w}$ obtained previously, we obtain the desired result. ■

To summarize, by assimilating all of these results, we may conclude that under the stated conditions, at any locally

minimizing solution $\boldsymbol{w} = \bar{\boldsymbol{w}}$ it must be that $\bar{w}_i = \widehat{w}_i + O(\epsilon^{2i})$ for all $i = 1, \ldots, r$ and $\bar{w}_i = O(\epsilon^{2r})$ for all $i > r$, where r is the number of nonzero rows in $\widehat{\boldsymbol{U}}$. It is then a simple matter to show that in fact $\bar{w}_i = 0$ for all $i > r$. One way to do this is to reconsider (22), only now with $\bar{\boldsymbol{\Sigma}}_x$ defined at the local solution converged to here and $\boldsymbol{\Psi}$ defined such that $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top = \sum_{i=r+1}^{\kappa} \bar{w}_i \boldsymbol{\phi}_{:i}\boldsymbol{\phi}_{:i}^\top$, i.e., the remaining small values of $\bar{\boldsymbol{w}}$ (or more precisely, we actually need only add up the remaining small values that are not already equal to zero).

With these changes we re-evaluate the derivative from (24). The first term will necessarily be $\Theta(1)$ under the stated conditions and assuming $\boldsymbol{\Psi} \neq 0$. In contrast, using similar analysis in deriving previous results, we also have that $\bar{\boldsymbol{\Sigma}}_x^{-1}\boldsymbol{X} = O(\epsilon^{(1-D)}$, and therefore the second term will be $O(\epsilon^2)$. Hence in combination, the derivative will always be positive unless $\boldsymbol{\Psi}$, and therefore $\bar{w}_i = 0$ for all $i > r$, equal zero.

Finally, the following result can be used to demonstrate that the remaining nonzero elements of $\bar{\boldsymbol{w}}$ at any minimum must be such that $\bar{w}_i = \widehat{w}_i \triangleq \frac{1}{n}\|\widehat{\boldsymbol{u}}_{i:}\|_2^2$, i.e., any deviations from $\widehat{\boldsymbol{w}}$ as defined previously evaporate.

Lemma C.6. *Consider the objective (18) with $\gamma \to 0$ and any $\kappa - d$ elements of $\boldsymbol{w}$ fixed at zero. Let $\boldsymbol{w}_d \in \mathbb{R}^d$ denote the unconstrained elements of $\boldsymbol{w}$, $\boldsymbol{\Phi}_d \in \mathbb{R}^{d\times d}$ the corresponding columns of $\boldsymbol{\Phi}$, and $\boldsymbol{U}_d \triangleq \boldsymbol{\Phi}_d^{-1}\boldsymbol{X}$. Then $\bar{\boldsymbol{w}}_d$ represents a minimum of the resulting constrained objective (global or local) iff $\bar{w}_d, i = \frac{1}{n}\|\boldsymbol{u}_{d,i:}\|_2^2$ (i.e., we are optimizing only those d elements of $\boldsymbol{w}$ that are not fixed at zero).*

Proof: Under the stated conditions, the constrained objective becomes

$$\mathcal{L}(\boldsymbol{w}_d, 0) = \text{tr}\left[\boldsymbol{X}\boldsymbol{X}^\top\left(\boldsymbol{\Phi}_d\boldsymbol{W}_d\boldsymbol{\Phi}_d^\top\right)^{-1}\right] + n\log\left|\boldsymbol{\Phi}_d\boldsymbol{W}_d\boldsymbol{\Phi}_d^\top\right|, \tag{37}$$

where $\boldsymbol{W}_d \triangleq \text{diag}[\boldsymbol{w}_d]$. After a few standard manipulations, (37) can be equivalently expressed as

$$\mathcal{L}(\boldsymbol{w}_d, 0) \equiv \sum_{i=1}^{d}\frac{\frac{1}{n}\|\boldsymbol{u}_{d,i:}\|_2^2}{w_{d,i}} + \sum_{i=1}^{d}\log w_{d,i} + C, \tag{38}$$

where C is a constant independent of $\boldsymbol{w}_d$. This function is separable, with a unique minimum each $w_{d,i}$ given by $\bar{w}_d, i = \frac{1}{n}\|\boldsymbol{u}_{d,i:}\|_2^2$. The latter follows by simply taking gradients and equating to zero. ■

Applying Lemma C.6 to the solution defined previously with nonzeros aligned with $\widehat{\boldsymbol{w}}$ (note that we can always fill out $d - \kappa$ additional unconstrained elements to enter the regime where the lemma applies), ensures that, if $\bar{\boldsymbol{w}}$ is any minimum to (18) (whether global or local), it must be that

$$\bar{w}_i = \widehat{w}_i = \tfrac{1}{n}\|\widehat{\boldsymbol{u}}_{i:}\|_2^2, \;\; \forall i = 1, \ldots, \kappa. \tag{39}$$

C.3. Combining Pieces

We are now positioned to revisit the original four conclusions of Theorem 4.1. In brief, we first demonstrated that at any minimum w.r.t the encoder parameters $\phi = \{\boldsymbol{W}_z, \boldsymbol{S}\}$ (global or local), the VAE energy under the stated conditions reduces to (18) for any arbitrary value of the decoder $\theta = \{\boldsymbol{w}_x, \gamma\}$. We then showed that (18) has a unique minimum w.r.t $\boldsymbol{w}_x$ in the limit $\gamma \to 0$ (with the limit taken *outside* of the minimization), namely $\boldsymbol{w}^* = \bar{\boldsymbol{w}}$ such that Conclusion (i) of Theorem 4.1 is satisfied.

Conclusions (ii) and (iii) directly follow from Conclusion (i) and (17). Specifically, for (ii) we have that

$$\boldsymbol{\Phi}\text{diag}\left[\boldsymbol{w}_x^*\right]\boldsymbol{\mu}_z(\boldsymbol{x}_{:i};\phi^*) = \boldsymbol{\Phi}\text{diag}\left[\bar{\boldsymbol{w}}_x\right]\left(\boldsymbol{\Phi}\text{diag}\left[\bar{\boldsymbol{w}}_x\right]\right)^\dagger\boldsymbol{x}_{:i} = \boldsymbol{x}_{:i}, \;\; \forall i. \tag{40}$$

such that zero reconstruction error is achieved. Similarly, given that the support of $\bar{\boldsymbol{w}}$ and $\widehat{\boldsymbol{u}}_{:i}$ are equivalent, and the uniqueness of any feasible solution defined w.r.t. the corresponding columns of $\boldsymbol{\Phi}$, then for (iii) we have that

$$\text{diag}\left[\boldsymbol{w}_x^*\right]\boldsymbol{\mu}_z(\boldsymbol{x}_{:i};\phi^*) = \text{diag}\left[\bar{\boldsymbol{w}}_x\right]\left(\boldsymbol{\Phi}\text{diag}\left[\bar{\boldsymbol{w}}_x\right]\right)^\dagger\boldsymbol{x}_{:i} = \widehat{\boldsymbol{u}}_{:i}, \;\; \forall i. \tag{41}$$

And lastly, Conclusion (iv) follows by repurposing the analysis that was used to arrive at Lemma C.1. Specifically, suppose that $\boldsymbol{U}_0 \neq \widehat{\boldsymbol{U}}$. This means that at least one of the r nonzero rows of $\widehat{\boldsymbol{U}}$ must be zero in $\boldsymbol{U}_0$; if this were not the case, then either $\boldsymbol{U}_0 = \widehat{\boldsymbol{U}}$ or $\rho(\boldsymbol{U}_0) > \rho(\widehat{\boldsymbol{U}})$ which is a contradiction. Let K denote the index of the largest nonzero row of $\widehat{\boldsymbol{U}}$ that is zero in $\boldsymbol{U}_0$. If $K = 1$, then $\boldsymbol{U}_0$ will necessarily fall into the first category of candidate solutions specified by Lemma C.1, and hence $\rho(\boldsymbol{U}_0) > \rho(\widehat{\boldsymbol{U}})$ which is a contradiction. For $K > 1$, let $\boldsymbol{\Phi}'$ denote the submatrix of $\boldsymbol{\Phi}$ with the first $K-1$ rows and columns removed; similarly $\boldsymbol{X}'$ and $\boldsymbol{U}_0'$ the submatrices of $\boldsymbol{X}$ and $\boldsymbol{U}_0$ with the first $K-1$ rows removed. In this revised scenario, we may leverage the same analysis as was used to prove Lemma $C.1$ to show that $\rho(\boldsymbol{U}_0') \geq d - K + 1$, and therefore it must be that $\rho(\boldsymbol{U}_0) = K - 1 + \rho(\boldsymbol{U}_0') \geq d > \rho(\widehat{\boldsymbol{U}})$, which again is a contradiction. Hence $\boldsymbol{U}_0 = \widehat{\boldsymbol{U}}$ completing the proof. ■

D. Proof of Theorem 4.2

We first present the following lemma, which allows us to relax the constraint from (8).

Lemma D.1. *If $\bar{\boldsymbol{Z}}$ is any minimum (global or local) of the function*

$$\mathcal{L}(\boldsymbol{Z}) \triangleq \sum_{i=1}^{n} \tfrac{1}{\gamma} \left\| \boldsymbol{x}_{:i} - \boldsymbol{A}\boldsymbol{z}_{:i} \right\|_2^2 + \sum_{j=1}^{\kappa} h\left(\|\boldsymbol{z}_{j:}\|_2^2 \right) \tag{42}$$

defined via arbitrary matrix $\boldsymbol{A} \in \mathbb{R}^{d \times \kappa}$ and concave, non-decreasing function h, then it must be that $\bar{\boldsymbol{Z}} = \boldsymbol{B}\boldsymbol{X}$ for some matrix $\boldsymbol{B} \in \mathbb{R}^{\kappa \times d}$.

Proof: If $\bar{\boldsymbol{Z}}$ is any global or local minimum of (42), then it must also be a minimum of any convex upper bound $\tilde{\mathcal{L}}(\boldsymbol{Z}) \geq \mathcal{L}(\boldsymbol{Z})$ defined such that $\tilde{\mathcal{L}}(\bar{\boldsymbol{Z}}) = \mathcal{L}(\bar{\boldsymbol{Z}})$. If this were not the case, we could further minimize $\tilde{\mathcal{L}}(\boldsymbol{Z})$ along some descent path, which would violate our assumption that the bound is tight at $\bar{\boldsymbol{Z}}$. In the present case, we can always construct such abound using a quadratic approximation to h.

For example, because h is a concave, non-decreasing function, it can always be expressed as

$$h(x) = \min_{\lambda \geq 0} \; [\lambda x - h^*(\lambda)] \leq \lambda x - h^*(\lambda), \; \forall \lambda \geq 0, \tag{43}$$

where h^* is the concave conjugate of h (Rockafellar, 1970). Furthermore, per the basic rules of Fenchel duality, at any point x' we obtain equality with righthand-side upper bound when λ is set equal to $\lambda' \triangleq \partial h(x)/\partial x|_{x=x'}$, i.e., $h(x') = \lambda' x' - h^*(\lambda')$. Consequently, if $\bar{\boldsymbol{Z}}$ is a local minimum to (42), it must also be a minimum to the convex objective

$$\tilde{\mathcal{L}}(\boldsymbol{Z}) \triangleq \sum_{i=1}^{n} \tfrac{1}{\gamma} \left\| \boldsymbol{x}_{:i} - \boldsymbol{A}\boldsymbol{z}_{:i} \right\|_2^2 + \sum_{j=1}^{\kappa} \bar{\lambda}_j \|\boldsymbol{z}_{j:}\|_2^2 + C, \tag{44}$$

where $\bar{\lambda}_j \triangleq \partial h(x)/\partial x|_{x=\|\bar{\boldsymbol{z}}_{j:}\|_2^2}$ and C is a constant independent of $\boldsymbol{Z}$. But this is just a generalized form of penalized ridge regression with optimal solution

$$\boldsymbol{Z} = \bar{\boldsymbol{\Lambda}}\boldsymbol{A}^\top \left(\gamma \boldsymbol{I} + \boldsymbol{A}\bar{\boldsymbol{\Lambda}}\boldsymbol{A}^\top \right)^{-1} \boldsymbol{X}, \tag{45}$$

where $\bar{\boldsymbol{\Lambda}} \triangleq \text{diag}[\bar{\boldsymbol{\lambda}}]$. Therefore, it follows that any locally minimizing $\boldsymbol{Z}$ must be in the form $\bar{\boldsymbol{Z}} = \boldsymbol{B}\boldsymbol{X}$. ■

Per Lemma D.1, we can relax the constraint $\boldsymbol{Z} = \boldsymbol{W}_z\boldsymbol{X}$ and simply optimize over $\boldsymbol{Z}$ directly. In doing so, any local minimum we enter will necessarily also be a local minimum to the original constrained objective. Hence we instead consider the more convenient, relaxed objective

$$\mathcal{L}_{AE}(\boldsymbol{w}_x, \boldsymbol{Z}) \triangleq \sum_{i=1}^{n} \tfrac{1}{\gamma} \left\| \boldsymbol{x}_{:i} - \boldsymbol{\Phi}\text{diag}\left[\boldsymbol{w}_x\right] \boldsymbol{z}_{:i} \right\|_2^2 + \|\boldsymbol{w}_x\|_2^2 + \sum_{j=1}^{\kappa} h\left(\|\boldsymbol{z}_{j:}\|_2^2 \right). \tag{46}$$

If we define $\boldsymbol{U} \triangleq \text{diag}\left[\boldsymbol{w}_x\right] \boldsymbol{Z} \in \mathbb{R}^{\kappa \times n}$ then we may also obtain an equivalent reparameterization of (46) given by

$$\mathcal{L}_{AE}(\boldsymbol{U}, \boldsymbol{Z}) \triangleq \sum_{i=1}^{n} \tfrac{1}{\gamma} \left\|\boldsymbol{x}_{:i} - \boldsymbol{\Phi}\boldsymbol{u}_{:i}\right\|_2^2 + \sum_{j=1}^{\kappa} \left[\frac{\|\boldsymbol{u}_{j:}\|_2^2}{\|\boldsymbol{z}_{j:}\|_2^2} + h\left(\|\boldsymbol{z}_{j:}\|_2^2\right)\right]. \tag{47}$$

From (47) we observe that dependency on $\boldsymbol{Z}$ is now restricted to the separable row norms which can be conveniently optimized away. In particular, we apply the following simple lemma:

Lemma D.2. *Let $\widetilde{h}(x) \triangleq \inf_{\alpha>0} \left[\frac{x^2}{\alpha} + h(\alpha)\right]$, where h is a concave non-deceasing function. Then $\widetilde{h}$ will be a concave non-decreasing function of $|x|$ defined on the domain $[0, \infty)$.*

Proof: This result can be inferred from (Wipf & Zhang, 2014)[Theorem 2]. ■

Note that we can always minimize (47) with respect to $\boldsymbol{Z}$ without encountering spurious minima, and based on Lemma D.2, the effective objective that emerges will be

$$\mathcal{L}_{AE}(\boldsymbol{U}) \triangleq \sum_{i=1}^{n} \tfrac{1}{\gamma} \left\|\boldsymbol{x}_{:i} - \boldsymbol{\Phi}\boldsymbol{u}_{:i}\right\|_2^2 + \sum_{j=1}^{\kappa} \widetilde{h}\left(\|\boldsymbol{u}_{j:}\|_2\right), \tag{48}$$

where $\widetilde{h}$ is a concave, non-decreasing function of the row norms of $\boldsymbol{U}$ (no longer the row-norms squared as with the function h applied to $\boldsymbol{Z}$). Such functions are well-known to favor row-sparse solutions, meaning matrices with many rows pushed to exactly zero (Cotter et al., 2005). And finally, if we allow $\gamma \to 0$, then (48) can be recast in the more interpretable equivalent form

$$\mathcal{L}_{AE}(\boldsymbol{U}) \equiv \sum_{j=1}^{\kappa} \widetilde{h}\left(\|\boldsymbol{u}_{j:}\|_2\right), \;\; \text{s.t. } \boldsymbol{X} = \boldsymbol{\Phi}\boldsymbol{U}. \tag{49}$$

It has been shown (Wipf et al., 2011)[Theorem 9] that in the special case of $n = 1$, any possible objective in the form of (49) can have minima (either local or global) that do not produce an optimal sparse representation under an analogous scaling constraint. This is sufficient to prove Theorem 4.2. ■

E. Proof of Lemma 4.3

Under the stated conditions, if we follow the encoder optimizations analogous to those used in the proof of Theorem 4.1, we arrive at the VAE loss

$$\mathcal{L}(\theta, \phi) = \sum_{i=1}^{n} \tfrac{1}{\gamma} \left\|\boldsymbol{x}_{:i} - \boldsymbol{\Phi}\text{diag}\left[\boldsymbol{w}_x\right]\boldsymbol{m}_{:i}\right\|_2^2 + n \sum_{j=1}^{\kappa} \log\left(\gamma + w_{x,j}^2 \|\boldsymbol{\phi}_{:j}\|_2^2\right) + \|\boldsymbol{M}\|_{\mathcal{F}}^2, \tag{50}$$

where $w_{x,j}$ denotes the j-th element of $\boldsymbol{w}_x$ and we have aggregated encoder means into the matrix $\boldsymbol{M} \triangleq [\boldsymbol{\mu}_z(\boldsymbol{x}_{:1}; \phi), \ldots, \boldsymbol{\mu}_z(\boldsymbol{x}_{:n}; \phi)] \in \mathbb{R}^{d \times n}$. Then after taking the limit $\gamma \to 0$ and regrouping terms, this becomes equivalent to

$$\mathcal{L}(\theta, \phi) \equiv \sum_{j=1}^{\kappa} \left[\log\left(w_{x,j}^2\right) + \tfrac{1}{n}\|\boldsymbol{m}_{j:}\|_2^2\right], \;\; \text{s.t.} \;\; \begin{array}{c} \boldsymbol{m}_{:i} = \boldsymbol{W}_z \boldsymbol{x}_{:i} \\ \boldsymbol{x}_{:i} = \boldsymbol{\Phi}\text{diag}\left[\boldsymbol{w}_x\right]\boldsymbol{m}_{:i} \end{array}, \;\; \forall i \tag{51}$$

excluding irrelevant constants. From this expression we observe that the penalization applied to $\boldsymbol{w}_x$ and each $\boldsymbol{m}_{j:}$ is nearly equivalent to that applied to $\boldsymbol{w}_x$ and $\boldsymbol{z}_{j:}$ in (9) when $h(\cdot) = \log(\cdot)$; the only difference is that the two penalty function nonlinearities $(\cdot)^2$ and $\log(\cdot)^2$ are flipped. However, since the reconstruction only depends on the product $w_{x,j}\frac{1}{\sqrt{n}}\|\boldsymbol{m}_{j:}\|_2$ for all j (not each factor in isolation), the aggregate penalization on this product is actually equivalent, i.e., switching the two functions makes no difference beyond an inconsequential reparameterization. ■

F. Proof of Theorem 4.4

Following the optimization over encoder parameters discussed in the proof of Theorem 4.1, the VAE energy from (5) reduces to (18). From this expression, w.l.o.g. we decompose $\boldsymbol{\Sigma}_x$ as

$$\boldsymbol{\Sigma}_x = w_{x,i}^2 \boldsymbol{\phi}_{:i}\boldsymbol{\phi}_{:i}^\top + \boldsymbol{\Psi}_i, \;\; \text{with } \boldsymbol{\Psi}_i \triangleq \sum_{j \neq i} w_{x,j}^2 \boldsymbol{\phi}_{:j}\boldsymbol{\phi}_{:j}^\top + \gamma \boldsymbol{I}. \tag{52}$$

We then reexpress $\boldsymbol{\Psi}_i$ as

$$\boldsymbol{\Psi}_i = \gamma_i' \boldsymbol{\phi}_{:i}\boldsymbol{\phi}_{:i}^\top + \boldsymbol{R}_i \boldsymbol{R}_i^\top, \tag{53}$$

where $\boldsymbol{R}_i \in \mathbb{R}^{d \times d-1}$ is a matrix with columns spanning the orthogonal complement of $\boldsymbol{\phi}_{:i}$ and $\gamma_i' \geq \gamma / \|\boldsymbol{\phi}_{:i}\|_2^2$ is a projection weight of $\boldsymbol{\Psi}_i$ in the direction of $\boldsymbol{\phi}_{:i}$. In this way, by applying standard determinant and matrix inverse identities we can reformulate (18) as

$$\begin{aligned}
\mathcal{L}(\boldsymbol{w}, \gamma) &= \text{tr}\left[\boldsymbol{X}\boldsymbol{X}^\top \left([w_{x,i}^2 + \gamma']\boldsymbol{\phi}_{:i}\boldsymbol{\phi}_{:i}^\top + \boldsymbol{R}_i\boldsymbol{R}_i^\top\right)^{-1}\right] + n \log \left|[w_{x,i}^2 + \gamma']\boldsymbol{\phi}_{:i}\boldsymbol{\phi}_{:i}^\top + \boldsymbol{R}_i\boldsymbol{R}_i^\top\right| \\
&= \frac{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{x}}_i}{\left(w_{x,i}^2 + \gamma'\right)\|\boldsymbol{\phi}_{:i}\|_2^4} + n \log\left(w_{x,i}^2 + \gamma'\right) + C \\
&\equiv \frac{a}{\left(w_{x,i}^2 + \gamma'\right)} + \log\left(w_{x,i}^2 + \gamma'\right),
\end{aligned} \tag{54}$$

where C is a constant independent of $w_{x,i}^2$ and $\widetilde{\boldsymbol{x}}_i \in \mathbb{R}^n$ denotes a vector whose j-th element is $\boldsymbol{x}_{:j}^\top \boldsymbol{\phi}_{:i}$ and $a \triangleq \widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{x}}_i / \left(n\|\boldsymbol{\phi}_{:i}\|_2^4\right)$. It is then a simple matter to show via differentiation that (54) either has a single stationary point at $w_{x,i}^2 = a - \gamma'$ when $a - \gamma' \geq 0$, which serves as the unique minimum. Otherwise if $a - \gamma' < 0$, then the minimum occurs as $w_{x,i}^2 = 0$ and monotonically increases from there. Hence there is a single unique minimum, which proves Theorem 4.4. ■

G. Corollary 4.5

For the first part of the corollary, we optimize (50) from the proof of Lemma 4.3 over $\boldsymbol{M}$, and then adopt the decompositions (52) and (53) defined in the proof of Theorem 4.4. This leads to a modified version of (54) given by

$$\begin{aligned}
\mathcal{L}(\boldsymbol{w}, \gamma) &= \text{tr}\left[\boldsymbol{X}\boldsymbol{X}^\top \left([w_{x,i}^2 + \gamma']\boldsymbol{\phi}_{:i}\boldsymbol{\phi}_{:i}^\top + \boldsymbol{R}_i\boldsymbol{R}_i^\top\right)^{-1}\right] + n \sum_{j=1}^{\kappa} \log\left(\gamma + w_{x,j}^2 \|\boldsymbol{\phi}_{:j}\|_2^2\right) \\
&\equiv \frac{a}{\left(w_{x,i}^2 + \gamma'\right)} + \log\left(\gamma'' + w_{x,i}^2\right),
\end{aligned} \tag{55}$$

where $\gamma'' \triangleq \gamma / \|\boldsymbol{\phi}_{:i}\|_2^2$ and terms independent of $w_{x,i}^2$ have been omitted.

However, because $\gamma'' \leq \gamma'$, it is now actually possible for (55) to have multiple disconnected local minima. To see this, we can take the gradient w.r.t. $w_{x,i}^2$ and equate to zero, revealing that two feasible stationary points with $w_{x,i}^2 > 0$ are possible. Specifically, we have

$$w_{x,i}^2 = \frac{a - 2\gamma' \pm \sqrt{a^2 + 4a(\gamma'' - \gamma')}}{2} \tag{56}$$

as candidate solutions. Note that to guarantee no suboptimal local minima requires that the smaller candidate solution from (56) is a non-negative real for all feasible values of $\{a, \gamma', \gamma''\}$; the latter consists of all non-negative selections satisfying the constraint $\gamma' \geq \gamma''$. However, it is straightforward to find feasible sets $\{a, \gamma', \gamma''\}$ such that both $w_{x,i}^2$ solutions from (56) are non-negative (e.g., $a = 10,\ \gamma' = 1,\ \gamma'' = 0.01$) and therefore, we cannot rule out suboptimal local minima when $\boldsymbol{\Sigma}_z$ is diagonalized under the conditions of Theorem 4.4.

We now turn to the second part of the corollary. For the $\boldsymbol{s}$ case, we begin with (50) (where $\boldsymbol{s}$ has already been optimized away with no bad local minima). We assume $\boldsymbol{W}_{z,\setminus i}$ is fixed, which for present purposes is tantamount to treating $\boldsymbol{M}_{\setminus i}$ as

fixed (as opposed to optimizing it away). After defining

$$\bar{\boldsymbol{X}}_i \;\triangleq\; \boldsymbol{X} - \sum_{j \neq i} \boldsymbol{\phi}_{:j} w_{x,j} \boldsymbol{m}_{j:} \tag{57}$$

and excluding all terms that are independent of $w_{x,i}$ and $\boldsymbol{m}_{i:} = \boldsymbol{w}_{z,i:}\boldsymbol{X}$, (50) can be reexpressed as

$$\mathcal{L}(w_{x,i}, \boldsymbol{w}_{z,i:}, \gamma) = \tfrac{1}{\gamma} \left\| \bar{\boldsymbol{X}}_i - \boldsymbol{\phi}_{:i} w_{x,i} \boldsymbol{m}_{i:} \right\|_{\mathcal{F}}^2 + n \log\left(\gamma + w_{x,i}^2 \|\boldsymbol{\phi}_{:i}\|_2^2\right) + \|\boldsymbol{m}_{i:}\|_2^2. \tag{58}$$

This function is convex in $\boldsymbol{m}_{i:}$, and once optimized away, we obtain

$$\mathcal{L}(w_{x,i}, \gamma) = \frac{\bar{a}}{\gamma + w_{x,i}^2 \|\boldsymbol{\phi}_{:i}\|_2^2} + \log\left(\gamma + w_{x,i}^2 \|\boldsymbol{\phi}_{:i}\|_2^2\right), \tag{59}$$

where $\bar{a} \triangleq \boldsymbol{\phi}_{:i}^\top \bar{\boldsymbol{X}}_i \bar{\boldsymbol{X}}_i^\top \boldsymbol{\phi}_{:i} / (n \|\boldsymbol{\phi}_{:i}\|_2^2)$. With a simple rescaling, this expression is in the same basic form as (54), and hence has a single unique minima (global or local). And finally, for the $\boldsymbol{S}$ case, we follow similar steps as above, only now the log-det term has an additional factor stemming from the decomposition from (53). The resulting simplified loss is then structurally analogous to (55), with the critical difference that now $\gamma'' \geq \gamma'$. Upon differentiation and inspection, we then observe that there is a unique minimum like the $\boldsymbol{s}$ case. ■

H. Proof of Corollary 4.6

Given

$$\bar{\boldsymbol{X}}_i \;\triangleq\; \boldsymbol{X} - \sum_{j \neq i} \boldsymbol{\phi}_{:j} w_{x,j} \boldsymbol{z}_{j:} \tag{60}$$

and excluding all terms that are independent of $w_{x,i}$ and $\boldsymbol{z}_{i:} = \boldsymbol{w}_{z,i:}\boldsymbol{X}$, (8) can be reexpressed as

$$\mathcal{L}_{AE}(w_{x,i}, \boldsymbol{w}_{z,i:}) = \tfrac{1}{\gamma} \left\| \bar{\boldsymbol{X}}_i - \boldsymbol{\phi}_{:i} w_{x,i} \boldsymbol{z}_{i:} \right\|_{\mathcal{F}}^2 + w_{x,i}^2 + h\left(\tfrac{1}{n} \|\boldsymbol{z}_{i:}\|_2^2\right). \tag{61}$$

Using an analogous reparameterization as adopted in the proof of Theorem 4.2, (61) can be converted to the equivalent objective

$$\mathcal{L}_{AE}(\boldsymbol{u}_{i:}) = \tfrac{1}{\gamma} \left\| \bar{\boldsymbol{X}}_i - \boldsymbol{\phi}_{:i} \boldsymbol{u}_{i:} \right\|_{\mathcal{F}}^2 + \widetilde{h}\left(\tfrac{1}{n} \|\boldsymbol{u}_{i:}\|_2^2\right), \tag{62}$$

where it follows by construction that the composite $\widetilde{h}\left([\cdot]^2\right)$ must be concave, non-decreasing, and nonlinear if $h([\cdot]^2)$ is. From here we can trivially choose any number of simple counter-examples. For example, if we choose $n = 1$, then (62) reduces to a function $f : \mathbb{R} \rightarrow \mathbb{R}$ in the general form

$$f(u) = u^2 - 2au + \gamma \widetilde{h}(u^2), \tag{63}$$

where $a \in \mathbb{R}$ can be arbitrary. We can then apply (Wipf et al., 2011)[Theorem 6] which demonstrates that if $h([\cdot]^2)$ is strictly concave, $f(x)$ can have multiple local minima. However, from examination of the proof, it is trivial to extend to any concave, non-decreasing, and nonlinear $h([\cdot]^2)$. ■