Deep Denoising: Rate-Optimal Recovery of Structured Signals with a Deep Prior

Reinhard Heckel* Wen Huang* Paul Hand* Vladislav Voroninski†

Department of ECE*, Rice University
Department of CAAM*, Rice University
Helm.ai†

May 22, 2018

Abstract

Deep neural networks provide state-of-the-art performance for image denoising, where the goal is to map a noisy image to a near noise-free image. The underlying principle is simple: images are well described by priors that map a low-dimensional latent representations to image. Based on a prior, a noisy image can be denoised by finding a close image in the range of the prior. Since deep networks trained on large set of images have empirically been shown to be good priors, they enable effective denoisers. However, there is little theory to justify this success, let alone to predict the denoising performance. In this paper we consider the problem of denoising an image from additive Gaussian noise with variance σ^2, assuming the image is well described by a deep neural network with ReLu activations functions, mapping a k-dimensional latent space to an n-dimensional image. We provide an iterative algorithm minimizing a non-convex loss that provably removes noise energy by a fraction $\sigma^2 k/n$. We also demonstrate in numerical experiments that this denoising performance is, indeed, achieved by generative priors learned from data.

1 Introduction

We consider the image or data denoising problem, where the goal is to remove noise from an unknown image or data point. In more detail, our goal is to obtain an estimate of an image or vector $y_* \in \mathbb{R}^n$ from

$$y = y_* + \eta,$$

where η is unknown noise, often modeled as a zero-mean white Gaussian random variable with covariance matrix σ^2/nI. Image denoising relies on prior assumptions on the image y_*. For example, if the signal y_* lies in a k-dimensional subspace $\mathcal{Y}$, then the estimate $\hat{y}$ that selects the closest point in ℓ_2-distance to the noisy observation y on the subspace $\mathcal{Y}$ as an estimate obeys

$$\|\hat{y} - y_*\|^2 \lesssim \sigma^2 \frac{k}{n}, \tag{1}$$

with high probability (throughout, $\|\cdot\|$ denotes the ℓ_2-norm). Thus, the noise is reduced by a factor of k/n over the trivial estimate $\hat{y} = y$ which does not use any prior knowledge of the signal. The denoising rate (1) shows that the more concise the image prior or image representation (i.e., the smaller k), the more noise can be removed. If on the other hand the prior (the subspace, in this example) does not include the original image y_*, then the error bound (1) increases as we would remove a significant part of the signal along with noise when projecting onto the range of the signal prior. Thus a concise and accurate prior is crucial for denoising.

Since real world signals rarely lie in subspaces, the last few decades of image denoising research have developed more sophisticated and accurate priors and algorithms, for example based on sparse representations in overcomplete dictionaries such as wavelets [Don95] and curvelets [Sta+02] and

based on exploiting self-similarity within images [Dab+07]. A prominent example of the former class is the BM3D [Dab+07] algorithm, which achieves state-of-the-art performance for certain denoising problems. However, the nuances of real world images are difficult to describe. Thus, starting with the paper [EA06] that proposes to learn sparse representation based on training data, it has become common to learn concise representation for denoising from a set of training images.

In 2012, Burger et al. [Bur+12] applied deep networks to the denoising problem, by training a deep network on a large set of images. Since then, deep learning based denoisers [Zha+17] have set the standard for denoising. Intuitively, the success of deep network priors can be attributed to their ability to efficiently represent and learn realistic image priors, for example via auto-decoders [HS06] and generative adversarial models [Goo+14]. Over the last few years, the quality of deep priors has significantly improved [Kar+17; Uly+17]. As this field matures, priors will be developed with even smaller latent code dimensionality and more accurate approximation of natural signal manifolds. Consequently, the representation error from deep priors will decrease, and thereby enable even more powerful denoisers.

While deep networks can model complex priors through many parameters and non-linearities, those non-linearities also make their analysis inherently difficult. There is a corresponding lack of theory explaining the success of deep network based priors.

Contributions: The goal of this paper is to analytically quantify the denoising performance of deep-prior based denoisers. Specifically, we characterize the denoising performance of a simple and efficient algorithm for denoising based on a d-layer generative neural network $G\colon \mathbb{R}^k \to \mathbb{R}^n$, with $k < n$, and random weights. In more detail, we propose a gradient method with a tweak that attempts to minimize the least-squares loss $f(x) = \frac{1}{2}\|G(x) - y\|$ between the noisy image y and an image in the range of the prior, $G(x)$. Albeit f is non-convex, we show that our algorithm yields an estimate $\hat{x}$ obeying

$$\|G(\hat{x}) - y_*\|^2 \lesssim \sigma^2 \frac{k}{n},$$

where the notation $\lesssim$ absorbs a constant factor depending on the number of layers of the network, and its expansitivity, discussed in more detail later. Thus, the denoising rate of deep prior based denoisers is determined by the dimension of the latent representation. We also show in numerical experiments, that this rate—shown to be analytically achieved for random priors—is also experimentally achieved for priors learned from real imaging data.

2 Problem formulation

We consider the problem of estimating a vector $y_* \in \mathbb{R}^n$ from a noisy observation $y = y_* + \eta$. We assume, as a prior, that the vector y_* belongs to the range of a d-layer generative neural network $G\colon \mathbb{R}^k \to \mathbb{R}^n$, with $k < n$. That is, $y_* = G(x_*)$ for some $x_* \in \mathbb{R}^k$. Specifically, we consider a generative network modeled by

$$G(x) = \mathrm{relu}(W_d \ldots \mathrm{relu}(W_2\, \mathrm{relu}(W_1 x_*)) \ldots),$$

where $\mathrm{relu}(x) = \max(x, 0)$ applies entrywise, $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$, n_i is the number of neurons in the ith layer, and $k = n_0 < n_1 < \cdots < n_d = n$. The problem at hand is: Given $W_1 \ldots W_d$ and a noisy observation y, obtain an estimate $\hat{y}$ of y_* such that $\|\hat{y} - y_*\|$ is small.

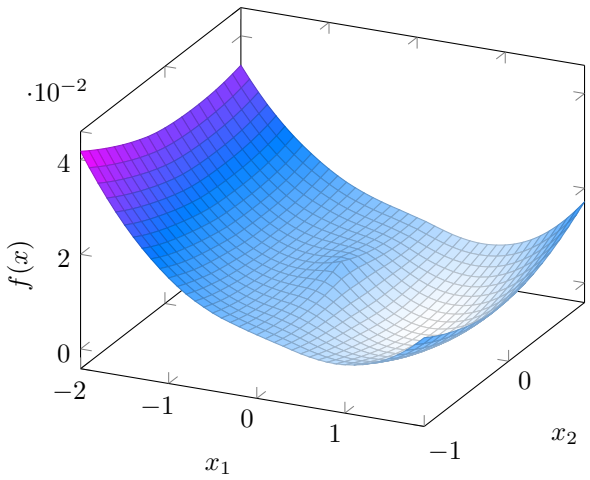

Figure 1: Loss surface $f(x) = \|G(x) - G(x_*)\|$, $x_* = [1, 0]$, of an expansive network G with ReLu activation functions with $k = 2$ nodes in the input layer and $n_2 = 300$ and $n_3 = 784$ nodes in the hidden and output layers, respectively, with random Gaussian weights in each layer. The surface has a critical point near $-x_*$, a global minimum at x_*, and a local maximum at 0.

3 Denoising via empirical risk minimization

As a way to solve the above problem, we first obtain an estimate of x_*, denoted by $\hat{x}$, and then estimate y_* as $G(\hat{x})$. In order to estimate x_*, we consider minimizing the empirical risk objective

$$f(x) := \frac{1}{2}\|G(x) - y\|^2. \tag{2}$$

As this objective is nonconvex, there is no *a priori* guarantee of efficiently finding the global minimum. Approaches such as gradient methods could in principle get stuck in local minima, instead of finding a global minimizer that is close to x_*.

However, as we show in this paper, under appropriate conditions, a gradient method with a little tweak—introduced next—finds a point that is very close to x_*, with the distance to x_* controlled by the noise. In order to state the algorithm, we first introduce a useful quantity. For analyzing which rows of a matrix W are active when computing $\text{relu}(Wx)$, we let

$$W_{+,x} = \text{diag}(Wx > 0)W.$$

For a fixed W, the matrix $W_{+,x}$ zeros out the rows of W that do not have a positive dot product with x. Alternatively put, $W_{+,x}$ contains weights from only the neurons that are active for the input x. We also define $W_{1,+,x} = (W_1)_{+,x} = \text{diag}(W_1 x > 0)W_1$ and

$$W_{i,+,x} = \text{diag}(W_i W_{i-1,+,x} \cdots W_{2,+,x} W_{1,+,x} x > 0)W_i.$$

The matrix $W_{i,+,x}$ consists only of the weights of the neurons in the ith layer that are active if the input to the first layer is x.

We are now ready to state our algorithm: a gradient method with a tweak. Given a noisy observation y, the algorithm starts with an arbitrary initial point $x_0 \neq 0$. At each iteration $i = 0, 1, \ldots$, the algorithm computes the step direction

$$\tilde{v}_{x_i} = (\Pi_{i=d}^{1} W_{i,+,x_i})^t (G(x_i) - y),$$

which is equal to the gradient of f if f is differentiable at x_i. It then takes a small step opposite to $\tilde{v}_{x_i}$. The tweak is that before each iteration, the algorithm checks whether $f(-x_i)$ is smaller than $f(x_i)$, and if so, negates the sign of the current iterate x_i. To understand this step, it is instructive to examine the loss surface for the *noiseless* case in Figure 1. It can be seen that while the loss function has a *global* minimum at x_*, it is relatively flat close to $-x_*$. In expectation, there is a critical point that is a negative multiple of x_* with the property that the curvature in the $\pm x_*$ direction is positive, and the curvature in orthogonal directions is zero. Further, around approximately $-x_*$, the loss function is larger than around the optimum x_* As a gradient descent could potentially get stuck in this region, the negation check provides a way to avoid converging to this region. Our algorithm is formally summarized as Algorithm 1 below.

Algorithm 1 Gradient method

Require: Weights of the network W_i, noisy observation y, and step size $\alpha > 0$

1: Choose an arbitrary initial point $x_0 \in \mathbb{R}^k \backslash \{0\}$
2: **for** $i = 0, 1, \ldots$ **do**
3: **if** $f(-x_i) < f(x_i)$ **then**
4: $x_i \leftarrow -x_i$;
5: **end if**
6: Compute $\tilde{v}_{x_i} = (\Pi_{i=d}^{1} W_{i,+,x_i})^t (G(x_i) - y)$
7: $x_{i+1} = x_i - \alpha \tilde{v}_{x_i}$
8: **end for**

Other variations of the tweak are also possible. For example, the negation check in Step 2 could be performed after a convergence criterion is satisfied, and if a lower objective is achieved by negating the latent code, then the gradient descent can be continued again until the convergence criterion is again satisfied.

4 Main results

For our analysis, we consider a fully-connected generative network $G \colon \mathbb{R}^k \to \mathbb{R}^n$ with Gaussian weights and no bias terms. Specifically, we assume the W_i are independently and identically distributed as $\mathcal{N}(0, 1/n_i)$, but do not require them to be independent across layers. Moreover, we assume that the network is sufficiently *expansive*:

Expansivity condition. *We say that the expansivity condition with constant $\epsilon > 0$ holds if*

$$n_i \geqslant c\epsilon^{-2} \log(1/\epsilon) n_{i-1} \log n_{i-1}, \quad \textit{for all } i,$$

where c is a particular numerical constant.

The motivation for selecting Gaussian weights for our analysis is severalfold:

1. the empirical distribution of weights from deep neural networks often have statistics consistent with Gaussians, AlexNet is a concrete example [Aro+15];
2. random convolutional generative networks have show impressive denoising, inpainting, and superresolution performance when trained against a single image, as done with the Deep Image Prior [Uly+17]; and
3. the field of theoretical analysis of recovery guarantees for deep learning is nascent, and Gaussian networks can permit theoretical results because of well developed theories for random matrices.

We are now ready to state our main result, pertaining to denoising zero-mean Gaussian noise with co-variance matrix σ/nI.

Theorem 1. *Consider a network with the weights in the i-th layer, $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$, i.i.d. $\mathcal{N}(0, 1/n_i)$ distributed, and suppose that the network satisfies the expansivity condition for some $\epsilon \leqslant K/d^{90}$. Also, suppose that the noise variance obeys*

$$\omega \leqslant \frac{\|x_*\| K_1 2^{-d/2}}{d^{16}}, \quad \omega := \sqrt{18\sigma \frac{k}{n} \log(n_1^d n_2^{d-1} \dots n_d)}.$$

Consider the iterates of Algorithm 1 with stepsize $\alpha = K_4 \frac{2^d}{d^2}$. Then, there exists a number of steps N upper bounded by

$$N \leqslant \frac{K_2}{d^4 \epsilon} \frac{f(x_*) 2^d}{\|x_*\|}$$

such that after N steps, the iterates of Algorithm 1 obey

$$\|x_i - x_*\| \leqslant K_5 d^9 \|x_*\| \sqrt{\epsilon} + K_6 d^6 2^{d/2} \omega, \quad \text{for all } i \geqslant N, \tag{3}$$

with probability at least $1 - 2e^{-2k \log n} - \sum_{i=2}^{d} 8n_i e^{-K_7 n_{i-2}} - 8n_1 e^{-K_7 \epsilon^2 \log(1/\epsilon) k}$. Here, $K_1, K_2, ..$ are numerical constants.

The error term in the bound (3) consists of two terms—the first is controlled by ϵ, and the second depends on the noise. The first term is negligible if ϵ is chosen sufficiently small, but that comes at the expense of the expansivity condition being more stringent. The second term in the bound (3) is more interesting and controls the effect of noise. Specifically, for ϵ sufficiently small, our result guarantees that after sufficiently many iterations,

$$\|x_i - x_*\|^2 \lesssim 2^d \sigma^2 \frac{k}{n},$$

where the notation $\lesssim$ absorbs a factor logarithmic in n and linear in d. Noting that G is approximately $2/2^{d/2}$-Lipschitz in a region around x_*, (as established in [Hua+18]), our result guarantees that for ϵ sufficiently small, after sufficiently many iterates,

$$\|G(x_i) - G(x_*)\|^2 \lesssim \sigma^2 \frac{k}{n},$$

Thus, the theorem guarantees that our algorithm yields the denoising rate of $\sigma k/n$, and, as a consequence, denoising based on a generative deep prior provably reduces the energy of the noise in the original image by a factor of k/n.

We hasten to add that the factors of 2^d in the theorem are present because the weights of the coefficients of the matrices W_i have variance $1/n_i$. As a result, the operation relu(Wx) returns approximately half of the entries of Wx. Because of this, $G(x)$ scales like $2^{-d/2}\|x\|$, the noise ω scales like $2^{-d/2}$, ∇f scales like 2^d, and α scales like 2^d. All of these scalings would be unity with an alternate choice of the variance of the entries of W_i.

4.1 The Weight Distribution Condition (WDC)

To prove our main result, we make use of a deterministic condition on G, called the Weight Distribution Condition (WDC), and then show that Gaussian W_i of appropriate sizes satisfy the WDC with the appropriate probability, provided the expansivity condition holds. Our main result, Theorem 1, continues to hold for any matrices W_i that satisfy the WDC.

The condition is on the spatial arrangement of the network weights within each layer. We say that the matrix $W \in \mathbb{R}^{n\times k}$ satisfies the *Weight Distribution Condition* with constant ϵ if for all nonzero $x, y \in \mathbb{R}^k$,

$$\left\| \sum_{i=1}^{n} 1_{\langle w_i, x\rangle > 0} 1_{\langle w_i, y\rangle > 0} \cdot w_i w_i^t - Q_{x,y} \right\| \leqslant \epsilon, \text{ with } Q_{x,y} = \frac{\pi - \theta_0}{2\pi} I_k + \frac{\sin\theta_0}{2\pi} M_{\hat{x}\leftrightarrow\hat{y}}, \tag{4}$$

where $w_i \in \mathbb{R}^k$ is the ith row of W; $M_{\hat{x}\leftrightarrow\hat{y}} \in \mathbb{R}^{k\times k}$ is the matrix[1] such that $\hat{x} \mapsto \hat{y}$, $\hat{y} \mapsto \hat{x}$, and $z \mapsto 0$ for all $z \in \text{span}(\{x, y\})^\perp$; $\hat{x} = x/\|x\|_2$ and $\hat{y} = y/\|y\|_2$; $\theta_0 = \angle(x, y)$; and 1_S is the indicator function on S. The norm in the left hand side of (4) is the spectral norm. Note that an elementary calculation[2] gives that $Q_{x,y} = \mathbb{E}[\sum_{i=1}^n 1_{\langle w_i,x\rangle>0} 1_{\langle w_i,y\rangle>0} \cdot w_i w_i^t]$ for $w_i \sim \mathcal{N}(0, I_k/n)$. As the rows w_i correspond to the neural network weights of the ith neuron in a layer given by W, the WDC provides a deterministic property under which the set of neuron weights within the layer given by W are distributed approximately like a Gaussian. The WDC could also be interpreted as a deterministic property under which the neuron weights are distributed approximately like a uniform random variable on a sphere of a particular radius. Note that if $x = y$, $Q_{x,y}$ is an isometry up to a factor of $1/2$.

5 Extensions

In this section we briefly discuss another important scenario to which our results apply to, namely regularizing inverse problems using deep generative priors. Approaches that regularize inverse problems using deep generative models [Bor+17] have empirically been shown to improve over sparsity-based approaches, see [Luc+18] for a review for applications in imaging, and [Mar+17] for an application in Magnetic Resonance Imaging showing a significant performance improvement over conventional methods.

Consider an inverse problem, where the goal is to reconstruct an unknown vector $y_* \in \mathbb{R}^n$ from $m < n$ noisy linear measurements:

$$z = Ay_* + \eta \quad \in \mathbb{R}^m,$$

where $A \in \mathbb{R}^{m\times n}$ is called the measurement matrix and η is zero mean Gaussian noise with covariance matrix σ^2/nI, as before. As before, assume that y_* lies in the range of a generative prior G, i.e., $y_* = G(x_*)$ for some x_*. As a way to recover x_*, consider minimizing the empirical risk objective $f(x) = \frac{1}{2}\|AG(x) - z\|$, using Algorithm 1, with Step 6 substituted by $\tilde{v}_{x_i} = (A\Pi_{i=d}^1 W_{i,+,x_i})^t(AG(x_i) - y)$, to account for the fact that measurements were taken with the matrix A.

Suppose that A is a random projection matrix, for concreteness assume that A has i.i.d. Gaussian entries with variance $1/m$. Using an analogous argument than used for the proof of Theorem 1, we can show that Theorem 1 continues to hold with $\omega = \sqrt{18\sigma\frac{k}{m}\log(n_1^d n_2^{d-1}\ldots n_d)}$, (note that n has been replaced by m), and slightly adopting the success probability of the statement. This extension

[1] A formula for $M_{\hat{x}\leftrightarrow\hat{y}}$ is as follows. If $\theta_0 = \angle(\hat{x}, \hat{y}) \in (0, \pi)$ and R is a rotation matrix such that $\hat{x}$ and $\hat{y}$ map to e_1 and $\cos\theta_0 \cdot e_1 + \sin\theta_0 \cdot e_2$ respectively, then $M_{\hat{x}\leftrightarrow\hat{y}} = R^t \begin{pmatrix} \cos\theta_0 & \sin\theta_0 & 0 \\ \sin\theta_0 & -\cos\theta_0 & 0 \\ 0 & 0 & 0_{k-2} \end{pmatrix} R$, where 0_{k-2} is a $k-2\times k-2$ matrix of zeros. If $\theta_0 = 0$ or π, then $M_{\hat{x}\leftrightarrow\hat{y}} = \hat{x}\hat{x}^t$ or $-\hat{x}\hat{x}^t$, respectively.

[2] To do this calculation, take $x = e_1$ and $y = \cos\theta_0 \cdot e_1 + \sin\theta_0 \cdot e_2$ without loss of generality. Then each entry of the matrix can be determined analytically by an integral that factors in polar coordinates.

shows that, provided ϵ is chosen sufficiently small, that our algorithm yields an iterate x_i obeying

$$\|G(x_i) - G(x_*)\|^2 \lesssim \sigma^2 \frac{k}{m},$$

where again $\lesssim$ absorbs factors logarithmic in the n_i's, and linear in d.

This extension of our result shows that Algorithm 1 enables solving inverse problems under noise efficiently, and quantifies the effect of the noise.

6 Experimental results

In this section we provide experimental evidence that corroborates our theoretical claims that denoising with deep priors achieves a denoising rate proportional to $\sigma^2 k/n$. We consider both a synthetic, random prior, as studied theoretically in the paper, as well as a prior learned from data. All our results are reproducible with the code provided in the supplement.

6.1 Denoising with a synthetic prior

We start with a synthetic generative network prior with ReLu-activation functions, and draw its weights independently from a Gaussian distribution. We consider a three layer network with $n = 782$ neurons in the output layer, 300 in the middle layer, and vary the number of input neurons, k. We next present simulations showing that if k is sufficiently small, our algorithm recovers the latent representation and image in the noiseless case, and, in the noisy case, achieves a denoising rate proportional to $\sigma k/n$ as guaranteed by our theory.

Towards this goal, we generate Gaussian inputs x_* to the network and observe the noisy image $y = G(x_*) + \eta$, $\eta \sim \mathcal{N}(0, \sigma^2/nI)$. From the noisy image, i) we obtain an estimate $\hat{x}$ of the latent representation by running Algorithm 1 until convergence, and then ii) we obtain an estimate of the image as $\hat{y} = G(\hat{x})$. In the right panel of Figure 3, we depict the normalized mean squared error of the latent representation, $\text{MSE}(\hat{x}, x_*)$, and the mean squared error in the image domain, $\text{MSE}(G(\hat{x}), G(x_*))$, where we defined $\text{MSE}(z, z') = \|z - z'\|^2/\|z'\|^2$.

The results in Figure 3, left panel, show that, if the network is sufficiently expansive, guaranteed by k being sufficiently small, then in the noiseless case ($\sigma^2 = 0$), the latent representation and image are perfectly recovered. In the noisy case ($\sigma^2 = 0.25$, corresponding to a signal-to-noise ratio of 4), we achieve a MSE proportional to $\sigma^2 k/n$. Moreover, even if the network is not sufficiently expansive, we achieve a MSE of less than $\sigma^2 k/n$, albeit only in the image domain.

We also observed that for the problem instances considered here, the negation trick in step 3-4 of Algorithm 1 is often not necessary, in that even without that step the algorithm typically converges to the global minimum.

6.2 Denoising with a learned prior

We next consider a prior learned from data, and show that even when using that learned prior we achieve the rate predicted by our theory pertaining to a random prior. Towards this goal, we consider two different fully-connected autoencoders parameterized by k, consisting of an decoder and encoder with ReLu activation functions and fully connected layers. We choose the number of neurons in the three layers of the encoder as $784, 400, k$, and those of the decoder as $k, 400, 784$. We set $k = 10$ and $k = 20$ for the two different autoencoders. We trained both autoencoders on the MNIST [Lec+98] *training set*.

We then take an image y_* from the MNIST *test set*, add Gaussian noise to it, and denoise it using our method based on the learned decoder-network G for $k = 10$ and $k = 20$. Specifically, we

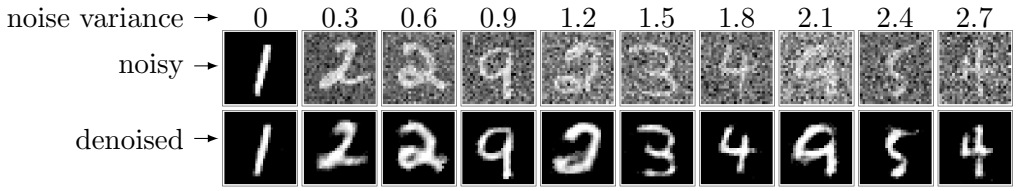

Figure 2: Denosing with a learned generative prior: Even when the number is barely visible, the denoiser recovers a sharp image.

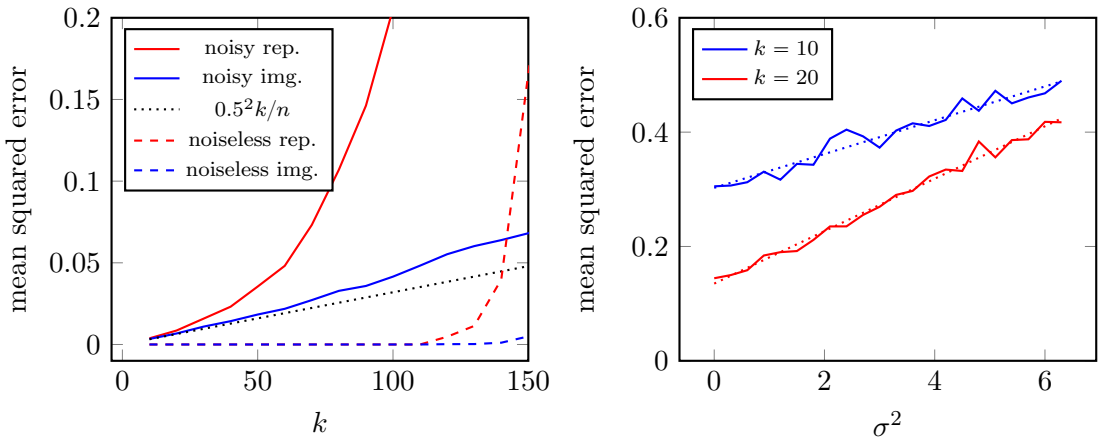

Figure 3: **Left panel:** Mean square error in the image domain, $\text{MSE}(G(\hat{x}), x_*)$, and in the latent representation, $\text{MSE}(\hat{x}, x_*)$, as a function of the dimension of the latent representation, k, for denoising without noise, and with noise $\sigma^2 = 0.5^2$. As predicted by the theory, if k is sufficiently small, and thus the network is sufficiently expansive, the denoising rate is proportional to $\sigma^2 k/n$. **Right panel:** Denoising of handwritten digits based on a learned decoder with $k = 10$ and $k = 20$, along with the least-squares fit as dotted lines. The learned decoder with $k = 20$ has more parameters and thus represents the images with a smaller error; therefore the MSE at $\sigma = 0$ is smaller. However, the denoising rate for the decoder with $k = 20$, which is the slope of the curve is larger as well, as predicted by our theory.

estimate the latent representation $\hat{x}$ by running Algorithm 1, and then set $\hat{y} = G(\hat{x})$. See Figure 2 for a few examples demonstrating the performance of our approach for different noise levels.

We next show that our method achieves a mean squared error (MSE) proportional to $\sigma^2 k/n$, as predicted by our theory. We add noise to the images with noise variance ranging from $\sigma^2 = 0$ to $\sigma^2 = 6$. In the right panel of Figure 3 we show the MSE in the image domain, $\text{MSE}(G(\hat{x}), G(x_*))$, averaged over a number of images for the learned decoders with $k = 10$ and $k = 20$. We observe an interesting tradeoff: The decoder with $k = 10$ has fewer parameters, and thus does not represent the digits as well, therefore the MSE is larger than that for $k = 20$ for the noiseless case (i.e., for $\sigma = 0$). On the other hand, the smaller number of parameters results in a better denoising rate (by about a factor of two), corresponding to the steeper slope of the MSE as a function of the noise variance, σ^2.

7 Proofs

In this section we prove our main result, Theorem 1. As mentioned in Section 4.1, our proof makes use of a deterministic condition, called the Weight Distribution Condition (WDC), formally defined in Section 4.1. The following proposition establishes that the expansivity condition ensures that the WDC holds:

Lemma 2 (Lemma 9 in [HV18]). *Fix $\epsilon \in (0,1)$. If the entires of $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ are i.i.d. $\mathcal{N}(0, 1/n_i)$ and the expansivity condition $n_i > c\epsilon^{-2}\log(1/\epsilon) n_{i-1} \log n_{i-1}$ holds, then W_i satisfies the WDC with constant ϵ with probability at least $1 - 8n_i e^{-K\epsilon^2 n_{i-1}}$. Here, c and K are numerical constants.*

We note that the form of dependence of n_i on ϵ can be read off the proofs of Lemma 10 in [HV18]. It follows from Lemma 2, that the WDC holds for all W_i with probability at least $1 - \sum_{i=2}^{d} 8n_i e^{-K_7 n_{i-2}} - 8n_1 e^{-K_7 \epsilon^2 \log(1/\epsilon) k}$.

In the remainder of the proof we work on the event that the WDC holds for all W_i.

7.1 Preliminaries

Recall that the goal of our algorithm is to minimize the empirical risk objective

$$f(x) = \frac{1}{2}\|G(x) - y\|^2, \tag{5}$$

where $y := G(x_*) + \eta$, with $\eta \sim \mathcal{N}(0, \sigma^2/nI)$.

Our results rely on the fact that outside of two balls around $x = x_*$ and $x = -\rho_d x_*$, with ρ_d a constant defined below, the direction chosen by the algorithm is a descent direction, with high probability. Towards this goal, we use a concentration argument, similar to the arguments used in [HV18]. First, define $\Lambda_x := \Pi_{i=d}^{1} W_{i,+,x}$ (with $W_{i,+,x}$ defined in Section 3) for notational convenience, and note that the step direction of our algorithm can be written as

$$\tilde{v}_x = \overline{v}_x + \bar{q}_x, \quad \text{with} \quad \overline{v}_x := \Lambda_x^t \Lambda_x x - (\Lambda_x)^t (\Lambda_{x_*}) x_*, \quad \text{and} \quad \bar{q}_x := \Lambda_x^t \eta. \tag{6}$$

Note that at points x where G (and hence f) is differentiable, we have that $\tilde{v}_x = \nabla f(x)$.

The proof is based on showing that $\tilde{v}_x$ concentrates around a particular $h_x \in \mathbb{R}^k$, defined below, that is a continuous function of nonzero x, x_* and is zero only at $x = x_*$ and $x = -\rho_d x_*$. The definition of h_x depends on a function that is helpful for controlling how the operator $x \mapsto W_{+,x} x$ distorts angles, defined as:

$$g(\theta) := \cos^{-1}\Big(\frac{(\pi - \theta)\cos\theta + \sin\theta}{\pi}\Big). \tag{7}$$

With this notation, we define

$$h_x := -\frac{1}{2^d}\Big(\prod_{i=0}^{d-1} \frac{\pi - \overline{\theta}_i}{\pi}\Big) x_* + \frac{1}{2^d}\left[x - \sum_{i=0}^{d-1} \frac{\sin\overline{\theta}_i}{\pi}\Big(\prod_{j=i+1}^{d-1} \frac{\pi - \overline{\theta}_j}{\pi}\Big)\frac{\|x_*\|_2}{\|x\|_2} x\right],$$

where $\overline{\theta}_0 = \angle(x, x_*)$ and $\overline{\theta}_i = g(\overline{\theta}_{i-1})$. Note that h_x is deterministic and only depends on x, x_*, and the number of layers, d.

In order to bound the deviation of $\tilde{v}_x$ from h_x we use the following two lemmas, bounding the deviation controlled by the WDC and the deviation from the noise:

Lemma 3 (Lemma 6 in [HV18]). *Suppose that the WDC holds with $\epsilon < 1/(16\pi d^2)^2$. Then, for all nonzero $x, x_* \in \mathbb{R}^k$,*

$$\|\overline{v}_x - h_x\|_2 \leqslant K\frac{d^3\sqrt{\epsilon}}{2^d}\max(\|x\|_2, \|x_*\|_2), \text{ and} \tag{8}$$

$$\left\langle (\Pi_{i=d}^1 W_{i,+,x})x, (\Pi_{i=d}^1 W_{i,+,x_*})x_* \right\rangle \geqslant \frac{1}{4\pi}\frac{1}{2^d}\|x\|_2\|x_*\|_2, \text{ and} \tag{9}$$

$$\left\|\Pi_{i=d}^1 W_{i,+,x}\right\|^2 \leqslant \frac{1}{2^d}(1+2\epsilon)^d \leqslant \frac{13}{12}2^{-d}. \tag{10}$$

Proof. Equation (8) and (9) are Lemma 6 in [HV18]. Regarding (10), note that the WDC implies that $\|W_{i,+,x}\|^2 \leqslant 1/2 + \epsilon$. It follows that

$$\left\|\Pi_{i=d}^1 W_{i,+,x}\right\|^2 \leqslant \frac{1}{2^d}(1+2\epsilon)^d = \frac{1}{2^d}e^{d\log(1+2\epsilon)} \leqslant \frac{1+4\epsilon d}{2^d} \leqslant \frac{13}{12}2^{-d},$$

where the last inequalities follow by our assumption on ϵ. □

Lemma 4. *Suppose the WDC holds with $\epsilon < 1/(16\pi d^2)^2$, that any subset of n_{i-1} rows of W_i are linearly independent for each i, and that $\eta \sim \mathcal{N}(0, \sigma^2/nI)$. Then the event*

$$\mathcal{E}_{noise} := \left\{ \left\|(\Pi_{i=d}^1 W_{i,+,x})^t \eta\right\| \leqslant \frac{\omega}{2^{d/2}}, \text{ for all } x \right\}, \quad \omega := \sqrt{16\sigma\frac{k}{n}\log(n_1^d n_2^{d-1}\dots n_d)} \tag{11}$$

holds with probability at least $1 - 2e^{-2k\log n}$.

As the cost function f is not differentiable everywhere, we will make use of the generalized subdifferential in order to reference the subgradients at nondifferentiable points. For a Lipschitz function $\tilde{f}$ defined from a Hilbert space $\mathcal{X}$ to $\mathbb{R}$, the Clarke generalized directional derivative of $\tilde{f}$ at the point $x \in \mathcal{X}$ in the direction u, denoted by $\tilde{f}^o(x;u)$, is defined by $\tilde{f}^o(x;u) = \limsup_{y\to x, t\downarrow 0} \frac{\tilde{f}(y+tu)-\tilde{f}(y)}{t}$, and the generalized subdifferential of $\tilde{f}$ at x, denoted by $\partial\tilde{f}(x)$, is defined by

$$\partial\tilde{f}(x) = \{v \in \mathbb{R}^k \mid \langle v, u\rangle \leqslant \tilde{f}^o(x;u), \text{ for all } u \in \mathcal{X}\}.$$

Since $f(x)$ is a piecewise quadratic function, we have

$$\partial f(x) = \text{conv}(v_1, v_2, \dots, v_t), \tag{12}$$

where conv denotes the convex hull of the vectors $v_1, \dots, v_t$, t is the number of quadratic functions adjoint to x, and v_i is the gradient of the i-th quadratic function at x.

Lemma 5. *Under the assumption of Lemma 4, and assuming that $\mathcal{E}_{noise}$ holds, we have that, for any $x \neq 0$ and any $v_x \in \partial f(x)$,*

$$\|v_x - h_x\| \leqslant K\frac{d^3\sqrt{\epsilon}}{2^d}\max(\|x\|_2, \|x_*\|_2) + \frac{\omega}{2^{d/2}}.$$

In particular, this holds for the subgradient $v_x = \tilde{v}_x$.

Proof. By (12), $\partial f(x) = \text{conv}(v_1, \dots v_t)$ for some finite t, and thus $v_x = a_1 v_1 + \dots a_t v_t$ for some $a_1, \dots, a_t \geqslant 0$, $\sum_i a_i = 1$. For each v_i, there exists a w such that $v_i = \lim_{t\downarrow 0} \tilde{v}_{x+tw}$. On the event $\mathcal{E}_{\text{noise}}$, we have that for any $x \neq 0$, for any $\tilde{v}_x \in \partial f(x)$

$$\begin{aligned}
\|\tilde{v}_x - h_x\| &= \|\overline{v}_x + \bar{q}_x - h_x\| \\
&\leqslant \|\overline{v}_x - h_x\| + \|\bar{q}_x\| \\
&\leqslant K\frac{d^3\sqrt{\epsilon}}{2^d}\max(\|x\|_2, \|x_*\|_2) + \frac{\omega}{2^{d/2}},
\end{aligned}$$

where the last inequality follows from Lemmas 3 and 4 above. The proof is concluded by appealing to the continuity of h_x with respect to nonzero x, and by noting that

$$\|v_x - h_x\| \leqslant \sum_i a_i \|v_i - h_x\| \leqslant K \frac{d^3\sqrt{\epsilon}}{2^d} \max(\|x\|_2, \|x_*\|_2) + \frac{\omega}{2^{d/2}},$$

where we used the inequality above and that $\sum_i a_i = 1$. □

We will also need an upper bound on the norm of the step direction of our algorithm:

Lemma 6. *Suppose that the WDC holds with $\epsilon < 1/(16\pi d^2)^2$ and that the event $\mathcal{E}_{noise}$ holds with $\omega \leqslant \frac{2^{-d/2}\|x_*\|}{8\pi}$. Then, for all x,*

$$\|\tilde{v}_x\| \leqslant \frac{dK}{2^d} \max(\|x\|, \|x_*\|), \tag{13}$$

where K is a numerical constant.

Proof. Define for convenience $\zeta_j = \prod_{i=j}^{d-1} \frac{\pi - \bar{\theta}_{j,x,x_*}}{\pi}$. We have

$$\begin{aligned}
\|\tilde{v}_x\| \leqslant & \|h_x\| + \|h_x - \tilde{v}_x\| \\
\leqslant & \left\| \frac{1}{2^d} x - \frac{1}{2^d}\zeta_0 x_* - \frac{1}{2^d} \sum_{i=0}^{d-1} \frac{\sin\bar{\theta}_{i,x}}{\pi} \zeta_{i+1} \frac{\|x_*\|}{\|x\|} x \right\| + K_1 \frac{d^3\sqrt{\epsilon}}{2^d} \max(\|x\|_2, \|x_*\|_2) + \frac{\omega}{2^{d/2}} \\
\leqslant & \frac{1}{2^d}\|x\| + \left(\frac{1}{2^d} + \frac{d}{\pi 2^d}\right)\|x_*\| + K_1 \frac{d^3\sqrt{\epsilon}}{2^d} \max(\|x\|, \|x_*\|) + \frac{\omega}{2^{d/2}} \\
\leqslant & \frac{dK}{2^d} \max(\|x\|, \|x_*\|),
\end{aligned}$$

where the second inequality follows from the definition of h_x and Lemma 5, the third inequality uses $|\zeta_j| \leqslant 1$, and the last inequality uses the assumption $\omega \leqslant \frac{2^{-d/2}\|x_*\|}{8\pi}$. □

7.2 Proof of Theorem 1

We are now ready to prove Theorem 1. The logic of the proof is illustrated in Figure 4. Recall that x_i is the ith iterate of x as per Algorithm 1. We first ensure that we can assume throughout that x_i is bounded away from zero:

Lemma 7. *Suppose that WDC holds with $\epsilon < 1/(16\pi d^2)^2$ and that $\mathcal{E}_{noise}$ holds with ω in (11) obeying $\omega \leqslant \frac{2^{-d/2}\|x_*\|}{8\pi}$. Moreover, suppose that the step size in Algorithm 1 satisfies $0 < \alpha < \frac{K2^d}{d^2}$, where K is a numerical constant. Then, after at most $N = (\frac{38\pi K_0 2^d}{\alpha})^2$ steps, we have that for all $i > N$ that $x_i \notin \mathcal{B}(0, K_0\|x_*\|)$, $K_0 = \frac{1}{32\pi}$.*

In particular, if $\alpha = K2^d/d^2$, then N is bounded by a constant times d^4.

We can therefore assume throughout this proof that $x_i \notin \mathcal{B}(0, K_0\|x_*\|)$, $K_0 = \frac{1}{32\pi}$. We prove Theorem 1 by showing that if $\|h_x\|$ is sufficiently large, i.e., if the iterate x_i is outside of set

$$\mathcal{S}_\beta = \left\{ x \in \mathbb{R}^k \mid \|h_x\| \leqslant \frac{1}{2^d}\beta \max(\|x\|, \|x_*\|) \right\},$$

with

$$\beta = 4Kd^3\sqrt{\epsilon} + 13\omega 2^{d/2}/\|x_*\|, \tag{14}$$

then the algorithm makes progress in the sense that $f(x_{i+1}) - f(x_i)$ is smaller than a certain negative value. The set $\mathcal{S}_\beta$ is contained in two balls around x_* and $-\rho x_*$, whose radius is controlled by β:

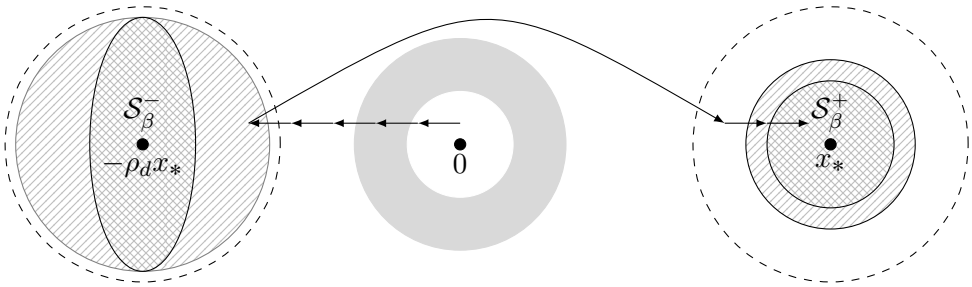

Figure 4: Logic of the proof: Starting at an arbitrary point, Algorithm 1 moves away from 0, at least till its iterates are outside the gray ring, as 0 is a local maximum; and once an iterate x_i leaves the gray ring around 0, all subsequent iterates will never be in the white circle around 0 again (see Lemma 7). Then the algorithm might move towards $-\rho_d x_*$, but once it enters the dashed ball around $-\rho_d x_*$, it enters a region where the function value is strictly larger than that of the dashed ball around x_*, by Lemma 9. Thus steps 3-5 of the algorithm will ensure that the next iterate x_i is in the dashed ball around x_*. From there, the iterates will move into the region $\mathcal{S}_\beta^+$, since outside of $\mathcal{S}_\beta^+ \cup \mathcal{S}_\beta^-$ the algorithm chooses a descent direction in each step (see the argument around equation (17)). The region $\mathcal{S}_\beta^+$ is covered by a ball of radius r, by Lemma 8, determined by the noise and ϵ.

Lemma 8. *For any* $\beta \leqslant \frac{1}{64^2 d^{12}}$,

$$\mathcal{S}_\beta \subset \mathcal{B}(x_*, 5000 d^6 \beta \|x_*\|_2) \cup \mathcal{B}(-\rho_d x_*, 500 d^{11} \sqrt{\beta} \|x_*\|_2). \tag{15}$$

Here, $\rho_d > 0$ *is defined in the proof and obeys* $\rho_d \to 1$ *as* $d \to \infty$.

Note that by the assumption $\omega \leqslant \frac{\|x_*\| K_1 2^{-d/2}}{d^{16}}$ and $K d^{45} \sqrt{\epsilon} \leqslant 1$, our choice of β in (14) obeys $\beta \leqslant \frac{1}{64^2 d^{12}}$ for sufficiently small K_1, K, and thus Lemma 8 yields:

$$\mathcal{S}_\beta \subset \mathcal{B}(x_*, r) \cup \mathcal{B}(-\rho_d x_*, \sqrt{r \|x_*\|} d^8).$$

were we define the radius $r = K_2 d^9 \sqrt{\epsilon} \|x_*\| + K_3 d^6 \omega 2^{d/2}$, where K_2, K_3 are numerical constants. Note that hat the radius r is equal to the right hand side in the error bound (3) in our theorem. In order to guarantee that the algorithm converges to a ball around x_*, and not to that around $-\rho_d x_*$, we use the following lemma:

Lemma 9. *Suppose that the WDC holds with* $\epsilon < 1/(16\pi d^2)^2$. *Moreover suppose that* $\mathcal{E}_{noise}$ *holds, and that* ω *in the event* $\mathcal{E}_{noise}$ *obeys* $\frac{\omega}{2^{-d/2}\|x_*\|_2} \leqslant K_9/d^2$, *where* $K_9 < 1$ *is a universal constant. Then for any* $\phi_d \in [\rho_d, 1]$, *it holds that*

$$f(x) < f(y) \tag{16}$$

for all $x \in \mathcal{B}(\phi_d x_*, K_3 d^{-10} \|x_*\|)$ *and* $y \in \mathcal{B}(-\phi_d x_*, K_3 d^{-10} \|x_*\|)$, *where* $K_3 < 1$ *is a universal constant.*

In order to apply Lemma 9, define for convenience the two sets:

$$\mathcal{S}_\beta^+ := \mathcal{S}_\beta \cap \mathcal{B}(x_*, r), \text{ and}$$
$$\mathcal{S}_\beta^- := \mathcal{S}_\beta \cap \mathcal{B}(-\rho_d x_*, \sqrt{r \|x_*\|} d^8).$$

By the assumption that $Kd^{45}\sqrt{\epsilon} \leqslant 1$ and $\omega \leqslant K_1 d^{-16} 2^{-d/2}\|x_*\|$, we have that for sufficiently small K_1, K,

$$\mathcal{S}_\beta^+ \subseteq \mathcal{B}(x_*, K_3 d^{-10}\|x_*\|) \quad \text{and} \quad \mathcal{S}_\beta^- \subseteq \mathcal{B}(-\rho_d x_*, K_3 d^{-10}\|x_*\|).$$

Thus, the assumptions of Lemma 9 are met, and the lemma implies that for any $x \in \mathcal{S}_\beta^-$ and $y \in \mathcal{S}_\beta^+$, it holds that $f(x) > f(y)$. We now show that the algorithm converges to a point in $\mathcal{S}_\beta^+$. This fact and the negation step in our algorithm (line 3-5) establish that the algorithm converges to a point in $\mathcal{S}_\beta^+$ if we prove that the objective is nonincreasing with iteration number, which will form the remainder of this proof.

Consider i such that $x_i \notin \mathcal{S}_\beta$. By the mean value theorem [Cla17, Theorem 8.13], there is a $t \in [0,1]$ such that for $\hat{x}_i = x_i - t\alpha\tilde{v}_{x_i}$ there is a $v_{\hat{x}_i} \in \partial f(\hat{x}_i)$, where ∂f is the generalized subdifferential of f, obeying

$$\begin{aligned} f(x_i - \alpha\tilde{v}_{x_i}) - f(x_i) =& \langle v_{\hat{x}_i}, -\alpha\tilde{v}_{x_i}\rangle \\ =& \langle \tilde{v}_{x_i}, -\alpha\tilde{v}_{x_i}\rangle + \langle v_{\hat{x}_i} - \tilde{v}_{x_i}, -\alpha\tilde{v}_{x_i}\rangle \\ \leqslant& -\alpha\|\tilde{v}_{x_i}\|^2 + \alpha\|v_{\hat{x}_i} - \tilde{v}_{x_i}\|\|\tilde{v}_{x_i}\| \\ =& -\alpha\|\tilde{v}_{x_i}\|(\|\tilde{v}_{x_i}\| - \|v_{\hat{x}_i} - \tilde{v}_{x_i}\|). \end{aligned} \tag{17}$$

In the next subsection, we guarantee that for any $t \in [0,1]$, $v_{\hat{x}_i}$ with $\hat{x}_i = x_i - t\alpha\tilde{v}_{x_i}$ is close to $\tilde{v}_{x_i}$:

$$\|v_{\hat{x}_i} - \tilde{v}_{x_i}\| \leqslant \left(\frac{5}{6} + \alpha K_7 \frac{d^2}{2^d}\right)\|\tilde{v}_{x_i}\|, \text{ for all } v_{\hat{x}_i} \in \partial f(\hat{x}_i). \tag{18}$$

Applying (18) to (17) yields

$$f(x_i - \alpha\tilde{v}_{x_i}) - f(x_i) \leqslant -\frac{1}{12}\alpha\|\tilde{v}_{x_i}\|_2^2,$$

where we used that $\alpha K_7 \frac{d^2}{2^d} \leqslant \frac{1}{12}$, by our assumption on the stepsize α being sufficiently small.

Thus, the maximum number of iterations for which $x_i \notin \mathcal{S}_\beta$ is $f(x_0)12/(\alpha \min_i \|\tilde{v}_{x_i}\|^2)$. We next lower-bound $\|\tilde{v}_{x_i}\|$. We have that on $\mathcal{E}_{\text{noise}}$, for all $x \notin \mathcal{S}_\beta$, with β given by (14).

$$\begin{aligned} \|\tilde{v}_x\|_2 &\geqslant \|h_x\| - \|h_x - \tilde{v}_x\| \\ &\geqslant 2^{-d}\max(\|x\|, \|x_*\|)\left(\beta - K_1 d^3\sqrt{\epsilon} - \omega\frac{2^{d/2}}{\|x_*\|}\right) \\ &\geqslant 2^{-d}\max(\|x\|, \|x_*\|)\left(3Kd^3\sqrt{\epsilon} + 12\omega\frac{2^{d/2}}{\|x_*\|}\right) \\ &\geqslant 2^{-d}\|x_*\|3Kd^3\sqrt{\epsilon} \end{aligned} \tag{19}$$

where the second inequality follows by the definition of $\mathcal{S}_\beta$ and Lemma 5, and the third inequality follows from our definition of β in (14). Thus,

$$f(x_i - \alpha\tilde{v}_{x_i}) - f(x_i) \leqslant -\alpha K_5 2^{-2d} d^6 \epsilon\|x_*\|^2 \leqslant -2^{-d} d^4 K_6 \epsilon\|x_*\|^2$$

where we used $\alpha = K_4\frac{2^d}{d^2}$. Hence, there can be at most $\frac{f(x_0)2^d}{K_6 d^4\epsilon\|x_*\|^2}$ iterations for which $x_i \notin \mathcal{S}_\beta$.

In order to conclude our proof, we remark that once x_i is inside a ball of radius r around x_*, the iterates do not leave a ball of radius $2r$ around x_*. To see this, note that by (13) and our choice of stepsize,

$$\alpha\|\tilde{v}_{x_i}\| \leqslant \frac{K}{d}\max(\|x_i\|, \|x_*\|).$$

This concludes our proof.

The remainder of the proof is devoted to prove the lemmas used in this section.

7.3 Proof of Equation (18)

Our proof relies on h_x being Lipschitz, as formalized by the lemma below, which is proven in Section A.5:

Lemma 10. *For any $x, y \notin \mathcal{B}(0, K_0\|x_*\|)$, where K_0 and K_4 are numerical constants,*

$$\|h_x - h_y\| \leqslant \frac{K_4 d^2}{2^d}\|x - y\|.$$

By Lemma 10, for all $t \in [0,1]$ and $i > N$ (recall that by Lemma 7, after at most N steps, $x_i \neq \mathcal{B}(0, K_0\|x_*\|)$):

$$\|h_{\hat{x}_i} - h_{x_i}\| \leqslant \frac{K_4 d^2}{2^d}\|\hat{x}_i - x_i\|, \tag{20}$$

where $\hat{x}_i = x_i - t\alpha\tilde{v}_{x_i}$. Thus, we have that on $\mathcal{E}_{\text{noise}}$, for any $v_{\hat{x}_i} \in \partial f(\hat{x}_i)$ by Lemma 5,

$$\begin{aligned}
\|v_{\hat{x}_i} - \tilde{v}_{x_i}\| \leqslant& \|v_{\hat{x}_i} - h_{\hat{x}_i}\| + \|h_{\hat{x}_i} - h_{x_i}\| + \|h_{x_i} - \tilde{v}_{x_i}\| \\
\leqslant& K_1\frac{d^3\sqrt{\epsilon}}{2^d}\max(\|\hat{x}_i\|, \|x_*\|) + \frac{\omega}{2^{d/2}} + \frac{K_4 d^2}{2^d}\|\hat{x}_i - x_i\| + K_1\frac{d^3\sqrt{\epsilon}}{2^d}\max(\|x_i\|, \|x_*\|) + \frac{\omega}{2^{d/2}} \\
\leqslant& K_1\frac{d^3\sqrt{\epsilon}}{2^d}\max(\|x_i\| + \alpha\|\tilde{v}_{x_i}\|, \|x_*\|) + \frac{K_4 d^2}{2^d}\alpha\|\tilde{v}_{x_i}\| + K_1\frac{d^3\sqrt{\epsilon}}{2^d}\max(\|x_i\|, \|x_*\|) + 2\frac{\omega}{2^{d/2}} \\
\leqslant& K_1\frac{d^3\sqrt{\epsilon}}{2^d}\left(2 + \frac{\alpha d K}{2^d}\right)\max(\|x_i\|, \|x_*\|) + \frac{K_4 d^2}{2^d}\alpha\|\tilde{v}_{x_i}\| + 2\frac{K_9/d^2}{2^d}\|x_*\|
\end{aligned} \tag{21}$$

where the second inequality is from Lemma 5 and Equation 20, and the fourth inequality is from (13) and the assumption $\frac{\omega}{2^{-d/2}\|x_*\|_2} \leqslant K_9/d^2$.

Combining (21) and (19), we get that

$$\|v_{\hat{x}_i} - \tilde{v}_{x_i}\| \leqslant \left(\frac{5}{6} + \alpha K_7\frac{d^2}{2^d}\right)\|\tilde{v}_{x_i}\|,$$

with the appropriate constants chosen sufficiently small. This concludes the proof of Equation (18).

7.4 Proof of Lemma 7

First suppose that $x_i \in \mathcal{B}(0, 2K_0\|x_*\|)$. We show that after a polynomial number of iterations N, we have that $x_{i+N} \notin \mathcal{B}(0, 2K_0\|x_*\|)$. Below, we prove that

$$\langle x, \tilde{v}_x\rangle < 0 \text{ and } \|\tilde{v}_x\| \geqslant \frac{1}{2^d 16\pi}\|x_*\| \text{ for all } x \in \mathcal{B}(0, 2K_0\|x_*\|). \tag{22}$$

It follows that for any $x_i \in \mathcal{B}(0, 2K_0\|x_*\|)$, x_i and the next iterate produced by the algorithm, $x_{i+1} = x_i - \alpha\tilde{v}_{x_i}$, form an obtruse triangle. As a consequence,

$$\begin{aligned}
\|x_{i+1}\|^2 &\geqslant \|x_i\|^2 + \alpha^2\|\tilde{v}_{x_i}\|^2 \\
&\geqslant \|x_i\|^2 + \alpha^2\frac{1}{(2^d 16\pi)^2}\|x_*\|^2,
\end{aligned}$$

where the last inequality follows from (22). Thus, the norm of the iterates x_i will increase until after $\left(\frac{2K_0 2^d 16\pi}{\alpha}\right)^2$ iterations, we have $x_{i+N} \notin \mathcal{B}(0, 2K_0\|x_*\|)$.

The proof of the lemma is concluded by showing that

$$x_i \notin \mathcal{B}(0, 2K_0\|x_*\|) \text{ implies } x_{i+1} \notin \mathcal{B}(0, K_0\|x_*\|) \tag{23}$$

As a consequence, in a polynomial number N of steps, for each iterate, we have that $x_i \notin \mathcal{B}(0, K_0\|x_*\|)$, for all $i \geqslant N$, as claimed.

We next prove the implication (23). Consider $x_i \notin \mathcal{B}(0, 2K_0\|x_*\|)$, and note that

$$\begin{aligned}
\|x_{i+1}\| = \|x_i - \alpha\tilde{v}_{x_i}\| &\geqslant \|x_i\| - \alpha\|\tilde{v}_{x_i}\| \\
&\geqslant \|x_i\| - \alpha\frac{dK}{2^d}\max(\|x_i\|, \|x_*\|) \\
&\geqslant \|x_i\| - \alpha\frac{dK}{2^d}\frac{\|x_i\|}{2K_0} \\
&\geqslant \|x_i\| - \frac{1}{2}\|x_i\|
\end{aligned}$$

where the second inequality follows from (13), the third inequality from $\|x_i\| \geqslant 2K_0\|x_*\|$, and finally the last inequality from our assumption on the stepsize α. This concludes the proof of (23).

Proof of (22): It remains to prove (22). We start with proving $\langle x, \tilde{v}_x\rangle < 0$. For brevity of notation, let $\Lambda_z = \prod_{i=d}^{1} W_{i,+,z}$. We have

$$\begin{aligned}
x^T\tilde{v}_x &= \left\langle \Lambda_x^T\Lambda_x x - \Lambda_x^T\Lambda_{x_*}x_* + \Lambda_x^T\eta, x\right\rangle \\
&\leqslant \frac{13}{12}2^{-d}\|x\|^2 - \frac{1}{4\pi}\frac{1}{2^d}\|x\|\|x_*\| + \|x\|\frac{\omega}{2^{d/2}} \\
&\leqslant \|x\|\left(\frac{13}{12}2^{-d}\|x\| + \frac{1/(8\pi)}{2^d}\|x_*\| - \frac{1}{4\pi}\frac{1}{2^d}\|x_*\|\right) \\
&\leqslant \|x\|\frac{1}{2^d}\left(2\|x\| - \frac{1}{8\pi}\|x_*\|\right).
\end{aligned}$$

The first inequality follows from (9) and (10), and the second inequality follows from our assumption on ω. Therefore, for any $x \in \mathcal{B}(0, \frac{1}{16\pi}\|x_*\|)$, $\langle x, \tilde{v}_x\rangle < 0$, as desired.

We next show that, for any $x \in \mathcal{B}(0, \frac{1}{16\pi}\|x_*\|)$

$$\begin{aligned}
\|\tilde{v}_x\| =& \|\Lambda_x^T\Lambda_x x - \Lambda_x^T\Lambda_{x_*}x_* + \Lambda_x^T\eta\| \geqslant \|\Lambda_x^T\Lambda_{x_*}x_*\| - \|\Lambda_x^T\Lambda_x x\| - \|\Lambda_x^T\eta\| \\
\geqslant& \frac{1}{4\pi}\frac{1}{2^d}\|x_*\| - \frac{13}{12}\frac{1}{2^d}\|x\| - \frac{w}{2^{d/2}} \\
\geqslant& \frac{1}{2^d}\left(\frac{1}{8\pi} - \frac{1}{16\pi}\right)\|x_*\|.
\end{aligned}$$

where the second inequality is from (9) and (10). This concludes the proof of (22).

Acknowledgements

PH is partially supported by NSF Grant DMS-1464525, and the authors would like to thank Tan Nguyen for helpful discussions.

References

[Aro+15] S. Arora, Y. Liang, and T. Ma. "Why are deep nets reversible: A simple theory, with implications for training". In: *arXiv:1511.05653* (2015).

[Bor+17] A. Bora, A. Jalal, E. Price, and A. G. Dimakis. "Compressed sensing using generative models". In: *arXiv:1703.03208* (2017).

[Bur+12] H. C. Burger, C. J. Schuler, and S. Harmeling. "Image denoising: Can plain neural networks compete with BM3D?" In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012, pp. 2392–2399.

[Cla17] C. Clason. "Nonsmooth Analysis and Optimization". In: *arXiv:1708.04180* (2017).

[Dab+07] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. "Image denoising by sparse 3-D transform-domain collaborative filtering". In: *IEEE Transactions on Image Processing* 16.8 (2007), pp. 2080–2095.

[Don95] D. L. Donoho. "De-noising by soft-thresholding". In: *IEEE Transactions on Information Theory* 41.3 (1995), pp. 613–627.

[EA06] M. Elad and M. Aharon. "Image denoising via sparse and redundantd representations over learned dictionaries". In: *IEEE Transactions on Image Processing* 15.12 (2006), pp. 3736–3745.

[Goo+14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A Courville, and Y. Bengio. "Generative adversarial nets". In: *Advances in Neural Information Processing Systems 27*. 2014, pp. 2672–2680.

[HV18] P. Hand and V. Voroninski. "Global guarantees for enforcing deep generative priors by empirical risk". In: *Conference on Learning Theory*. arXiv:1705.07576. 2018.

[HS06] G. E. Hinton and R. R. Salakhutdinov. "Reducing the dimensionality of data with neural networks". In: *Science* 313.5786 (2006), pp. 504–507.

[Hua+18] W. Huang, P. Hand, and V. Voroninski. "A provably convergent scheme for inverting random generative networks subject to compressive observations". In: (2018).

[Kar+17] T. Karras, T. Aila, S. Laine, and J. Lehtinen. "Progressive growing of GANs for improved quality, stability, and variation". In: *arXiv: 1710.10196* (2017).

[Lec+98] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[Luc+18] A. Lucas, M. Iliadis, R. Molina, and A. K. Katsaggelos. "Using deep neural networks for inverse problems in imaging: Beyond analytical methods". In: *IEEE Signal Processing Magazine* 35.1 (2018), pp. 20–36.

[Lug+13] G. Lugosi, P. Massart, and S. Boucheron. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford: Oxford University Press, 2013.

[Mar+17] M. Mardani, H. Monajemi, V. Papyan, S. Vasanawala, D. Donoho, and J. Pauly. "Recurrent generative adversarial networks for proximal learning and automated compressive image recovery". In: *arXiv:1711.10046* (2017).

[Sta+02] J.-L. Starck, E. J. Candes, and D. L. Donoho. "The curvelet transform for image denoising". In: *IEEE Transactions on Image Processing* 11.6 (2002), pp. 670–684.

[Uly+17] D. Ulyanov, A. Vedaldi, and V. Lempitsky. "Deep Image Prior". In: *arXiv:1711.10925* (2017).

[Zha+17] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising". In: *IEEE Transactions on Image Processing* 26.7 (2017), pp. 3142–3155.

A Appendix

A.1 Proof of Lemma 4

Let $\Lambda_x = \Pi_{i=d}^1 W_{i,+,x}$. We have that

$$\|\bar{q}_x\|^2 = \left\|\Lambda_x^t \eta\right\|^2 \leqslant \|\Lambda_x\|^2 \|P_{\Lambda_x}\eta\|^2,$$

where P_{Λ_x} is a projector onto the span of Λ_x. As a consequence, $\|P_{\Lambda_x}\eta\|^2$ is χ^2-distributed random variable with k-degrees of freedom scaled by σ/n. A standard tail bound (see [Lug+13, p. 43]) yields that, for any $\beta \geqslant k$,

$$\mathbb{P}\left[\|P_{\Lambda_x}\eta\|^2 \geqslant 4\beta\right] \leqslant 2e^{-\beta}.$$

Next, we note that by applying Lemmas 13-14 from [HV18, Proof of Lem. 15])[3], with probability one, that the number of different matrices Λ_x can be bounded as

$$|\{\Lambda_x | x \neq 0\}| = |\{\Pi_{i=d}^1 W_{i,+,x} | x \neq 0\}| \leqslant 10^{d^2}(n_1^d n_2^{d-1} \dots n_d)^k \leqslant (n_1^d n_2^{d-1} \dots n_d)^{2k},$$

where the second inequality holds for $\log(10) \leqslant k/4 \log(n_1)$. To see this, note that $(n_1^d n_2^{d-1} \dots n_d)^k \geqslant 10^{d^2}$ is implied by $k(d\log(n_1) + (d-1)\log(n_2) + \dots \log(n_d)) \geqslant kd^2/4 \log(n_1) \geqslant d^2 \log(10)$. Thus, by the union bound,

$$\mathbb{P}\left[\|P_{\Lambda_x}\eta\|^2 \leqslant 16k \log(n_1^d n_2^{d-1} \dots n_d), \text{ for all } x\right] \geqslant 1 - 2e^{-2k\log(n)},$$

where $n = n_d$. Recall from (10) that $\|\Lambda_x\| \leqslant \frac{13}{12}$. Combining this inequality with $\|\bar{q}_x\|^2 \leqslant \|\Lambda_x\|^2 \|P_{\Lambda_x}\eta\|^2$ concludes the proof.

A.2 Proof of Lemma 8

We now show that h_x is away from zero outside of a neighborhood of x_* and $-\rho_d x_*$. We prove Lemma 8 by establishing the following:

Lemma 11. *Suppose* $64d^6\sqrt{\beta} \leqslant 1$. *Define*

$$\rho_d := \sum_{i=0}^{d-1} \frac{\sin \breve{\theta}_i}{\pi} \left(\prod_{j=i+1}^{d-1} \frac{\pi - \breve{\theta}_j}{\pi} \right),$$

where $\breve{\theta}_0 = \pi$ *and* $\breve{\theta}_i = g(\breve{\theta}_{i-1})$. *If* $x \in \mathcal{S}_\beta$, *then we have that either*

$$|\bar{\theta}_0| \leqslant 32d^4\beta \quad \textit{and} \quad |\|x\|_2 - \|x_*\|_2| \leqslant 132d^6\beta\|x_*\|_2$$

or

$$|\bar{\theta}_0 - \pi| \leqslant 8\pi d^4\sqrt{\beta} \quad \textit{and} \quad |\|x\|_2 - \|x_*\|_2\rho_d| \leqslant 200d^7\sqrt{\epsilon}\|x_*\|_2.$$

In particular, we have

$$\mathcal{S}_\beta \subset \mathcal{B}(x_*, 5000d^6\beta\|x_*\|_2) \cup \mathcal{B}(-\rho_d x_*, 500d^{11}\sqrt{\beta}\|x_*\|_2). \tag{24}$$

Additionally, $\rho_d \to 1$ *as* $d \to \infty$.

[3] The proof in that argument only uses the assumption of independence of subsets of rows of the weight matrices.

Proof. Without loss of generality, let $\|x_*\| = 1$, $x_* = e_1$ and $\hat{x} = r\cos\overline{\theta}_0 \cdot e_1 + r\sin\overline{\theta}_0 \cdot e_2$ for $\overline{\theta}_0 \in [0,\pi]$. Let $x \in \mathcal{S}_\beta$.

First we introduce some notation for convenience. Let

$$\xi = \prod_{i=0}^{d-1} \frac{\pi - \overline{\theta}_i}{\pi}, \quad \zeta = \sum_{i=0}^{d-1} \frac{\sin\overline{\theta}_i}{\pi} \prod_{j=i+1}^{d-1} \frac{\pi - \overline{\theta}_j}{\pi}, \quad r = \|x\|_2, \quad M = \max(r,1).$$

Thus, $h_x = -\frac{1}{2^d}\xi\hat{x}_0 + \frac{1}{2^d}(r-\zeta)\hat{x}$. By inspecting the components of h_x, we have that $x \in \mathcal{S}_\beta$ implies

$$|-\xi + \cos\overline{\theta}_0(r-\zeta)| \leqslant \beta M \tag{25}$$
$$|\sin\overline{\theta}_0(r-\zeta)| \leqslant \beta M \tag{26}$$

Now, we record several properties. We have:

$$\overline{\theta}_i \in [0,\pi/2] \text{ for } i \geqslant 1$$
$$\overline{\theta}_i \leqslant \overline{\theta}_{i-1} \text{ for } i \geqslant 1$$
$$|\xi| \leqslant 1 \tag{27}$$
$$|\zeta| \leqslant \frac{d}{\pi}\sin\theta_0 \tag{28}$$
$$\breve{\theta}_i \leqslant \frac{3\pi}{i+3} \text{ for } i \geqslant 0 \tag{29}$$
$$\breve{\theta}_i \geqslant \frac{\pi}{i+1} \text{ for } i \geqslant 0 \tag{30}$$
$$\xi = \prod_{i=0}^{d-1} \frac{\pi - \overline{\theta}_i}{\pi} \geqslant \frac{\pi - \overline{\theta}_0}{\pi} d^{-3} \tag{31}$$
$$\overline{\theta}_0 = \pi + O_1(\delta) \Rightarrow \overline{\theta}_i = \breve{\theta}_i + O_1(i\delta) \tag{32}$$
$$\overline{\theta}_0 = \pi + O_1(\delta) \Rightarrow |\xi| \leqslant \frac{\delta}{\pi} \tag{33}$$
$$\overline{\theta}_0 = \pi + O_1(\delta) \Rightarrow \zeta = \rho_d + O_1(3d^3\delta) \text{ if } \frac{d^2\delta}{\pi} \leqslant 1 \tag{34}$$

We now establish (29). Observe $0 < g(\theta) \leqslant \left(\frac{1}{3\pi} + \frac{1}{\theta}\right)^{-1} =: \tilde{g}(\theta)$ for $\theta \in (0,\pi]$. As g and $\tilde{g}$ are monotonic increasing, we have $\breve{\theta}_i = g^{\circ i}(\breve{\theta}_0) = g^{\circ i}(\pi) \leqslant \tilde{g}^{\circ i}(\pi) = \left(\frac{i}{3\pi} + \frac{1}{\pi}\right)^{-1} = \frac{3\pi}{i+3}$. Similarly, $g(\theta) \geqslant (\frac{1}{\pi} + \frac{1}{\theta})^{-1}$ implies that $\breve{\theta}_i \geqslant \frac{\pi}{i+1}$, establishing (30).

We now establish (31). Using (29) and $\overline{\theta}_i \leqslant \breve{\theta}_i$, we have

$$\prod_{i=1}^{d-1}\left(1 - \frac{\overline{\theta}_i}{\pi}\right) \geqslant \prod_{i=1}^{d-1}\left(1 - \frac{3}{i+3}\right) \geqslant d^{-3},$$

where the last inequality can be established by showing that the ratio of consecutive terms with respect to d is greater for the product in the middle expression than for d^{-3}.

We establish (32) by using the fact that $|g'(\theta)| \leqslant 1$ for all $\theta \in [0,\pi]$ and using the same logic as for [HV18, Eq. 17].

We now establish (34). As $\overline{\theta}_0 = \pi + O_1(\delta)$, we have $\overline{\theta}_i = \breve{\theta}_i + O_1(i\delta)$. Thus, if $\frac{d^2\delta}{\pi} \leqslant 1$,

$$\prod_{j=i+1}^{d-1} \frac{\pi - \overline{\theta}_j}{\pi} = \prod_{j=i+1}^{d-1}\left(\frac{\pi - \breve{\theta}_j}{\pi} + O_1(\frac{i\delta}{2\pi})\right) = \left(\prod_{j=i+1}^{d-1} \frac{\pi - \breve{\theta}_j}{\pi}\right) + O_1(d^2\delta)$$

So

$$\zeta = \sum_{i=0}^{d-1} \Big(\frac{\sin \breve{\theta}_i}{\pi} + O_1(\frac{i\delta}{\pi}) \Big) \Big[\Big(\prod_{j=i+1}^{d-1} \frac{\pi - \breve{\theta}_j}{\pi} \Big) + O_1(d^2\delta) \Big] \tag{35}$$

$$= \rho_d + O_1\Big(d^2\delta/\pi + d^3\delta/\pi + d^4\delta^2/\pi \Big) \tag{36}$$

$$= \rho_d + O_1(3d^3\delta). \tag{37}$$

Thus (34) holds.

Next, we establish that $x \in \mathcal{S}_\beta \Rightarrow r \leqslant 4d$, and thus $M \leqslant 4d$. Suppose $r > 1$. At least one of the following holds: $|\sin\overline{\theta}_0| \geqslant 1/\sqrt{2}$ or $|\cos\overline{\theta}_0| \geqslant 1/\sqrt{2}$. If $|\sin\overline{\theta}_0| \geqslant 1/\sqrt{2}$ then (26) implies that $|r-\zeta| \leqslant \sqrt{2}\beta r$. Using (28), we get $r \leqslant \frac{d/\pi}{1-\sqrt{2}\beta} \leqslant d/2$ if $\beta < 1/4$. If $|\cos\overline{\theta}_0| \geqslant 1/\sqrt{2}$, then (25) implies that $|r-\zeta| \leqslant \sqrt{2}(\beta r + |\xi|)$. Using (27), (28), and $\beta < 1/4$, we get $r \leqslant \frac{\sqrt{2}|\xi|+\zeta}{1-\sqrt{2}\beta} \leqslant \frac{d+\sqrt{2}}{1-\sqrt{2}\beta} \leqslant 4d$. Thus, we have $x \in S_\beta \Rightarrow r \leqslant 4d \Rightarrow M \leqslant 4d$.

Next, we establish that we only need to consider the small angle case ($\overline{\theta}_0 \approx 0$) and the large angle case ($\overline{\theta}_0 \approx \pi$), by considering the following three cases:

(Case I) $\sin\overline{\theta}_0 \leqslant 16d^4\beta$: We have $\overline{\theta}_0 = O_1(32d^4\beta)$ or $\overline{\theta}_0 = \pi + O_1(32d^4\beta)$, as $32d^4\beta < 1$.

(Case II) $|r-\zeta| < \sqrt{\beta}M$: Applying case II to inequality (25) yields $|\xi| \leqslant 2\sqrt{\beta}M$. Using (31), we get $\overline{\theta}_0 = \pi + O_1(2\pi d^3\sqrt{\beta}M)$.

(Case III) $\sin\overline{\theta}_0 > 16d^4\beta$ and $|r-\zeta| \geqslant \sqrt{\beta}M$: Finally, consider Case III. By (26), we have $|r-\zeta| \leqslant \frac{\beta M}{\sin\overline{\theta}_0}$. Using this inequality in (25), we have $|\xi| \leqslant \beta M + \frac{\beta M}{\sin\overline{\theta}_0} \leqslant \frac{2\beta M}{\sin\overline{\theta}_0} \leqslant \frac{1}{8}d^{-4}M \leqslant \frac{1}{2}d^{-3}$, where the second to last inequality uses $\sin\overline{\theta}_0 > 16d^4\beta$ and the last inequality uses $M \leqslant 4d$. By (31), we have $\frac{\pi-\overline{\theta}_0}{\pi}d^{-3} \leqslant \xi \leqslant \frac{1}{2}d^{-3}$, which implies that $\overline{\theta}_0 \geqslant \pi/2$. Now, as $|r-\zeta| \geqslant \sqrt{\beta}M$, then by (26), we have $|\sin\overline{\theta}_0| \leqslant \sqrt{\beta}$. Hence, $\overline{\theta}_0 = \pi + O_1(2\sqrt{\beta})$, as $\overline{\theta}_0 \geqslant \pi/2$ and as $\beta < 1$.

At least one of the Cases I,II, or III hold. Thus, we see that it suffices to consider the small angle case $\overline{\theta}_0 = O_1(32d^4\beta)$ or the large angle case $\overline{\theta}_0 = \pi + O_1(8\pi d^4\sqrt{\beta})$.

Small Angle Case. Assume $\overline{\theta}_0 = O_1(\delta)$ with $\delta = 32d^4\beta$. As $\overline{\theta}_i \leqslant \overline{\theta}_0 \leqslant \delta$ for all i, we have $1 \geqslant \xi \geqslant (1-\frac{\delta}{\pi})^d = 1 + O_1(\frac{2\delta d}{\pi})$ provided $\delta d/\pi \leqslant 1/2$ (which holds by our choice $\delta = 32d^4\beta$ by assumption $64d^6\sqrt{\beta} \leqslant 1$). By (28), we also have $\zeta = O_1(\frac{d}{\pi}\delta)$. By (25), we have

$$|-\xi + \cos\overline{\theta}_0(r-\zeta)| \leqslant \beta M.$$

Thus, as $\cos\overline{\theta}_0 = 1 + O_1(\overline{\theta}_0^2/2) = 1 + O_1(\delta^2/2)$,

$$-\Big(1 + O_1(\frac{2\delta d}{\pi})\Big) + (1 + O_1(\frac{2\delta d}{\pi}))(r + O_1(\frac{\delta d}{\pi})) = O_1(4d\beta),$$

and $r \leqslant M \leqslant 4d$ (shown above) provides,

$$r - 1 = O_1(4d\beta + \frac{2\delta d}{\pi} + \frac{\delta d}{\pi} + \frac{2\delta d}{\pi}4d + \frac{2\delta^2 d^2}{\pi^2}) \tag{38}$$

$$= O_1(4\beta d + 4\delta d^2). \tag{39}$$

By plugging in that $\delta = 32d^4\beta$, we have that $r - 1 = O_1(132d^6\beta)$, where we have used that $\frac{32d^5\beta}{\pi} \leqslant 1/2$.

Large Angle Case. Assume $\theta_0 = \pi + O_1(\delta)$ where $\delta = 8\pi d^4\sqrt{\beta}$. By (33) and (34), we have $\xi = O_1(\delta/\pi)$, and we have $\zeta = \rho_d + O_1(3d^3\delta)$ if $8d^6\sqrt{\beta} \leqslant 1$. By (25), we have

$$|-\xi + \cos\theta_0(r-\zeta)| \leqslant \beta M,$$

so, as $\cos\theta_0 = 1 - O_1(\theta_0^2/2)$,

$$O_1(\delta/\pi) + (1 + O_1(\delta^2/2))(r - \rho_d + O_1(3d^3\delta)) = O_1(\beta M),$$

and thus, using $r \leqslant 4d$, $\rho_d \leqslant d$, and $\delta = 8\pi d^4\sqrt{\beta} \leqslant 1$,

$$\begin{aligned} r - \rho_d &= O_1(\beta M + \delta/\pi + 3d^3\delta + \frac{5}{2}\delta^2 d + \frac{3}{2}d^3\delta^3) && (40)\\ &= O_1\Big(4\beta d + \delta(\frac{1}{\pi} + 3d^3 + \frac{5}{2}d + \frac{3}{2}d^3)\Big) && (41)\\ &= O_1(200d^7\sqrt{\beta}) && (42)\end{aligned}$$

To conclude the proof of (24), we use the fact that

$$\|x - x_*\|_2 \leqslant \big|\|x\|_2 - \|x_*\|_2\big| + (\|x_*\|_2 + \big|\|x\|_2 - \|x_*\|_2\big|)\overline{\theta}_0.$$

This fact simply says that if a 2d point is known to have magnitude within Δr of some r and is known to be within angle $\Delta\theta$ from 0, then its Euclidean distance to the point of polar coordinates $(r, 0)$ is no more than $\Delta r + (r + \Delta r)\Delta\theta$.

Finally, we establish that $\rho_d \to 1$ as $d \to \infty$. Note that $\rho_{d+1} = (1 - \frac{\breve{\theta}_d}{\pi})\rho_d + \frac{\sin\breve{\theta}_d}{\pi}$ and $\rho_0 = 0$. It suffices to show $\tilde{\rho}_d \to 0$, where $\tilde{\rho}_d := 1 - \rho_d$. The following recurrence relation holds: $\tilde{\rho}_d = (1 - \frac{\breve{\theta}_{d-1}}{\pi})\tilde{\rho}_{d-1} + \frac{\breve{\theta}_{d-1} - \sin\breve{\theta}_{d-1}}{\pi}$, with $\tilde{\rho}_0 = 1$. Using the recurrence formula [HV18, Eq. (15)] and the fact that $\breve{\theta}_0 = \pi$, we get that

$$\tilde{\rho}_d = \sum_{i=1}^{d} \frac{\breve{\theta}_{i-1} - \sin\breve{\theta}_{i-1}}{\pi} \prod_{j=i+1}^{d} (1 - \frac{\breve{\theta}_{j-1}}{\pi}) \tag{43}$$

using (30), we have that

$$\prod_{j=i+1}^{d}\Big(1 - \frac{\breve{\theta}_{j-1}}{\pi}\Big) \leqslant \prod_{j=i+1}^{d}\Big(1 - \frac{1}{j}\Big) = \exp\Big(-\sum_{j=i+1}^{d}\frac{1}{j}\Big) \leqslant \exp\Big(-\int_{i+1}^{d+1}\frac{1}{s}ds\Big) = \frac{i+1}{d+1}$$

Using (29) and the fact that $\breve{\theta}_{i-1} - \sin\breve{\theta}_{i-1} \leqslant \breve{\theta}_{i-1}^3/6$, we have that $\tilde{\rho}_d \leqslant \sum_{i=1}^{d}\frac{\breve{\theta}_{i-1}^3}{6\pi}\cdot\frac{i+1}{d+1} \to 0$ as $d \to \infty$.

□

A.3 Proof of Lemma 9

Consider the function

$$f_\eta(x) = f_0(x) - \langle G(x) - G(x_*), \eta\rangle,$$

and note that $f(x) = f_\eta(x) - \|\eta\|^2$. Consider $x \in \mathcal{B}(\phi_d x_*, \varphi\|x_*\|)$, for a φ that will be specified later. Note that

$$\begin{aligned} |\langle G(x) - G(x_*), \eta\rangle| &\leqslant |\left\langle \Pi_{i=d}^1 W_{i,+,x}x, \eta\right\rangle| + |\left\langle \Pi_{i=d}^1 W_{i,+,x_*}x_*, \eta\right\rangle|\\ &= |\left\langle x, (\Pi_{i=d}^1 W_{i,+,x})^t\eta\right\rangle| + |\left\langle x_*, (\Pi_{i=d}^1 W_{i,+,x_*})^t\eta\right\rangle|\\ &\leqslant (\|x\| + \|x_*\|)\frac{\omega}{2^{d/2}}\\ &\leqslant (\varphi + \|x_*\|)\frac{\omega}{2^{d/2}},\end{aligned}$$

where the second inequality holds on the event $\mathcal{E}_{\text{noise}}$, by Lemma 4, and the last inequality holds by our assumption on x. Thus, for $x \in \mathcal{B}(\phi_d x_*, \varphi\|x_*\|)$

$$\begin{aligned}
f_\eta(x) \leqslant & \mathbb{E}f_0(x) + |f_0(x) - \mathbb{E}f_0(x)| + |\langle G(x) - G(x_*), \eta\rangle| \\
\leqslant & \frac{1}{2^{d+1}}\left(\phi_d^2 - 2\phi_d + \frac{10}{K_2^3}d\varphi\right)\|x_*\|^2 + \frac{1}{2^{d+1}}\|x_*\|^2 \\
& + \frac{\epsilon(1+4\epsilon d)}{2^d}\|x\|^2 + \frac{\epsilon(1+4\epsilon d) + 48d^3\sqrt{\epsilon}}{2^{d+1}}\|x\|\|x_*\| + \frac{\epsilon(1+4\epsilon d)}{2^d}\|x_*\|^2 \\
& + (\varphi + \|x_*\|)\frac{\omega}{2^{d/2}} \\
\leqslant & \frac{1}{2^{d+1}}\left(\phi_d^2 - 2\phi_d + \frac{10}{K_2^3}d\varphi\right)\|x_*\|^2 + \frac{1}{2^{d+1}}\|x_*\|^2 \\
& + \frac{\epsilon(1+4\epsilon d)}{2^d}(\phi_d + \varphi)^2\|x_*\|^2 + \frac{\epsilon(1+4\epsilon d) + 48d^3\sqrt{\epsilon}}{2^{d+1}}(\phi_d + \varphi)\|x_*\|^2 + \frac{\epsilon(1+4\epsilon d)}{2^d}\|x_*\|^2 \\
& + (\varphi + \|x_*\|)\frac{\omega}{2^{d/2}} \\
\leqslant & \frac{\|x_*\|^2}{2^{d+1}}\left(1 + \phi_d^2 - 2\phi_d + \frac{10}{K_2^3}d\epsilon + 68d^2\sqrt{\epsilon}\right) + (\varphi + \|x_*\|)\frac{\omega}{2^{d/2}} \qquad (44)
\end{aligned}$$

where the last inequality follows from $\epsilon < \sqrt{\epsilon}$, $\rho_d \leqslant 1$, $4\epsilon d < 1$, $\varphi < 1$ and assuming $\varphi = \epsilon$.

Similarly, we have that for any $y \in \mathcal{B}(-\phi_d x_*, \varphi\|x_*\|)$

$$\begin{aligned}
f_\eta(y) \geqslant & \mathbb{E}[f(y)] - |f(y) - \mathbb{E}[f(y)]| - |\langle G(x) - G(x_*), \eta\rangle| \\
\geqslant & \frac{1}{2^{d+1}}\left(\phi_d^2 - 2\phi_d\rho_d - 10d^3\varphi\right)\|x_*\|^2 + \frac{1}{2^{d+1}}\|x_*\|^2 \\
& - \left(\frac{\epsilon(1+4\epsilon d)}{2^d}\|y\|^2 + \frac{\epsilon(1+4\epsilon d) + 48d^3\sqrt{\epsilon}}{2^{d+1}}\|y\|\|x_*\| + \frac{\epsilon(1+4\epsilon d)}{2^d}\|x_*\|^2\right) \\
& - (\varphi + \|x_*\|)\frac{\omega}{2^{d/2}} \\
\geqslant & \frac{\|x_*\|^2}{2^{d+1}}\left(1 + \phi_d^2 - 2\phi_d\rho_d - 10d^3\varphi - 68d^2\sqrt{\epsilon}\right) - (\varphi + \|x_*\|)\frac{\omega}{2^{d/2}} \qquad (45)
\end{aligned}$$

Using $\epsilon < \sqrt{\epsilon}$, $\rho_d \leqslant 1$, $4\epsilon d < 1$, $\varphi < 1$ and assuming $\varphi = \epsilon$, the right side of (44) is smaller than the right side of (45) if

$$\varphi = \epsilon \leqslant \left(\frac{\phi_d - \rho_d\phi_d - 13\|\overline{\eta}\|_2}{\left(125 + \frac{5}{K_2^3}\right)d^3}\right)^2. \qquad (46)$$

We can establish that:

Lemma 12. *For all $d \geqslant 2$, that*

$$1/\left(K_1(d+2)^2\right) \leqslant 1 - \rho_d \leqslant 250/(d+1).$$

Thus, it suffices to have $\varphi = \epsilon = \frac{K_3}{d^{10}}$ and $13\|\overline{\eta}\|_2 \leqslant \frac{K_9}{d^2} \leqslant \frac{1}{2}\frac{K_2}{K_1(d+2)^2}$ for an appropriate universal constant K_9, and for an appropriate universal constant K_3.

A.4 Proof of Lemma 12

It holds that

$$\|x-y\| \geqslant 2\sin(\theta_{x,y}/2)\min(\|x\|,\|y\|), \qquad \forall x,y \tag{47}$$
$$\sin(\theta/2) \geqslant \theta/4, \qquad \forall \theta \in [0,\pi] \tag{48}$$
$$\frac{d}{d\theta}g(\theta) \in [0,1] \qquad \forall \theta \in [0,\pi] \tag{49}$$
$$\log(1+x) \leqslant x \qquad \forall x \in [-0.5,1] \tag{50}$$
$$\log(1-x) \geqslant -2x \qquad \forall x \in [0,0.75] \tag{51}$$

where $\theta_{x,y} = \angle(x,y)$. We recall the results (36), (37), and (50) in [HV18]:

$$\check{\theta}_i \leqslant \frac{3\pi}{i+3} \qquad \text{and} \qquad \check{\theta}_i \geqslant \frac{\pi}{i+1} \qquad \forall i \geqslant 0$$
$$1-\rho_d = \prod_{i=1}^{d-1}\left(1-\frac{\check{\theta}_i}{\pi}\right) + \sum_{i=1}^{d-1}\frac{\check{\theta}_i - \sin\check{\theta}_i}{\pi}\prod_{j=i+1}^{d-1}\left(1-\frac{\check{\theta}_j}{\pi}\right).$$

Therefore, we have for all $0 \leqslant i \leqslant d-2$,

$$\prod_{j=i+1}^{d-1}\left(1-\frac{\check{\theta}_j}{\pi}\right) \leqslant \prod_{j=i+1}^{d-1}\left(1-\frac{1}{j+1}\right) = e^{\sum_{j=i+1}^{d-1}\log\left(1-\frac{1}{j+1}\right)} \leqslant e^{-\sum_{j=i+1}^{d-1}\frac{1}{j+1}} \leqslant e^{-\int_{i+1}^{d}\frac{1}{s+1}ds} = \frac{i+2}{d+1},$$
$$\prod_{j=i+1}^{d-1}\left(1-\frac{\check{\theta}_j}{\pi}\right) \geqslant \prod_{j=i+1}^{d-1}\left(1-\frac{3}{j+3}\right) = e^{\sum_{j=i+1}^{d-1}\log\left(1-\frac{3}{j+3}\right)} \geqslant e^{-\sum_{j=i+1}^{d-1}\frac{6}{j+3}} \geqslant e^{-\int_{i}^{d-1}\frac{6}{s+3}ds} = \left(\frac{i+3}{d+2}\right)^6,$$

where the second and the fifth inequalities follow from (50) and (51) respectively. Since $\pi^3/(12(i+1)^3) \leqslant \check{\theta}_i^3/12 \leqslant \check{\theta}_i - \sin\check{\theta}_i \leqslant \check{\theta}_i^3/6 \leqslant 27\pi^3/(6(i+3)^3)$, we have that for all $d \geqslant 3$

$$1-\rho_d \leqslant \frac{2}{d+1} + \sum_{i=1}^{d-1}\frac{27\pi^3}{6(i+3)^3}\frac{i+2}{d+1} \leqslant \frac{2}{d+1} + \frac{3\pi^5}{4(d+1)} \leqslant \frac{250}{d+1}$$

and

$$1-\rho_d \geqslant \left(\frac{3}{(d+2)}\right)^6 + \sum_{i=1}^{d-1}\frac{\pi^3}{12(i+3)^3}\left(\frac{i+3}{d+2}\right)^6 \geqslant \frac{1}{K_1(d+2)^2},$$

where we use $\sum_{i=4}^{\infty}\frac{1}{i^2} \leqslant \frac{\pi^2}{6}$ and $\sum_{i=1}^{n} i^3 = O(n^4)$.

A.5 Proof of Lemma 10

To establish Lemma 10, we prove the following:

Lemma 13. *For all $x, y \neq 0$,*

$$\|h_x - h_y\| \leqslant \left(\frac{1}{2^d} + \frac{6d+4d^2}{\pi 2^d}\max\left(\frac{1}{\|x\|},\frac{1}{\|y\|}\right)\|x_*\|\right)\|x-y\|$$

Lemma 10 follows by noting that if $x, y \notin \mathcal{B}(0, r\|x_*\|)$, then $\|h_x - h_y\| \leqslant \left(\frac{1}{2^d} + \frac{6d+4d^2}{\pi r 2^d}\right)\|x-y\|$.

Proof of Lemma 13. For brevity of notation, let $\zeta_{j,z} = \prod_{i=j}^{d-1} \frac{\pi - \bar{\theta}_{i,z}}{\pi}$. Combining (47) and (48) gives $|\bar{\theta}_{0,x} - \bar{\theta}_{0,y}| \leqslant 4 \max\left(\frac{1}{\|x\|}, \frac{1}{\|y\|}\right) \|x - y\|$. Inequality (49) implies $|\bar{\theta}_{i,x} - \bar{\theta}_{i,y}| \leqslant |\bar{\theta}_{j,x} - \bar{\theta}_{j,y}|$ for all $i \geqslant j$. It follows that

$$\begin{aligned}\|h_x - h_y\| \leqslant & \frac{1}{2^d}\|x - y\| + \frac{1}{2^d} \underbrace{|\zeta_{0,x} - \zeta_{0,y}|}_{T_1} \|x_*\| \\ & + \frac{1}{2^d} \underbrace{\left| \sum_{i=0}^{d-1} \frac{\sin \bar{\theta}_{i,x}}{\pi} \zeta_{i+1,x} \hat{x} - \sum_{i=0}^{d-1} \frac{\sin \bar{\theta}_{i,y}}{\pi} \zeta_{i+1,y} \hat{y} \right|}_{T_2} \|x_*\|. \end{aligned} \tag{52}$$

By Lemma 14, we have

$$T_1 \leqslant \frac{d}{\pi} |\bar{\theta}_{0,x} - \bar{\theta}_{0,y}| \leqslant \frac{4d}{\pi} \max\left(\frac{1}{\|x\|}, \frac{1}{\|y\|}\right) \|x - y\|. \tag{53}$$

Additionally, it holds that

$$\begin{aligned} T_2 = & \left| \sum_{i=0}^{d-1} \frac{\sin \bar{\theta}_{i,x}}{\pi} \zeta_{i+1,x} \hat{x} - \frac{\sin \bar{\theta}_{i,x}}{\pi} \zeta_{i+1,x} \hat{y} + \frac{\sin \bar{\theta}_{i,x}}{\pi} \zeta_{i+1,x} \hat{y} - \sum_{i=0}^{d-1} \frac{\sin \bar{\theta}_{i,y}}{\pi} \zeta_{i+1,y} \hat{y} \right| \\ \leqslant & \frac{d}{\pi} \|\hat{x} - \hat{y}\| + \underbrace{\left| \sum_{i=0}^{d-1} \frac{\sin \bar{\theta}_{i,x}}{\pi} \zeta_{i+1,x} - \sum_{i=0}^{d-1} \frac{\sin \bar{\theta}_{i,y}}{\pi} \zeta_{i+1,y} \right|}_{T_3}. \end{aligned} \tag{54}$$

We have

$$\begin{aligned} T_3 \leqslant & \sum_{i=0}^{d-1} \left[\left| \frac{\sin \bar{\theta}_{i,x}}{\pi} \zeta_{i+1,x} - \frac{\sin \bar{\theta}_{i,x}}{\pi} \zeta_{i+1,y} \right| + \left| \frac{\sin \bar{\theta}_{i,x}}{\pi} \zeta_{i+1,y} - \frac{\sin \bar{\theta}_{i,y}}{\pi} \zeta_{i+1,y} \right| \right] \\ \leqslant & \sum_{i=0}^{d-1} \left[\frac{1}{\pi} \left(\frac{d - i - 1}{\pi} |\bar{\theta}_{i-1,x} - \bar{\theta}_{i-1,y}| \right) + \frac{1}{\pi} |\sin \bar{\theta}_{i,x} - \sin \bar{\theta}_{i,y}| \right] \\ \leqslant & \frac{d^2}{\pi} |\bar{\theta}_{0,x} - \bar{\theta}_{0,y}| \leqslant \frac{4d^2}{\pi} \max\left(\frac{1}{\|x\|}, \frac{1}{\|y\|}\right) \|x - y\|. \end{aligned} \tag{55}$$

Using (47) and (48) and noting $\|\hat{x} - \hat{y}\| \leqslant \theta_{x,y}$ yield

$$\|\hat{x} - \hat{y}\| \leqslant \theta_{x,y} \leqslant 2 \max\left(\frac{1}{\|x\|}, \frac{1}{\|y\|}\right) \|x - y\|. \tag{56}$$

Finally, combining (52), (53), (54), (55) and (56) yields the result. □

Lemma 14. *Suppose $a_i, b_i \in [0, \pi]$ for $i = 1, \ldots, k$, and $|a_i - b_i| \leqslant |a_j - b_j|, \forall i \geqslant j$. Then it holds that*

$$\left| \prod_{i=1}^{k} \frac{\pi - a_i}{\pi} - \prod_{i=1}^{k} \frac{\pi - b_i}{\pi} \right| \leqslant \frac{k}{\pi} |a_1 - b_1|.$$

Proof. Prove by induction. It is easy to verify that the inequality holds if $k = 1$. Suppose the inequality holds with $k = t - 1$. Then

$$\begin{aligned}\left|\prod_{i=1}^{t}\frac{\pi-a_i}{\pi}-\prod_{i=1}^{t}\frac{\pi-b_i}{\pi}\right| \leqslant & \left|\prod_{i=1}^{t}\frac{\pi-a_i}{\pi}-\frac{\pi-a_t}{\pi}\prod_{i=1}^{t-1}\frac{\pi-b_i}{\pi}\right| \\ & + \left|\frac{\pi-a_t}{\pi}\prod_{i=1}^{t-1}\frac{\pi-b_i}{\pi}-\prod_{i=1}^{t}\frac{\pi-b_i}{\pi}\right| \\ \leqslant & \frac{t-1}{\pi}|a_1-b_1|+\frac{1}{\pi}|a_t-b_t| \leqslant \frac{t}{\pi}|a_1-b_1|.\end{aligned}$$

□