Information and Inference: A Journal of the IMA (2020) **00**, 1–35
doi:10.1093/imaiai/iaaa011

Rate-optimal denoising with deep neural networks

Reinhard Heckel[†]
Department of Electrical and Computer Engineering, Rice University, Houston, Texas
[†]Corresponding author. Email: rh43@rice.edu

Wen Huang
School of Mathematical Sciences, Xiamen University, Xiamen, China

Paul Hand
Department of Mathematics and Khoury College of Computer Science, Northeastern University, Boston, Massachusetts

and

Vladislav Voroninski
Helm.ai, Menlo Park, California USA

[Received on 6 October 2019; revised on 31 March 2020; accepted on 2 April 2020]

Deep neural networks provide state-of-the-art performance for image denoising, where the goal is to recover a near noise-free image from a noisy observation. The underlying principle is that neural networks trained on large data sets have empirically been shown to be able to generate natural images well from a low-dimensional latent representation of the image. Given such a generator network, a noisy image can be denoised by (i) finding the closest image in the range of the generator or by (ii) passing it through an encoder-generator architecture (known as an autoencoder). However, there is little theory to justify this success, let alone to predict the denoising performance as a function of the network parameters. In this paper, we consider the problem of denoising an image from additive Gaussian noise using the two generator-based approaches. In both cases, we assume the image is well described by a deep neural network with ReLU activations functions, mapping a k-dimensional code to an n-dimensional image. In the case of the autoencoder, we show that the feedforward network reduces noise energy by a factor of $O(k/n)$. In the case of optimizing over the range of a generative model, we state and analyze a simple gradient algorithm that minimizes a non-convex loss function and provably reduces noise energy by a factor of $O(k/n)$. We also demonstrate in numerical experiments that this denoising performance is, indeed, achieved by generative priors learned from data.

Keywords: deep neural networks; denoising.

1. Introduction

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

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

Here, η is unknown noise, which we model as a zero-mean white Gaussian random variable with covariance matrix σ^2/nI. Image denoising relies on generative or prior assumptions on the image y_*,

such as self-similarity within images [5], sparsity in fixed [6] and learned bases [7] and most recently, by assuming the image can be generated by a pre-trained deep neural network [3;23]. Deep network-based approaches typically yield the best denoising performance. This success can be attributed to their ability to efficiently represent and learn realistic image priors, for example via auto-decoders [11] and generative adversarial models [8].

Motivated by this success story, we assume that the image y_* lies in the range of an image-generating network. In this paper, we propose the first algorithm for solving denoising with deep generative priors that provably finds an approximation of the underlying image. As the influence of deep networks in denoising and inverse problems grows, it becomes increasingly important to understand their performance at a theoretical level. Given that most optimization approaches for deep learning are first-order gradient methods, a justification is needed for why they do not get stuck in local minima.

The most related work that establishes theoretical reasons for why gradient methods might succeed when using deep generative priors for solving inverse problems is [9]. In it, the authors establish global favorability for optimization of a ℓ_2-loss function under a random neural network model. Specifically, they show existence of a descent direction outside a ball around the global optimizer and a negative multiple of it in the latent space of the generative model. This work does not justify why the one spurious point is avoided by gradient descent nor does it provide a specific algorithm which provably estimates the global minimizer nor does it provide an analysis of the robustness of the problem with respect to noise. This work was subsequently extended to include the case of generative convolutional neural networks by [18] but that work too does not prove convergence of a specific algorithm.

Contributions: The goal of this paper is to analytically quantify the denoising performance of deep prior-based denoisers. Specifically, we characterize the performance of two simple and efficient algorithms for denoising based on a d-layer generative neural network $G\colon \mathbb{R}^k \to \mathbb{R}^n$, with $k < n$.

We first provide a simple result for an encoder-generator network $G(E(y))$ where $E\colon \mathbb{R}^n \to \mathbb{R}^k$ is an encoder network. We show that if we pass noise through an encoder-decoder network $G(E(y))$ that acts as the identity on a class of images of interest, then it reduces the random noise by $O(k/n)$.

The second and main result of our paper pertains to denoising by optimizing over the latent code of a generator network with random weights. We propose a gradient method that attempts to minimize the least-squares loss $f(x) = \frac{1}{2}\|G(x) - y\|^2$ between the noisy image y and an image in the range of the generator, $G(x)$. Even though f is non-convex, we show that a gradient method yields an estimate $\hat{x}$ obeying

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

with high probability, where the notation $\lesssim$ absorbs a logarithmic factor dependent on the number of layers of the network and the network's expansivity, as discussed in more detail later. Our result shows that the denoising rate of a deep generator-based denoiser is optimal in terms of 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.

Related work: We hasten to add that a close theoretical work to the question considered in this paper is the paper [2], which solves a noisy compressive sensing problem with generative priors by minimizing an ℓ_2-loss. Under the assumption that the network is Lipschitz, they show that if the global optimizer can be found, which is in principle NP-hard, then a signal estimate is recovered to within the noise level.

While the Lipschitzness assumption is quite mild, the resulting theory does not provide justification for why global optimality can be reached. Another related work explored what properties of activation functions enables signal denoising with deep networks [20].

2. Background on denoising with classical and deep learning-based methods

As mentioned before, image denoising relies on modeling or prior assumptions on the image y_*. For example, suppose that the image y_* lies in a k-dimensional subspace of $\mathbb{R}^n$ denoted by $\mathcal{Y}$. Then we can estimate the original image by finding the closest point in ℓ_2-distance to the noisy observation y on the subspace $\mathcal{Y}$. The corresponding estimate, denoted by $\hat{y}$, obeys

$$\left\|\hat{y} - y_*\right\|^2 \lesssim \sigma^2 \frac{k}{n}, \tag{2.1}$$

with high probability (throughout, $\|\cdot\|$ denotes the ℓ_2-norm). Thus, the noise energy 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 (2.1] 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 image model (the subspace, in this example) does not include the original image y_*, then the error bound (2.1] increases, as we would remove a significant part of the signal along with noise when projecting onto the range of the image prior. Thus, a concise and accurate model is crucial for denoising.

Real-world signals rarely lie in *a priori* known subspaces, and the past few decades of image denoising research have developed sophisticated algorithms based on accurate image models. Examples include algorithms based on sparse representations in overcomplete dictionaries such as wavelets [6] and curvelets [21] and algorithms based on exploiting self-similarity within images [5]. A prominent example of the former class of algorithms is the (state-of-the-art) BM3D [5] algorithm. However, the nuances of real-world images are difficult to describe with handcrafted models. Thus, starting with the paper [7] that proposes to learn sparse representation based on training data, it has become common to learn concise representation for denoising (and other inverse problems) from a set of training images.

Burger *et al*. [3] applied deep networks to the denoising problem by training a plain neural network on a large set of images. Since then, deep learning-based denoisers [23] have set the standard for denoising. 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 [11] and generative adversarial models [8]. Over the past few years, the quality of deep priors has significantly improved [10;13;22]. 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.

3. Denoising with a neural network with an hourglass architecture

Perhaps the most straight-forward and classical approach to using deep networks for denoising is to train a deep network with an autoencoder or hourglass structure end-to-end to perform denoising. An autoencoder compresses data from the input layer into a low-dimensional code and then generates an output from that code. In this section, we analyze such networks from the perspective of denoising.

Specifically, we show mathematically that a simple model for neural networks with an hourglass structure achieves optimal denoising rates, as given by the dimensionality of the low-dimensional code.

An autoencoder $H(x) = G(E(x))$ consists of an encoder network $E\colon \mathbb{R}^n \to \mathbb{R}^k$ mapping an image to a low-dimensional latent representation, and a decoder or generator network $G\colon \mathbb{R}^k \to \mathbb{R}^n$ mapping the latent representation to an image. To see that the size of the low-dimensional code, k, determines the denoising rate, consider a simple one-layer encoder and multilayer decoder of the form

$$E(y) = \mathrm{relu}(W'y), \quad G(x) = \mathrm{relu}(W_d \dots \mathrm{relu}(W_2 \mathrm{relu}(W_1 x)) \dots), \tag{3.1}$$

where $\mathrm{relu}(x) = \max(x, 0)$ applies entrywise, W' are the weights of the encoder, $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ are the weights in the i-th layer of the decoder, and n_i is the number of neurons in the ith layer.

Typically, the autoencoder network H is trained such that $H(y) \approx y$ for some class of signals of interest (say, natural images). The following proposition shows that the structure of the network alone guarantees that an autoencoder 'filters out' most of the noise.

PROPOSITION 3.1 Let $H = G \circ E$ be an autoencoder of the form (3.1) and note that it is piecewise linear, i.e. $H(y) = Uy$ for some matrix U in a region around y. Suppose that $\|U\|^2 \leqslant 2$ for all such regions, where $\|U\|$ is the spectral norm, i.e. the largest singular value of U. Let η be Gaussian noise with covariance matrix σI, $\sigma > 0$. Then provided that $k \cdot 32 \log(2n_1 n_2 \dots n_d) \leqslant n$, we have with probability at least $1 - 2e^{-k \log(2n_1 n_2 \dots n_d)}$ that

$$\|H(\eta)\|_2^2 \leqslant 5 \frac{k}{n} \log(2n_1 n_2 \dots n_d) \|\eta\|_2^2.$$

Note that the assumption $\|U\|^2 \leqslant 2$ implies that $\|H(y)\|_2^2 / \|y\|_2^2 \leqslant 2$ for all y. This guarantees that the autoencoder does not 'amplify' a signal too much. The specific choice of the constant 2 is irrelevant and constant larger than one works and simply increases the bound by a constant factor. Note that we envision an autoencoder that is trained such that it obeys $H(y) \approx y$ for y in a class of images. The proposition would then justify why the feedforward network reduces noise by $O(k/n)$.

Note that the condition $k \cdot 32 \log(2n_1 n_2 \dots n_d) \leqslant n$ simply requires the autoencoder to have an hourglass structure. An autoencoder without hourglass structure does not denoise without additional assumptions on its weights or architecture.

The proof of Proposition 3.1, contained in the appendix, shows that H lies in the union of k-dimensional subspaces and then uses a standard concentration argument showing that the union of those subspaces can represent no more than a fraction of $O(k/n)$ of the noise.

In the remainder of the paper, we show that denoising via enforcing a generative prior gives us an analogous denoising rate.

4. Denoising via enforcing a generative model

We consider the problem of estimating a vector $y_* \in \mathbb{R}^n$ from a noisy observation $y = y_* + \eta$. We assume 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 $x_* \in \mathbb{R}^k$. We consider a generative network of the form

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

where $\mathrm{relu}(x) = \max(x, 0)$ applies entrywise, $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ are the weights in the i-th layer, n_i is the number of neurons in the ith layer, and the network is expansive in the sense that $k = n_0 < n_1 < \dots <$

$n_d = n$. The problem at hand is: given the weights of the network $W_1 \dots W_d$ and a noisy observation y, obtain an estimate $\hat{y}$ of the original image y_* such that $\|\hat{y} - y_*\|$ is small and $\hat{y}$ is in the range of G.

4.1 *Enforcing a generative model*

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 minimize the loss

$$f(x) := \frac{1}{2}\|G(x) - y\|^2. \tag{4.1}$$

Since 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—introduced next—finds a point that is very close to the original latent parameter x_*, with the distance to the parameter 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 weight matrix 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 informed by the loss surface of the function to be minimized. Given a noisy observation y, the algorithm starts with an arbitrary initial point $x_0 \neq 0$. At each iteration $i = 0, 1, \dots$, the algorithm computes the step direction

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

which is equal to the gradient of f if f is differentiable at x_i. It then takes a small step opposite to $\tilde{\nu}_{x_i}$. 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.

This tweak is informed by the loss surface. To understand this step, it is instructive to examine the loss surface for the *noiseless* case in Fig. 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 the orthogonal directions is zero. Further, around approximately $-x_*$, the loss function is larger than around the optimum x_*. As a simple gradient descent method (without the tweak) 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.

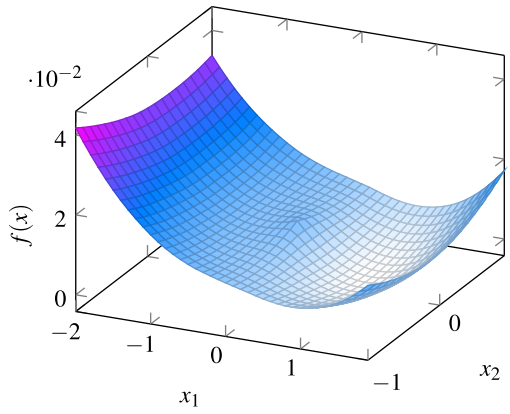

FIG. 1. Loss surface $f(x) = \|G(x) - G(x_*)\|$, $x_* = [1, 0]$, of an expansive network G with ReLU activation functions with $k = 2$ nodes in the input layer and $n_2 = 300$ and $n_3 = 784$ nodes in the hidden and output layers, respectively, with random Gaussian weights in each layer. The surface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

Other variations of the tweak are also possible. For example, the negation check in Step 3 could be

Algorithm 1 Gradient method for optimizing (4.1)

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

1: Choose an arbitrary initial point $x_0 \in \mathbb{R}^k \setminus \{0\}$
2: **for** $i = 0, 1, \ldots$ **do**
3: **if** $f(-x_i) < f(x_i)$ **then**
4: $\tilde{x}_i \leftarrow -x_i$;
5: **else**
6: $\tilde{x}_i \leftarrow x_i$;
7: **end if**
8: Compute $v_{\tilde{x}_i} \in \partial f(\tilde{x}_i)$, in particular, if G is differentiable at $\tilde{x}_i$, then set $v_{\tilde{x}_i} = \tilde{v}_{\tilde{x}_i}$, where

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

9: $x_{i+1} \leftarrow \tilde{x}_i - \alpha v_{\tilde{x}_i}$;
10: **end for**

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 a convergence criterion is again satisfied.

5. 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 that the weights W_i are independently and identically

distributed as $\mathcal{N}(0, 2/n_i)$ 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

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

where c is a particular numerical constant.

In a real-world generative network, the weights are learned from training data and are not drawn from a Gaussian distribution. Nonetheless, the motivation for selecting Gaussian weights for our analysis is as follows:

1. The empirical distribution of weights from deep neural networks often have statistics consistent with Gaussians. AlexNet is a concrete example [1].
2. 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.
3. It is not clear which non-Gaussian distribution for weights is superior from the joint perspective of realism and analytical tractability.

The network model we consider is fully connected. We anticipate that the analysis of this paper can be extended to the case of generative convolutional neural networks. This extension has already happened for theoretical results concerning the favorability of the optimization landscape for compressive sensing under generative priors [18], as mentioned previously.

We are now ready to state our main result.

THEOREM 5.1 Consider a network, G, with weights in the i-th layer given by $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$, i.i.d. $\mathcal{N}(0, 2/n_i)$ distributed, and suppose that the network satisfies the expansivity condition for $\epsilon = K/d^{90}$. Fix $x_* \in \mathbb{R}^k$. Let $y = G(x_*) + \eta$, where $\eta \sim \mathcal{N}(0, \frac{\sigma^2}{n} I_n)$. Also, suppose that the denoising bound ω, defined for notational convenience as

$$\omega := \sqrt{18\sigma^2 \frac{k}{n} \log\left(n_1^d n_2^{d-1} \dots n_d\right)},$$

obeys

$$\omega \leqslant \frac{\|x_*\| K_1}{d^{16}}.$$

Consider the iterates of Algorithm 1 with step size $\alpha = K_4 \frac{1}{d^2}$ for minimizing $f(x) = \frac{1}{2}\|G(x) - y\|^2$. Then there exists a number of steps N upper bounded by

$$N \leqslant K_2 d^{86} \frac{f(x_0)}{\|x_*\|}$$

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

$$\|x_i - x_*\| \leqslant \frac{K_5}{d^{36}} \|x_*\| + K_6 d^6 \omega, \quad \text{for all } i \geqslant N, \tag{5.1}$$

with probability at least $1-2e^{-2k\log n}-\sum_{i=2}^{d} 8n_i e^{-K_7 n_{i-2}}-8n_1 e^{-K_7 k\log d/d^{180}}$. In addition, for all $i > N$, we have

$$\|x_i - x_*\| \leqslant (1-\alpha 7/8)^{i-N}\|x_N - x_*\| + K_8\omega \text{ and} \tag{5.2}$$

$$\|G(x_i) - G(x_*)\| \leqslant 1.2(1-\alpha 7/8)^{i-N}\|x_N - x_*\| + 1.2K_8\omega, \tag{5.3}$$

where $\alpha < 1$ is the step size of the algorithm. Here, $K, K_1, K_2, \ldots$ are numerical constants, and x_0 is the initial point in the optimization.

Our result guarantees that after polynomially many iterations (with respect to d), the algorithm converges linearly to a region satisfying

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

where the notation $\lesssim$ absorbs a factor logarithmic in n and polynomial in d. To see this, note that for the number of iterations being sufficiently large, by Equation (5.2), $\left\|x_i - x_*\right\|^2 \leqslant K_8^2\omega^2 = K_8 18\sigma^2\frac{k}{n}\log(n_1^d n_2^{d-1}\ldots n_d)$. The authors made no attempt to optimize the dependence of this result on d and determining and proving the optimal scalings with respect to d are left as open problems.

The denoising rate can be observed in the $i \to \infty$ limit of (5.3), which holds because of (5.2) and Lipschitzness of G in a region around x_*.[1] Thus, the theorem guarantees that our algorithm yields the denoising rate of $\sigma^2 k/n$, 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, up to logarithmic and d-dependent factors.

In the case of $\sigma = 0$, the theorem guarantees convergence to the global minimizer x_*. We note that the intention of this paper is to show rate-optimality of recovery with respect to the noise power, the latent code dimensionality and the signal dimensionality. As a result, no attempt was made to establish optimal bounds with respect to the scaling of constants or to powers of d. The bounds provided in the theorem are highly conservative in the constants and dependency on the number of layers, d, in order to keep the proof as simple as possible. Numerical experiments shown later reveal that the parameter range for successful denoising are much broader than the constants suggest. As this result is the first of its kind for rigorous analysis of denoising performance by deep generative networks, we anticipate the results can be improved in future research, as has happened for other problems, such as sparsity-based compressed sensing and phase retrieval.

Finally, we remark that Theorem 5.1 can be generalized to the case where y_* only approximately lies in the range of the generator, i.e. $G(x_*) \approx y_*$. Specifically, if $\left\|G(x_*) - y_*\right\|$ is sufficiently small, the error induced by this perturbation is proportional to $\left\|G(x_*) - y_*\right\|$.

5.1 *The weight distribution condition*

While Theorem 5.1 is a statement about neural networks with Gaussian weights, we will prove it by establishing its conclusions for neural networks that satisfy a deterministic property known as the weight

[1] The proof of Lipschitzness follows from applying the WDC in Section 5.1.

distribution condition (WDC) [9]. The proof of the theorem will then follow because this property holds with high probability for the provided class of random neural networks.

We say that a matrix $W \in \mathbb{R}^{n\times k}$ satisfies the *WDC* with constant ϵ if for all nonzero $x, y \in \mathbb{R}^k$,

$$\left\| \sum_{i=1}^{n} 1_{<w_i x>0} 1_{<w_i y>0} \cdot w_i w_i^t - Q_{x,y} \right\| \leqslant \epsilon, \quad \text{with } Q_{x,y} = \frac{\pi - \theta_0}{2\pi} I_k + \frac{\sin\theta_0}{2\pi} M_{\hat{x}\leftrightarrow\hat{y}}, \tag{5.4}$$

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[2] 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 (5.4) is the spectral norm. Note that an elementary calculation[3] gives that $Q_{x,y} = \mathbb{E}[\sum_{i=1}^{n} 1_{<w_i x>0} 1_{<w_i y>0} \cdot w_i w_i^t]$ for $w_i \sim \mathcal{N}(0, I_k/n)$. Qualitatively, the WDC roughly means that the rows of W, once sampled, are approximately uniformly distributed over all directions.

Under the assumptions of Theorem 5.1, we note that $W_i/\sqrt{2}$ will satisfy the WDC with appropriate probability for all i, provided the expansivity condition holds.

5.2 *Sketch of proof of Theorem 5.1*

The proof relies on a characterization of the loss surface. We show that outside of two balls around $x = x_*$ and $x = -\rho_d x_*$, with ρ_d a constant defined in the proof, the direction chosen by the algorithm is a descent direction, with high probability.

We show that the step direction $\tilde{v}_x$ concentrates around a particular $h_x \in \mathbb{R}^k$ that is a continuous function of nonzero x, x_* and is zero only at $x = x_*$, $x = -\rho_d x_*$ and 0, using a concentration argument similar to [9]. Around $x = x_*$, the loss function has a global minimum, close to 0 it has a saddle point and close to $x = -\rho_d x_*$ potentially a local minimum. In a nutshell, we show that (i) the algorithm moves away from the saddle point at 0, and (ii) the algorithm escapes the local minimum close to $x = -\rho_d x_*$ with the twist in Steps 3–5 of the algorithm. Finally, the iterates end up close to the x_*.

The proof is organized as described next and as illustrated in Fig. 2. The algorithm is initialized at an arbitrary point, for example close to 0. 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 B.7 in the supplement). Then the algorithm might move toward $-\rho_d x_*$, 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 B.9 in the supplement. 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 (B.15) in the supplement). The region $\mathcal{S}_\beta^+$ is covered by a ball of radius r, by Lemma B.8 in the supplement, determined by the noise and ϵ. This establishes the bound (5.1) in the theorem.

---

[2] 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.

[3] 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.

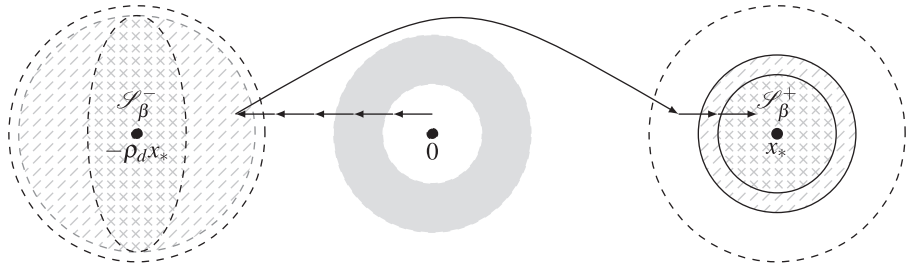

FIG. 2. Logic of the proof, explained in the text.

The proof proceeds be showing that within a ball around x_*, the algorithm then converges linearly, which establishes Equations (5.2) and (5.3).

6. Applications to compressed sensing

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 [2] have empirically been shown to improve over sparsity-based approaches, see [16] for a review for applications in imaging and [19] 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/mI, 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 $1/m$. One could prove an analogous result as Theorem 5.1, but with $\omega = \sqrt{18\sigma^2\frac{k}{m}\log(n_1^d n_2^{d-1}\dots n_d)}$, (note that n has been replaced by m). This extension shows that, provided ϵ is chosen sufficiently small, that our algorithm yields an iterate x_i obeying

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

where again $\lesssim$ absorbs factors logarithmic in the n_i's and polynomial in d. Proving this result would be analogous to the proof of Theorem 5.1, but with the additional assumption that the sensing matrix A acts like an isometry on the union of the ranges of $\Pi_{i=d}^1 W_{i,+,x_i}$, analogous to the proof in [9]. This extension of our result shows that Algorithm 1 enables solving inverse problems under noise efficiently and quantifies the effect of the noise.

We hasten to add that the paper [2] also derived an error bound for minimizing empirical loss. However, the corresponding result (for example Lemma 4.3) differs in two important aspects to our

result. First, the result in [2] only makes a statement about the *minimizer* of the empirical loss and does not provide justification that an *algorithm* can efficiently find a point near the global minimizer. As the program is non-convex, and as non-convex optimization is NP-hard in general, the empirical loss could have local minima at which algorithms get stuck. In contrast, the present paper presents a specific practical algorithm and proves that it finds a solution near the global optimizer regardless of initialization. Second, the result in [2] considers arbitrary noise η and thus cannot assert denoising performance. In contrast, we consider a random model for the noise and show the denoising behavior that the resulting error is no more than $O(k/n)$, as opposed to $\|\eta\|^2 \approx O(1)$, which is what we would get from direct application of the result in [2].

7. 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 provide denoising results for denoising with an hourglass architecture and via enforcing a generative prior. For enforcing a generative prior, 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.

7.1 *Denoising with enforcing 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 two-layer network with $n = 1500$ neurons in the output layer, 500 in the middle layer and vary the number of input neurons, k, and the noise level, σ. We next present simulations showing that if k is sufficiently small, our algorithm achieves a denoising rate proportional to $\sigma k/n$ as guaranteed by our theory.

Toward 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, we first obtain an estimate $\hat{x}$ of the latent representation by running Algorithm 1 until convergence, and second we obtain an estimate of the image as $\hat{y} = G(\hat{x})$. In the left and middle panel of Fig. 4, we depict the normalized mean squared error of the latent representation, $\mathrm{MSE}(\hat{x}, x_*)$ and the mean squared error in the image domain, $\mathrm{MSE}(G(\hat{x}), G(x_*))$, where we defined $\mathrm{MSE}(z, z') = \|z - z'\|^2$. For the left panel, we fix the noise variance to $\sigma^2 = 0.25$, and vary k, and for the middle panel we fix $k = 50$ and vary the noise variance. The results 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, we achieve an MSE proportional to $\sigma^2 k/n$, both in the representation and image domains.

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. Having said this, in general the negation step is necessary, since there exist problem instances that have a local minimum opposite of x_*.

7.2 *Denoising with enforcing a learned prior*

We next consider a prior learned from data. Technically, for such a prior our theory does not apply since we assume the weights to be chosen at random. However, the numerical results presented in this section show that even for the learned prior we achieve the rate predicted by our theory pertaining to a random prior. Toward this goal, we consider a fully connected autoencoder parameterized by k, consisting of an

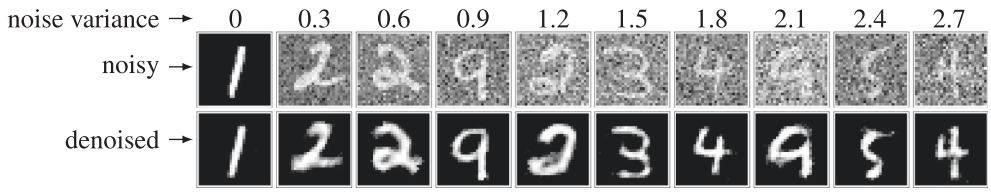

FIG. 3. Denosing with a learned generative prior: even when the number is barely visible, the denoiser recovers a sharp image.

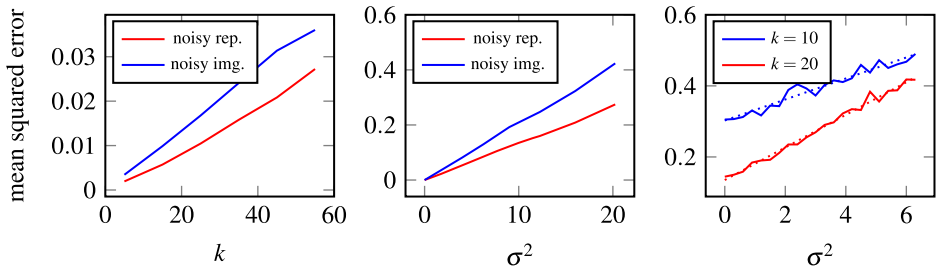

FIG. 4. MSE in the image domain, $\text{MSE}(G(\hat{x}), y_*)$ and in the latent representation, $\text{MSE}(\hat{x}, x_*)$, as a function of the dimension of the latent representation, k, with $\sigma^2 = 0.25$(left panel) and the noise variance, σ^2 with $k = 50$(middle panel). As suggested by the theory pertaining to decoders with random weights, 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 suggested by our theory.

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 $k = 20$ to obtain two different autoencoders. We train both autoencoders on the MNIST [**15**] *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 estimate the latent representation $\hat{x}$ by running Algorithm 1 and then set $\hat{y} = G(\hat{x})$. See Fig. 3 for a few examples demonstrating the performance of our approach for different noise levels.

We next show that this achieves a MSE proportional to $\sigma^2 k/n$, as suggested by our theory that applies for decoders with random weights. We add noise to the images with noise variance ranging from $\sigma^2 = 0$ to $\sigma^2 = 6$. In the right panel of Fig. 4, 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 bias-variance 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.

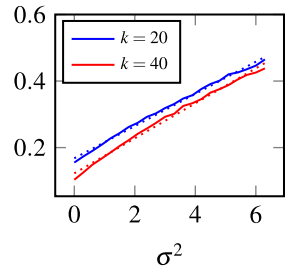

FIG. 5. Denoising of handwritten digits based on a learned hourglass (i.e. encoder-decoder) architecture with $k = 20$ and $k = 40$, along with the least-squares fit as dotted lines. The learned decoder with $k = 40$ represents the images with a smaller error; therefore, the MSE is slightly smaller for the noiseless case (i.e. $\sigma^2 = 0$) and for small noise variances. However, the denoising rate for the decoder with $k = 40$, which is the slope of the curve is larger as well (0.47 vs 0.52, illustrated by the least-squares fit as dotted lines), as suggested by our theory.

7.3 *Denoising with an autoencoder*

Finally, we consider a learned autoencoder for denoising. We consider the same setup as in the previous section; specifically, we train a fully connected autoencoder parameterized with five layers with $784, 400, k, 400, 787$ neurons in the five layers for $k \in \{20, 40\}$ and train both autoencoders on the MNIST training set. We then add Gaussian noise with variance $\sigma^2 \in [0, 6]$ and denoise the noisy image by passing it through the autoencoder. Figure 5 shows the MSE relative to the noiseless image, averaged over images from the MNIST test set. As suggested by our theory, the denoising rates are proportional to the noise variance σ^2, and the autoencoder with larger k has a larger denoising rate. Note that our theoretical denoising rates are worst case; therefore, an autoencoder with bottleneck k might show in simulations a smaller denoising rate than $k/n\sigma^2$, as is the case in the results from Fig. 5.

Funding

RH is partially supported by a NSF Grant ISS-1816986 and PH is partially supported by a NSF CAREER Grant DMS-1848087 as well as NSF Grant DMS-1464525.

Acknowledgements

The authors would like to thank Tan Nguyen and an anonymous reviewer for helpful discussions.

REFERENCES

1. ARORA, S., LIANG, Y. & MA, T. (2015) Why are deep nets reversible: as simple theory, with implications for training. arXiv:1511.05653.
2. BORA, A., JALAL, A., PRICE, E. & DIMAKIS, A. G. (2017) Compressed sensing using generative models. *Proceedings of the 34th International Conference on Machine Learning*, **70**. 537–546
3. BURGER, H. C., SCHULER, C. J. & HARMELING, S. (2012) Image denoising: can plain neural networks compete with BM3D? *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Rhode Island: Providence pp. 2392–2399.
4. CLASON, C. (2017) Nonsmooth analysis and optimization. Lecture Notes.. arXiv:1708.04180.
5. DABOV, K., FOI, A., KATKOVNIK, V. & EGIAZARIAN, K. (2007) Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Trans. Image Process.*, **16**, 2080–2095.

6. Donoho, D. L. (1995) De-noising by soft-thresholding. *IEEE Trans. Inf. Theory*, **41**, 613–627.

7. Elad, M. & Aharon, M. (2006) Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.*, **15**, 3736–3745.

8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014) Generative adversarial nets. *Advances in Neural Information Processing Systems* **27**, Montrèal, Canada, pp. 2672–2680.

9. Hand, P. & Voroninski, V. (2019) Global guarantees for enforcing deep generative priors by empirical risk. *IEEE Trans. Inf. Theory*, **66**, 401–418.

10. Heckel, R. & Hand, P. (2019) Deep decoder: concise image representations from untrained non-convolutional networks. *International Conference on Learning Representations*. New Orleans, Louisiana.

11. Hinton, G. E. & Salakhutdinov, R. R. (2006) Reducing the dimensionality of data with neural networks. *Science*, **313**, 504–507.

12. Jacques, L., Laska, J. N., Boufounos, P. T. & Baraniuk, R. G. (2013) Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Trans. Inf. Theory*, **59**, 2082–2102.

13. Karras, T., Aila, T., Laine, S. & Lehtinen, J. (2018) Progressive growing of GANs for improved quality, stability, and variation. *International Conference on Learning Representations*. Vancouver Canada: Vancouver Convention Center.

14. Laurent, B. & Massart, P. (2000) Adaptive estimation of a quadratic functional by model selection. *Ann. Stat.*, **28**, 1302–1338.

15. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998) Gradient-based learning applied to document recognition. *Proc. IEEE*, **86**, 2278–2324.

16. Lucas, A., Iliadis, M., Molina, R. & Katsaggelos, A. K. (2018) Using deep neural networks for inverse problems in imaging: beyond analytical methods. *IEEE Signal Process. Mag.*, **35**, 20–36.

17. Lugosi, G., Massart, P. & Boucheron, S. (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, England: Oxford University Press.

18. Ma, F., Ayaz, U. & Karaman, S. (2018) Invertibility of convolutional generative networks from partial measurements. *Advances in Neural Information Processing Systems*, Montreal, Canada: Montreal Convention Centre pp. 9651–9660.

19. Mardani, M., Monajemi, H., Papyan, V., Vasanawala, S., Donoho, D. & Pauly, J. (2018) Recurrent generative adversarial networks for proximal learning and automated compressive image recovery. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, Utah, United States.

20. Mixon, D. G. & Villar, S. (2018) SUNLayer: stable denoising with generative networks. arXiv:1803.09319 [cs, stat]. Salt Lake City, Utah, United States

21. Starck, J.-L., Candes, E. J. & Donoho, D. L. (2002) The curvelet transform for image denoising. *IEEE Trans. Image Process.*, **11**, 670–684.

22. Ulyanov, D., Vedaldi, A. & Lempitsky, V. (2018) Deep image prior. *IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, Utah, United States.

23. Zhang, K., Zuo, W., Chen, Y., Meng, D. & Zhang, L. (2017) Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.*, **26**, 3142–3155.

A. Proof of Proposition 3.1

We first show that $H(y)$ lies in the range of a union of k-dimensional subspaces and upper-bound the number of the subspaces. Toward this goal, first note that the effect of the ReLU operation $\text{relu}(z)$ can be described with a diagonal matrix D that contains a one on its diagonal if the respective entry of z is larger than zero, and zero otherwise, so that $Dz = \text{relu}(z)$. With this notation, we can write

$$H(y) = \underbrace{D_d W_d D_{d-1} \dots D_2 W_2 D_1 W_1 D' W'}_{U} y.$$

The matrix $U \in \mathbb{R}^{n\times n}$ has at most rank k, thus $H(y)$ lies in the range of a union of at most k-dimensional subspaces, where each subspace is determined by the matrices $D_d, \ldots, D_1, D'$. We next bound the number of subspaces. First note that since $W' \in \mathbb{R}^{n\times k}$, there are only 2^k many different choices for D', corresponding to all the sign patterns. Next note that, with the lemma below, we have that for a fixed $D'W'$, the number of different matrixes D_1 can be bounded by n_1^k. Likewise, for fixed $W_2D_1W_1D'W'$, the number of different matrices D_2 can be bounded by n_2^k and so forth. Thus, the total number of different choices of the matrices $D_d, \ldots, D_1, D'$ is upper bounded by

$$2^k n_1^k \ldots n_d^k.$$

LEMMA A.1 [Variant of Lemma 15 in [9]] For any $U \in \mathbb{R}^{n\times k}$ and $k \geqslant 5$,

$$|\{\mathrm{diag}(Uv > 0)U | v \in \mathbb{R}^k\}| \leqslant n^k.$$

We note that such lemmas are used in related signal recovery problems, such as 1-bit compressive sensing [12].

Next note that by assumption, we have that

$$\|U\eta\|_2^2/\|\eta\|_2^2 \leqslant 2, \tag{A.1}$$

for all vectors η and for all U defined by the matrices $D_d, \ldots, D_1, D'$. For fixed U, let S be the span of the right singular vectors of U, and note that S has dimension at most k. Let P_S be the orthogonal projector onto a subspace S. We have that

$$\frac{\|U\eta\|_2^2}{\|\eta\|_2^2} = \frac{\|UP_S\eta\|_2^2}{\|\eta\|_2^2} \leqslant 2\frac{\|P_S\eta\|_2^{2\,2}}{\|\eta\|_2^{}},$$

again for all η. Now, we make use of the following bound on the projection of the noise η onto a subspace, which follows from standard Gaussian concentration inequalities [14, Lemma 1].

LEMMA A.2 Let $S \subset \mathbb{R}^n$ be a subspace with dimension k. Let $\eta \sim \mathcal{N}(0, I_n)$ and $\beta \geqslant 1$. Then

$$\mathbb{P}\frac{\|P_S\eta\|_2^{2\,2}}{\|\eta\|_2^{}} \leqslant \frac{10\beta k}{n} \geqslant 1 - e^{-\beta k} - e^{-n/16}.$$

Taking a union bound over all subspaces, we obtain with the lemma above that

$$\mathbb{P}\frac{\|H(\eta)\|_2^{2\,2}}{\|\eta\|_2^{}} \leqslant \frac{20\beta k}{n} x \geqslant 1 - (2^k n_1^k \ldots n_d^k)(e^{-\beta k} + e^{-n/16}).$$

Choosing $\beta = 2\log(2n_1n_2\ldots n_d)$ concludes the proof.

B. Proof of Theorem 5.1

In this section, we prove our main result, Theorem 5.1. Instead of proving Theorem 5.1 as stated, we will prove the following equivalent rescaled statement for when W_i have i.i.d. $\mathcal{N}(0, 1/n_i)$ entries. Because of this rescaling, $G(x)$ scales like $2^{-d/2}\|x\|$, the noise ω is assumed to scale like $2^{-d/2}$, ∇f scales like 2^d and α scales like 2^d. Theorem 5.1 is the $\epsilon = K/d^{90}$ case of what follows.

THEOREM B.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

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

Consider the iterates of Algorithm 1 with step size $\alpha = K_4 \frac{2^d}{d^2}$.

A. Then there exists a number of steps N upper bounded by

$$N \leqslant \frac{K_2}{d^4 \epsilon} \frac{f(x_0) 2^d}{\|x_*\|}$$

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

$$\|x_i - x_*\| \leqslant K_5 d^9 \|x_*\| \sqrt{\epsilon} + K_6 d^6 2^{d/2} \omega, \quad \text{for all } i \geqslant N, \tag{B.1}$$

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

B. In addition, for all $i \geqslant N$, we have

$$\|x_{i+1} - x_*\| \leqslant (1 - \alpha 7/8)^{i+1-N} \|x_N - x_*\| + K_8 2^{d/2} \omega \text{ and} \tag{B.2}$$

$$\|G(x_{i+1}) - G(x_*)\| \leqslant \frac{1.2}{2^{d/2}} (1 - \alpha 7/8)^{i+1-N} \|x_N - x_*\| + 1.2 K_8 \omega, \tag{B.3}$$

where $\alpha < 1$ is the step size of the algorithm. Here, $K_1, K_2, ..$ are numerical constants, and x_0 is the initial point in the optimization.

As mentioned in Section 5.1, our proof makes use of a deterministic condition, called the WDC, formally defined in Section 5.1. The following proposition establishes that the expansivity condition ensures that the WDC holds:

LEMMA B.2 (Lemma 11 in [9]). Fix $\epsilon \in (0, 1)$. If the entries 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.

It follows from Lemma B.2, that the WDC holds for all W_i simultaneously 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.

B.1 *Preliminaries*

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

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

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. In

order to prove this, we use a concentration argument, similar to the arguments used in [9]. First, define

$$\Lambda_x := \Pi_{i=d}^{1} W_{i,+,x},$$

with $W_{i,+,x}$ defined in Section 4 for notational convenience, and note that the step direction of our algorithm can be written as

$$\tilde{v}_x = \bar{v}_x + \bar{q}_x, \quad \text{with} \quad \bar{v}_x := \Lambda_x^t \Lambda_x x - (\Lambda_x)^t (\Lambda_{x_*}) x_*, \quad \text{and} \quad \bar{q}_x := \Lambda_x^t \eta. \tag{B.4}$$

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:

$$g(\theta) := \cos^{-1}\Big(\frac{(\pi - \theta)\cos\theta + \sin\theta}{\pi}\Big). \tag{B.5}$$

With this notation, we define

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

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 B.3 (Lemma 8 in [9]). Suppose that the WDC holds with $\epsilon < 1/(16\pi d^2)^2$. Then for all nonzero $x, x_* \in \mathbb{R}^k$,

$$\|\bar{v}_x - h_x\|_2 \leqslant K\frac{d^3\sqrt{\epsilon}}{2^d}\max(\|x\|_2, \|x_*\|_2), \text{ and} \tag{B.6}$$

$$\left\langle \Lambda_x x, \Lambda_{x_*} x_* \right\rangle \geqslant \frac{1}{4\pi}\frac{1}{2^d}\|x\|_2\|x_*\|_2, \text{ and} \tag{B.7}$$

$$\left\|\Lambda_x\right\|^2 \leqslant \frac{1}{2^d}(1+2\epsilon)^d \leqslant \frac{13}{12}2^{-d}. \tag{B.8}$$

Proof. Equations (B.6) and (B.7) are Lemma 8 in [9]. Regarding (B.8), note that the WDC implies that $\left\|W_{i,+,x}\right\|^2 \leqslant 1/2 + \epsilon$. It follows that:

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

where the last inequalities follow by our assumption on ϵ (i.e. $\epsilon < 1/(16\pi d^2)^2$). □

LEMMA B.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

$$\mathcal{E}_{\text{noise}} := \left\{\left\|\Lambda_x^t \eta\right\| \leqslant \frac{\omega}{2^{d/2}}, \text{ for all } x\right\}, \quad \omega := \sqrt{16\sigma\frac{k}{n}\log(n_1^d n_2^{d-1}\ldots n_d)} \tag{B.9}$$

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

This lemma is proved in Section B.6.

As the cost function f is not differentiable everywhere, we 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 $\mathscr{X}$ to $\mathbb{R}$, the Clarke generalized directional derivative of $\tilde{f}$ at the point $x \in \mathscr{X}$ 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

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

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

$$\partial f(x) = \operatorname{conv}(v_1, v_2, \ldots, v_t), \tag{B.10}$$

where conv denotes the convex hull of the vectors $v_1, \ldots, v_t$, 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 B.5 Under the assumption of Lemma B.4, and assuming that $\mathscr{E}_{\text{noise}}$ holds, we have that, for any $x \neq 0$ and any $v_x \in \partial f(x)$,

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

Proof. By (B.10), $\partial f(x) = \operatorname{conv}(v_1, \ldots v_t)$ for some finite t, and thus $v_x = a_1v_1 + \ldots a_tv_t$ for some $a_1, \ldots, a_t \geqslant 0$, $\sum_i a_i = 1$. For each v_i, there exists a w such that $v_i = \lim_{t\downarrow 0} \tilde{v}_{x+tw}$. On the event $\mathscr{E}_{\text{noise}}$, we have that for any $x \neq 0$, for any $\tilde{v}_x \in \partial f(x)$

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

where the last inequality follows from Lemmas B.3 and B.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 B.6 Suppose that the WDC holds with $\epsilon < 1/(16\pi d^2)^2$ and that the event $\mathscr{E}_{\text{noise}}$ holds with $\omega \leqslant \frac{2^{-d/2}\|x_*\|}{8\pi}$. Then, for all x, and all $v_x \in \partial f(x)$,

$$\|v_x\| \leqslant \frac{dK}{2^d}\max(\|x\|, \|x_*\|), \tag{B.11}$$

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

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

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

B.2 *Proof of Theorem* B.1A

We are now ready to prove Theorem B.1A. The logic of the proof is illustrated in Fig. 2. 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 B.7 Suppose that WDC holds with $\epsilon < 1/(16\pi d^2)^2$ and that $\mathscr{E}_{\text{noise}}$ holds with ω in (B.9) 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 $i > N$ and all $t \in [0,1]$ that $t\tilde{x}_i + (1-t)x_{i+1} \notin \mathscr{B}(0, \frac{1}{32\pi}\|x_*\|)$.

This lemma is proven in Section B.5.

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

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

with

$$\beta = 4Kd^3\sqrt{\epsilon} + 13\omega 2^{d/2}/\|x_*\|, \tag{B.12}$$

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 $\mathscr{S}_\beta$ is contained in two balls around x_* and $-\rho x_*$, whose radius is controlled by β:

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

$$\mathscr{S}_\beta \subset \mathscr{B}(x_*, 5000d^6\beta\|x_*\|_2) \cup \mathscr{B}(-\rho_d x_*, 500d^{11}\sqrt{\beta}\|x_*\|_2). \tag{B.13}$$

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

This lemma is proven in Section B.7.

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

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

where 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 the radius r is equal to the right-hand side in the error bound (B.1) 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 B.9 Suppose that the WDC holds with $\epsilon < 1/(16\pi d^2)^2$. Moreover, suppose that $\mathscr{E}_{\text{noise}}$ holds, and that ω in the event $\mathscr{E}_{\text{noise}}$ 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

$$f(x) < f(y) \tag{B.14}$$

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

This lemma is proven in Section B.8.

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

$$\mathscr{S}_\beta^+ := \mathscr{S}_\beta \cap \mathscr{B}(x_*, r), \text{ and}$$

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

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,

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

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

Consider i such that $x_i \notin \mathscr{S}_\beta$. By the mean value theorem [4, 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

$$\begin{aligned} f(x_i - \alpha\tilde{v}_{x_i}) - f(x_i) &= \langle v_{\hat{x}_i} - \alpha\tilde{v}_{x_i}\rangle \\ &= \langle \tilde{v}_{x_i} - \alpha\tilde{v}_{x_i}\rangle + \langle v_{\hat{x}_i} - \tilde{v}_{x_i} - \alpha\tilde{v}_{x_i}\rangle \\ &\leqslant -\alpha\|\tilde{v}_{x_i}\|^2 + \alpha\|v_{\hat{x}_i} - \tilde{v}_{x_i}\|\|\tilde{v}_{x_i}\| \\ &= -\alpha\|\tilde{v}_{x_i}\|(\|\tilde{v}_{x_i}\| - \|v_{\hat{x}_i} - \tilde{v}_{x_i}\|). \end{aligned} \tag{B.15}$$

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 $\tilde{v}_{x_i}$:

$$\|v_{\hat{x}_i} - \tilde{v}_{x_i}\| \leqslant \left(\frac{5}{6} + \alpha K_7 \frac{d^2}{2^d}\right)\|\tilde{v}_{x_i}\|, \text{for all} v_{\hat{x}_i} \in \partial f(\hat{x}_i). \tag{B.16}$$

Applying (B.16) to (B.15) 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 step size α being sufficiently small.

Thus, the maximum number of iterations for which $x_i \notin \mathscr{S}_\beta$ 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 $\mathscr{E}_{\text{noise}}$, for all $x \notin \mathscr{S}_\beta$, with β given by (B.12) that

$$\begin{aligned}
\|\tilde{v}_x\|_2 &\geqslant \|h_x\| - \|h_x - \tilde{v}_x\| \\
&\geqslant 2^{-d}\max(\|x\|, \|x_*\|)\left(\beta - K_1 d^3\sqrt{\epsilon} - \omega\frac{2^{d/2}}{\|x_*\|}\right) \\
&\geqslant 2^{-d}\max(\|x\|, \|x_*\|)\left(3Kd^3\sqrt{\epsilon} + 12\omega\frac{2^{d/2}}{\|x_*\|}\right) \\
&\geqslant 2^{-d}\|x_*\|3Kd^3\sqrt{\epsilon},
\end{aligned} \tag{B.17}$$

where the second inequality follows by the definition of $\mathscr{S}_\beta$ and Lemma B.5, and the third inequality follows from our definition of β in Equation (B.12). 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 $x_i \notin \mathscr{S}_\beta$.

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 the bound on $\|v_x\|$ given in Equation (B.11) and our choice of step size,

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

This concludes the proof of Theorem B.1A.

B.3 *Proof of Theorem* B.1B

Theorem B.1A establishes that after N iterations, the iterates x_i are inside a ball of radius $2r$ around x_*. With the assumption that $\epsilon \leqslant K_1/d^{90}$ for sufficiently small K_1 and the definition of r, this implies that the iterates lie in a ball around x_* of radius at most $K_3 d^{-10}\|x_*\|$. In this proof of Theorem B.1B, we prove convergence within this ball.

In this proof, we show that for any $i \geqslant N$, it holds that $x_i \in \mathscr{B}(x_*, a_4 d^{-10}\|x_*\|)$, $\tilde{x}_i = x_i$ and

$$\|x_{i+1} - x_*\| \leqslant b_2^{i+1-N}\|x_N - x_*\| + b_4 2^{d/2}\omega,$$

where K_3 is defined in Lemma B.9, $b_2 = 1 - \frac{\alpha}{2^d}\frac{7}{8}$ and b_4 is a universal constant.

We need Lemma B.10 that guarantees that the search directions of the iterates afterward point to x_* only up to the noise ω:

LEMMA B.10 Suppose the WDC holds with $200d\sqrt{d\sqrt{\epsilon}} < 1$ and $x \in \mathscr{B}(x_*, d\sqrt{\epsilon}\|x_*\|)$. Then for all $x \neq 0$ and for all $v_x \in \partial f(x)$,

$$\left\| v_x - \frac{1}{2^d}(x - x_*) \right\| \leqslant \frac{1}{2^d}\frac{1}{8}\|x - x_*\| + \frac{1}{2^{d/2}}\omega.$$

This lemma is proven in Section B.11.

Suppose $\tilde{x}_i \in \mathscr{B}(x_*, K_3 d^{-10}\|x_*\|)$. By the assumption $\epsilon \leqslant K_1/d^{90}$ for sufficiently small K_1, the assumptions in Lemma B.10 are met. Therefore,

$$\begin{aligned}
\|x_{i+1} - x_*\| &= \|\tilde{x}_i - \alpha v_{\tilde{x}_i} - x_*\| \\
&= \|\tilde{x}_i - x_* - \frac{\alpha}{2^d}(\tilde{x}_i - x_*) - \alpha v_{\tilde{x}_i} + \frac{\alpha}{2^d}(\tilde{x}_i - x_*)\| \\
&\leqslant \left(1 - \frac{\alpha}{2^d}\right)\|\tilde{x}_i - x_*\| + \alpha\|v_{\tilde{x}_i} - \frac{1}{2^d}(\tilde{x}_i - x_*)\| \\
&\leqslant \left(1 - \frac{\alpha}{2^d}\right)\|\tilde{x}_i - x_*\| + \alpha\left(\frac{1}{8}\frac{1}{2^d}\|\tilde{x}_i - x_*\| + \frac{1}{2^{d/2}}\omega\right) \\
&= \left(1 - \frac{\alpha}{2^d}\frac{7}{8}\right)\|\tilde{x}_i - x_*\| + \alpha\frac{1}{2^{d/2}}\omega, \qquad \text{(B.18)}
\end{aligned}$$

where the second inequality holds by Lemma B.10. By the assumptions $\tilde{x}_i \in \mathscr{B}(x_*, K_3 d^{-10}\|x_*\|)$, $\omega \leqslant \frac{K_1\|x_*\|}{d^{16}2^{d/2}}$, and using (B.18), we have $x_{i+1} \in \mathscr{B}(x_*, K_3 d^{-10}\|x_*\|)$. In addition, using Lemma B.9 yields that $\tilde{x}_{i+1} = x_{i+1}$. Repeat the above steps yields that $x_i \in \mathscr{B}(x_*, K_3 d^{-10}\|x_*\|)$ and $\tilde{x}_i = x_i$ for all $i \geqslant N$.

Using (B.18) and $\alpha = K_4\frac{2^d}{d^2}$, we have

$$\|x_{i+1} - x_*\| \leqslant b_2\|x_i - x_*\| + b_3\frac{2^{d/2}}{d^2}\omega, \qquad \text{(B.19)}$$

where $b_2 = 1 - 7K_4/(8d^2)$ and b_3 is a universal constant. Repeatedly applying (B.19) yields

$$\begin{aligned}
\|x_{i+1} - x_*\| &\leqslant b_2^{i+1-N}\|x_N - x_*\| + (b_2^{i-N} + b_2^{i-N-1} + \cdots + 1)\frac{b_3 2^{d/2}}{d^2}\omega \\
&\leqslant b_2^{i+1-N}\|x_N - x_*\| + \frac{b_3 2^{d/2}}{(1 - b_2)d^2}\omega \\
&\leqslant b_2^{i+1-N}\|x_N - x_*\| + b_4 2^{d/2}\omega,
\end{aligned}$$

where the last inequality follows from the definition of b_2, and b_4 is a universal constant. This finishes the proof for (B.2). Inequality (B.3) follows from Lemma B.17.

This concludes the proof.

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

B.4 *Proof of Equation (B16)*

Our proof relies on h_x being Lipschitz, as formalized by the lemma below, which is proven in Section B.10:

LEMMA B.11 For any $x, y \notin \mathscr{B}(0, K_0\|x_*\|)$, where K_0 and K_4 are numerical constants,

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

By Lemma B.11, for all $t \in [0,1]$ and $i > N$ (recall that by Lemma B.7, after at most N steps, $x_i \neq \mathscr{B}(0, K_0\|x_*\|)$):

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

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

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

where the second inequality is from Lemma B.5 and Equation (B.20), and the fourth inequality is from (B.11) and the assumption $\frac{\omega}{2^{-d/2}\|x_*\|_2} \leqslant K_9/d^2$.

Combining (B.21) and (B.17), 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 (B.16).

B.5 *Proof of Lemma B.7*

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

$$< x v_x < 0 \text{ and } \|v_x\| \geqslant \frac{1}{2^d 16\pi}\|x_*\| \text{ for all } x \in \mathscr{B}(0, 2K_0\|x_*\|) \text{ and } v_x \in \partial f(x). \tag{B.22}$$

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

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

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

Consider $\tilde{x}_i \notin \mathscr{B}(0, 2K_0\|x_*\|)$, and note that

$$\alpha\|v_{\tilde{x}_i}\| \leqslant \alpha\frac{dK}{2^d}\max(\|\tilde{x}_i\|, \|x_*\|) \leqslant \alpha\frac{16\pi Kd}{2^d}\|\tilde{x}_i\| \leqslant \frac{1}{2}\|\tilde{x}_i\|,$$

where the first inequality follows from (B.11), the second inequality from $\|\tilde{x}_i\| \geqslant 2K_0\|x_*\|$ and finally the last inequality from our assumption on the sufficiently small step size α. Therefore, from $x_{i+1} = \tilde{x}_i - \alpha v_{\tilde{x}_i}$, we have that $t\tilde{x}_i + (1-t)x_{i+1} \notin \mathscr{B}(0, K_0\|x_*\|)$ for all $t \in [0, 1]$, which completes the proof.

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

$$\begin{aligned}x^T\tilde{v}_x = &< \Lambda_x^T\Lambda_x x - \Lambda_x^T\Lambda_{x_*}x_* + \Lambda_x^T\eta x \\ &\leqslant \frac{13}{12}2^{-d}\|x\|^2 - \frac{1}{4\pi}\frac{1}{2^d}\|x\|\|x_*\| + \|x\|\frac{\omega}{2^{d/2}} \\ &\leqslant \|x\|\left(\frac{13}{12}2^{-d}\|x\| + \frac{1/(16\pi)}{2^d}\|x_*\| - \frac{1}{4\pi}\frac{1}{2^d}\|x_*\|\right) \\ &\leqslant \|x\|\frac{1}{2^d}\left(2\|x\| - \frac{3}{16\pi}\|x_*\|\right).\end{aligned}$$

The first inequality follows from (B.7) and (B.8), and the second inequality follows from our assumption on ω. Therefore, for any $x \in \mathscr{B}(0, \frac{1}{16\pi}\|x_*\|)$,

$< x\tilde{v}_x < -\frac{1}{16\pi 2^d}\|x\|\|x_*\| \leqslant 0$,

as desired.

If $G(x)$ is differentiable at x, then $v_x = \tilde{v}_x$ and $\langle xv_x\rangle < 0$. If $G(x)$ is not differentiable at x, by Equation (B.10), we have

$$\begin{aligned}x^T v_x = x^T(c_1v_1 + c_2v_2 + \cdots + c_tv_t) &\leqslant (c_1 + c_2 + \ldots + c_t)\|x\|\frac{1}{2^d}\left(2\|x\| - \frac{3}{16\pi}\|x_*\|\right) \\ = \|x\|\frac{1}{2^d}\left(2\|x\| - \frac{3}{16\pi}\|x_*\|\right) &< -\frac{1}{16\pi 2^d}\|x\|\|x_*\| \leqslant 0,\end{aligned} \tag{B.23}$$

for all $v_x \in \partial f(x)$.

Using (B.23) yields

$$\|v_x\| = \max_{\|u\|=1}\langle v_x u\rangle \geqslant \langle v_x - x/\|x\|\rangle = -\frac{x^T v_x}{\|x\|} > \frac{1}{2^d 16\pi}\|x_*\|,$$

which concludes the proof of (B.22). □

B.6 *Proof of Lemma B.4*

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

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

where P_{Λ_x} is a projector onto the span of Λ_x. As a consequence, $\left\|P_{\Lambda_x}\eta\right\|^2$ is χ^2-distributed random variable with k-degrees of freedom scaled by σ/n. A standard tail bound (see [17, p. 43]) yields that, for any $\beta \geqslant k$,

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

Next, we note that by applying [9, Lem. 16])[4] , with probability one, that the number of different matrices Λ_x can be bounded as

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

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

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

where $n = n_d$. Recall from (B.8) that $\left\|\Lambda_x\right\| \leqslant \frac{13}{12}$. Combining this inequality with $\left\|\bar{q}_x\right\|^2 \leqslant \left\|\Lambda_x\right\|^2 \left\|P_{\Lambda_x}\eta\right\|^2$ concludes the proof.

B.7 *Proof of Lemma B.8*

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

LEMMA B.12 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 $x \in \mathscr{S}_\beta$, then we have that either

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

or

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

In particular, we have

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

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

[4] 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 $x \in \mathscr{S}_\beta$.

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 $x \in \mathscr{S}_\beta$ implies

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

$$|\sin\overline{\theta}_0(r - \zeta)| \leqslant \beta M. \tag{B.26}$$

Now, we record several properties. We have:

$$\overline{\theta}_i \in [0, \pi/2] \text{ for } i \geqslant 1$$

$$\overline{\theta}_i \leqslant \overline{\theta}_{i-1} \text{ for } i \geqslant 1$$

$$|\xi| \leqslant 1 \tag{B.27}$$

$$|\zeta| \leqslant \frac{d}{\pi}\sin\theta_0 \tag{B.28}$$

$$\check{\theta}_i \leqslant \frac{3\pi}{i+3} \text{ for } i \geqslant 0 \tag{B.29}$$

$$\check{\theta}_i \geqslant \frac{\pi}{i+1} \text{ for } i \geqslant 0 \tag{B.30}$$

$$\xi = \prod_{i=0}^{d-1} \frac{\pi - \overline{\theta}_i}{\pi} \geqslant \frac{\pi - \overline{\theta}_0}{\pi} d^{-3} \tag{B.31}$$

$$\overline{\theta}_0 = \pi + O_1(\delta) \Rightarrow \overline{\theta}_i = \check{\theta}_i + O_1(i\delta) \tag{B.32}$$

$$\overline{\theta}_0 = \pi + O_1(\delta) \Rightarrow |\xi| \leqslant \frac{\delta}{\pi} \tag{B.33}$$

$$\overline{\theta}_0 = \pi + O_1(\delta) \Rightarrow \zeta = \rho_d + O_1(3d^3\delta) \text{ if} \frac{d^2\delta}{\pi} \leqslant 1. \tag{B.34}$$

We now establish (B.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 $\check{\theta}_i = g^{\circ i}(\check{\theta}_0) = g^{\circ i}(\pi) \leqslant \tilde{g}^{\circ i}(\pi) = \left(\frac{i}{3\pi} + \frac{1}{\pi}\right)^{-1} = \frac{3\pi}{i+3}$. Similarly, $g(\theta) \geqslant (\frac{1}{\pi} + \frac{1}{\theta})^{-1}$ implies that $\check{\theta}_i \geqslant \frac{\pi}{i+1}$, establishing (B.30).

We now establish (B.31). Using (B.29) and $\overline{\theta}_i \leqslant \check{\theta}_i$, 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 (B.32) by using the fact that $|g'(\theta)| \leqslant 1$ for all $\theta \in [0, \pi]$ and using the same logic as for [9, Equation 38].

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

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

So

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

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

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

Thus (B.34) holds.

Next, we establish that $x \in \mathscr{S}_\beta \Rightarrow r \leqslant 4d$, 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 (B.26) implies that $|r - \zeta| \leqslant \sqrt{2}\beta r$. Using (B.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 (B.25) implies that $|r - \zeta| \leqslant \sqrt{2}(\beta r + |\xi|)$. Using (B.27), (B.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:

1. (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 (B.25) yields $|\xi| \leqslant 2\sqrt{\beta}M$. Using (B.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 (B.26), we have $|r - \zeta| \leqslant \frac{\beta M}{\sin \overline{\theta}_0}$. Using this inequality in (B.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 (B.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 (B.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 (B.28), we also have $\zeta = O_1(\frac{d}{\pi}\delta)$. By (B.25), we have

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

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

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

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

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

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

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 (B.33) and (B.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 (B.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 $\delta = 8\pi d^4\sqrt{\beta} \leqslant 1$,

$$r - \rho_d = O_1\left(\beta M + \delta/\pi + 3d^3\delta + \frac{5}{2}\delta^2 d + \frac{3}{2}d^3\delta^3\right) \tag{B.40}$$

$$= O_1\left(4\beta d + \delta\left(\frac{1}{\pi} + 3d^3 + \frac{5}{2}d + \frac{3}{2}d^3\right)\right) \tag{B.41}$$

$$= O_1(200d^7\sqrt{\beta}). \tag{B.42}$$

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

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

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 [**9**, Equation (15)] and the fact that $\breve{\theta}_0 = \pi$, we get that

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

using (B.30), we have that

$$\prod_{j=i+1}^{d}\left(1-\frac{\check{\theta}_{j-1}}{\pi}\right)\leqslant\prod_{j=i+1}^{d}\left(1-\frac{1}{j}\right)=\exp\left(-\sum_{j=i+1}^{d}\frac{1}{j}\right)\leqslant\exp\left(-\int_{i+1}^{d+1}\frac{1}{s}\mathrm{d}s\right)=\frac{i+1}{d+1}.$$

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

B.8 *Proof of Lemma B.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 $x\in\mathscr{B}(\phi_d x_*,\varphi\|x_*\|)$ for a φ that will be specified later. Note that

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

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

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

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 $y \in \mathscr{B}(-\phi_d x_*, \varphi\|x_*\|)$

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

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

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

We can establish that:

LEMMA B.13 For all $d \geqslant 2$, that

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

This lemma is proven in Section B.9.

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.

B.9 *Proof of Lemma B.13*

It holds that

$$\|x - y\| \geqslant 2\sin(\theta_{x,y}/2)\min(\|x\|, \|y\|), \qquad \forall x, y \tag{B.47}$$

$$\sin(\theta/2) \geqslant \theta/4, \qquad \forall\theta \in [0, \pi] \tag{B.48}$$

$$\frac{d}{d\theta}g(\theta) \in [0, 1] \qquad \forall\theta \in [0, \pi] \tag{B.49}$$

$$\log(1 + x) \leqslant x \qquad \forall x \in [-0.5, 1] \tag{B.50}$$

$$\log(1 - x) \geqslant -2x \qquad \forall x \in [0, 0.75], \tag{B.51}$$

where $\theta_{x,y} = \angle(x, y)$. We recall the results (35), (36) and (49) in [9]:

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

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

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

where the second and the fifth inequalities follow from (B.50) and (B.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 $\sum_{i=1}^{n} i^3 = O(n^4)$.

B.10 *Proof of Lemma B.11*

To establish Lemma B.11, we prove the following:

LEMMA B.14 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 B.11 follows by noting that if $x, y \notin \mathscr{B}(0, r\|x_*\|)$, 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 B.14. For brevity of notation, let $\zeta_{j,z} = \prod_{i=j}^{d-1}\frac{\pi-\bar{\theta}_{i,z}}{\pi}$. Combining (B.47) and (B.48) gives $|\bar{\theta}_{0,x} - \bar{\theta}_{0,y}| \leqslant 4\max\left(\frac{1}{\|x\|}, \frac{1}{\|y\|}\right)\|x-y\|$. Inequality (B.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:

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

By Lemma B.15, we have

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

Additionally, it holds that

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

We have

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

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

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

Finally, combining (B.52)–(B.56) yields the result. □

LEMMA B.15 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

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

□

B.11 *Proof of Lemma B.10*

We first need Lemmas B.16, B.17 and B.18.

LEMMA B.16 Suppose $W \in \mathbb{R}^{n\times k}$ satisfies the WDC with constant ϵ. Then for any $x, y \in \mathbb{R}^k$, it holds that

$$\|W_{+,x}x - W_{+,y}y\| \leqslant \left(\sqrt{\frac{1}{2} + \epsilon} + \sqrt{2(2\epsilon + \theta)}\right)\|x - y\|,$$

where $\theta = \angle(x, y)$.

Proof. We have

$$\begin{aligned}\|W_{+,x}x - W_{+,y}y\| &\leqslant \|W_{+,x}x - W_{+,x}y\| + \|W_{+,x}y - W_{+,y}y\| \\ &= \|W_{+,x}(x - y)\| + \|(W_{+,x} - W_{+,y})y\| \leqslant \|W_{+,x}\|\|x - y\| + \|(W_{+,x} - W_{+,y})y\|. \end{aligned} \tag{B.57}$$

By WDC assumption, we have

$$\begin{aligned}\|W_{+,x}^T(W_{+,x} - W_{+,y})\| &\leqslant \left\|W_{+,x}^T W_{+,x} - I/2\right\| + \left\|W_{+,x}^T W_{+,y} - Q_{x,y}\right\| + \left\|Q_{x,y} - I/2\right\| \\ &\leqslant 2\epsilon + \theta. \end{aligned} \tag{B.58}$$

We also have

$$\begin{aligned}\|(W_{+,x} - W_{+,y})y\|^2 &= \sum_{i=1}^{n}(1_{w_i\cdot x>0} - 1_{w_i\cdot y>0})^2(w_i \cdot y)^2 \\ &\leqslant \sum_{i=1}^{n}(1_{w_i\cdot x>0} - 1_{w_i\cdot y>0})^2((w_i \cdot x)^2 + (w_i \cdot y)^2 - 2(w_i \cdot x)(w_i \cdot y)) \\ &= \sum_{i=1}^{n}(1_{w_i\cdot x>0} - 1_{w_i\cdot y>0})^2(w_i \cdot (x - y))^2 \\ &= \sum_{i=1}^{n}1_{w_i\cdot x>0}1_{w_i\cdot y\leqslant 0}(w_i \cdot (x - y))^2 + \sum_{i=1}^{n}1_{w_i\cdot x\leqslant 0}1_{w_i\cdot y>0}(w_i \cdot (x - y))^2 \\ &= (x - y)^T W_{+,x}^T(W_{+,x} - W_{+,y})(x - y) + (x - y)^T W_{+,y}^T(W_{+,y} - W_{+,x})(x - y) \\ &\leqslant 2(2\epsilon + \theta)\|x - y\|^2. \quad \text{(by B.58)} \end{aligned} \tag{B.59}$$

Combining (B.57), (B.59) and $\|W_{i,+,x}\|^2 \leqslant 1/2 + \epsilon$ given in [9, (9)] yields the result. □

LEMMA B.17 Suppose $x \in \mathscr{B}(x_*, d\sqrt{\epsilon}\|x_*\|)$, and the WDC holds with $\epsilon < 1/(200)^4/d^6$. Then it holds that

$$\left\| \prod_{i=j}^{1} W_{i,+,x} x - \prod_{i=j}^{1} W_{i,+,x_*} x_* \right\| \leqslant \frac{1.2}{2^{\frac{j}{2}}} \|x - x_*\|.$$

Proof. In this proof, we denote θ_{i,x,x_*} and $\bar{\theta}_{i,x,x_*}$ by θ_i and $\bar{\theta}_i$, respectively. Since $x \in \mathscr{B}(x_*, d\sqrt{\epsilon}\|x_*\|)$, we have

$$\bar{\theta}_i \leqslant \bar{\theta}_0 \leqslant 2d\sqrt{\epsilon}. \tag{B.60}$$

By [9, (14)], we also have $|\theta_i - \bar{\theta}_i| \leqslant 4i\sqrt{\epsilon} \leqslant 4d\sqrt{\epsilon}$. It follows that:

$$\begin{aligned} 2\sqrt{\theta_i + 2\epsilon} &\leqslant 2\sqrt{\bar{\theta}_i + 4d\sqrt{\epsilon} + 2\epsilon} \leqslant 2\sqrt{2d\sqrt{\epsilon} + 4d\sqrt{\epsilon} + 2\epsilon} \\ &\leqslant 2\sqrt{8d\sqrt{\epsilon}} \leqslant \frac{1}{30d}. \text{ (by the assumption on } \epsilon) \end{aligned} \tag{B.61}$$

Note that $\sqrt{1+2\epsilon} \leqslant 1 + \epsilon \leqslant 1 + \sqrt{d\sqrt{\epsilon}}$. We have

$$\prod_{i=d-1}^{0} \left(\sqrt{1+2\epsilon} + 2\sqrt{\theta_i + 2\epsilon} \right) \leqslant \left(1 + 7\sqrt{d\sqrt{\epsilon}} \right)^d \leqslant 1 + 14d\sqrt{d\sqrt{\epsilon}} \leqslant \frac{107}{100} < 1.2,$$

where the second inequality is from that $(1+x)^d \leqslant 1 + 2dx$ if $0 < xd < 1$. Combining the above inequality with Lemma B.16 yields

$$\left\| \prod_{i=j}^{1} W_{i,+,x} x - \prod_{i=j}^{1} W_{i,+,x_*} x_* \right\| \leqslant \prod_{i=j-1}^{0} \left(\sqrt{\frac{1}{2} + \epsilon} + \sqrt{2}\sqrt{\theta_i + 2\epsilon} \right) \|x - x_*\| \leqslant \frac{1.2}{2^{\frac{j}{2}}} \|x - x_*\|.$$

□

LEMMA B.18 Suppose $x \in \mathscr{B}(x_*, d\sqrt{\epsilon}\|x_*\|)$, and the WDC holds with $\epsilon < 1/(200)^4/d^6$. Then it holds that

$$\left(\prod_{i=d}^{1} W_{i,+,x} \right)^T \left[\left(\prod_{i=d}^{1} W_{i,+,x} \right) x - \left(\prod_{i=d}^{1} W_{i,+,x_*} \right) x_* \right] = \frac{1}{2^d}(x - x_*) + \frac{1}{2^d}\frac{1}{16}\|x - x_*\| O_1(1).$$

Proof. For brevity of notation, let $\Lambda_{j,k,z} = \prod_{i=j}^{k} W_{i,+,z}$. We have

$$\begin{aligned} &\Lambda_{d,1,x}^T \left(\Lambda_{d,1,x} x - \Lambda_{d,1,x_*} x_* \right) \\ &= \Lambda_{d,1,x}^T \left[\Lambda_{d,1,x} x - \sum_{j=1}^{d} \left(\Lambda_{d,j,x} \Lambda_{j-1,1,x_*} x_* \right) + \sum_{j=1}^{d} \left(\Lambda_{d,j,x} \Lambda_{j-1,1,x_*} x_* \right) - \Lambda_{d,1,x_*} x_* \right] \\ &= \underbrace{\Lambda_{d,1,x}^T \Lambda_{d,1,x} (x - x_*)}_{T_1} + \underbrace{\Lambda_{d,1,x}^T \sum_{j=1}^{d} \Lambda_{d,j+1,x} \left(W_{j,+,x} - W_{j,+,x_*} \right) \Lambda_{j-1,1,x_*} x_*}_{T_2}. \end{aligned} \tag{B.62}$$

For T_1, we have

$$T_1 = \frac{1}{2^d}(x - x_*) + \frac{4d}{2^d}\|x - x_*\| O_1(\epsilon). \quad [9, (10)] \tag{B.63}$$

For T_2, we have

$$\begin{aligned} T_2 &= O_1(1) \sum_{j=1}^{d} \left(\frac{1}{2^{d-\frac{j}{2}}} + \frac{(4d-2j)\epsilon}{2^{d-\frac{j}{2}}} \right) \left\| (W_{j,+,x} - W_{j,+,x_*}) \Lambda_{j-1,1,x_*} x_* \right\| \\ &= O_1(1) \sum_{j=1}^{d} \left(\frac{1}{2^{d-\frac{j}{2}}} + \frac{(4d-2j)\epsilon}{2^{d-\frac{j}{2}}} \right) \left\| (\Lambda_{j-1,1,x} x - \Lambda_{j-1,1,x_*} x_*) \right\| \sqrt{2(\theta_{i,x,x_*} + 2\epsilon)} \\ &= O_1(1) \sum_{j=1}^{d} \left(\frac{1}{2^{d-\frac{j}{2}}} + \frac{(4d-2j)\epsilon}{2^{d-\frac{j}{2}}} \right) \frac{1.2}{2^{\frac{j}{2}}} \|x - x_*\| \frac{1}{30\sqrt{2}d} \\ &= \frac{1}{16} \frac{1}{2^d} \|x - x_*\| O_1(1), \end{aligned} \tag{B.64}$$

where the first Equation is by [**9**, (10)], the second equation is by (B.59), the third equation is by Lemma B.17 and (B.61). The result follows from (B.62), (B.63) and (B.64). □

Now, we are ready to prove Lemma (B.10). For brevity of notation, let $\Lambda_{j,z} = \prod_{i=j}^{1} W_{i,+,z}$. Using Lemma B.18 yields

$$\|\bar{v}_x - \frac{1}{2^d}(x - x_*)\| \leqslant \frac{1}{2^d} \frac{1}{16} \|x - x_*\|.$$

It follows that:

$$\left\| \tilde{v}_x - \frac{1}{2^d}(x - x_*) \right\| = \left\| \bar{v}_x + \bar{q}_x - \frac{1}{2^d}(x - x_*) \right\| \leqslant \frac{1}{2^d} \frac{1}{16} \|x - x_*\| + \frac{1}{2^{d/2}} \omega.$$

For any $x \neq 0$ and for any $v \in \partial f(x)$, by (B.10), there exist $c_1, c_2, \ldots, c_t \geqslant 0$ such that $c_1 + c_2 + \ldots + c_t = 1$ and $v = c_1 v_1 + c_2 v_2 + \ldots + c_t v_t$. It follows that $\|v - \frac{1}{2^d}(x - x_*)\| \leqslant \sum_{j=1}^{t} c_j \|v_j - \frac{1}{2^d}(x - x_*)\| \leqslant \frac{1}{2^d} \frac{1}{16} \|x - x_*\| + \frac{1}{2^{d/2}} \omega$.