Journal of Fourier Analysis and Applications (2021) 27:19
https://doi.org/10.1007/s00041-021-09830-5

A Provably Convergent Scheme for Compressive Sensing Under Random Generative Priors

Wen Huang[1] · Paul Hand[2] · Reinhard Heckel[3] · Vladislav Voroninski[4]

Received: 21 November 2019 / Revised: 14 November 2020 / Accepted: 5 February 2021 /
Published online: 11 March 2021

Abstract

Deep generative modeling has led to new and state of the art approaches for enforcing structural priors in a variety of inverse problems. In contrast to priors given by sparsity, deep models can provide direct low-dimensional parameterizations of the manifold of images or signals belonging to a particular natural class, allowing for recovery algorithms to be posed in a low-dimensional space. This dimensionality may even be lower than the sparsity level of the same signals when viewed in a fixed basis. What is not known about these methods is whether there are computationally efficient algorithms whose sample complexity is optimal in the dimensionality of the representation given by the generative model. In this paper, we present such an algorithm and analysis. Under the assumption that the generative model is a neural network that is sufficiently expansive at each layer and has Gaussian weights, we provide a gradient descent scheme and prove that for noisy compressive measurements of a signal in the range of the model, the algorithm converges to that signal, up to the noise level. The scaling of the sample complexity with respect to the input dimensionality of the generative prior is linear, and thus can not be improved except for constants and factors of other variables. To the best of the authors' knowledge, this is the first recovery guarantee for compressive sensing under generative priors by a computationally efficient algorithm.

Keywords Compressive sensing · Generative models · Convergence analysis · Gradient descent

Communicated by Roman Vershynin.

✉ Wen Huang
huwst08@gmail.com

1 Department of Mathematical Sciences, Xiamen University, Xiamen, China

2 Department of Mathematics and Khoury College of Computer Sciences, Northeastern University, Boston, USA

3 Department of Electrical and Computer Engineering, Rice University, Houston, USA

4 Helm.ai, Menlo Park, USA

1 Introduction

Generative models have greatly improved the state of the art in computer vision and image processing, including inpainting, superresolution, compression, compressive sensing, image manipulation, MRI imaging, and denoising [10,14,18,20–23,23,24, 24,27–31]. These models are learned in an unsupervised way from a dataset of images relevant for a particular domain, and they permit generation of new samples of the same distribution underlying the the training data. They can be trained using a variety of techniques, including Generative Adversarial Networks [10] and Variational Autoencoders [17]. The performance of generative models has improved substantially over the past several years. For example, multiple methods can now generate synthetic photorealistic images of human faces [15,16].

In many imaging inverse problems, an image is to be recovered from few and/or noisy measurements. In the case of undersampled linear measurements, this problem is known as compressive sensing. Structural assumptions are necessary in order to recover the desired image because of undersampling. Generative models can provide such a structural assumption, known as a prior, for inverse problems. Some generative models are of the form of a learned function $G : \mathbb{R}^k \to \mathbb{R}^n$, where n is the dimensionality of the measured image, and $k \ll n$. The domain of G is a low-dimensional space, known as a latent code space. The range of G is a manifold in $\mathbb{R}^n$ that approximates a domain-specific set of images, known as a natural signal manifold. An inverse problem can be regularized by seeking an image in the range of G that is most consistent with provided measurements.

A common structural prior for imaging inverse problems over the past few decades has been a sparsity prior [8,9]. With it, images are modeled to be approximately sparse in an appropriate bases, such as a Fourier of Wavelet basis. An inverse problem can be regularized by searching for a sparse solution to a provided set of measurements. This results in a combinatorially hard optimization problem. In the case of linear compressive measurements, a convex relaxation based on L_1 minimization can be solved instead, which admits signal recovery under generic measurements at optimal sample complexity with respect to the sparsity level of a signal.

In the context of compressive sensing under generative priors, recovery can be posed as a nonconvex empirical risk optimization, which can be solved by first order gradient methods. When solved this way, generative models have been shown to empirically outperform sparsity models in the sense that they can give comparable reconstruction error with 5 to 10 times fewer compressive measurements in some contexts [4]. This empirical result indicates both that representations from generative models are low dimensional and can be efficiently exploited. Nonetheless, this observation does not have a firm theoretical footing. In principle, such gradient algorithms for nonconvex programs could get stuck in local minima. Thus, it is important to provide algorithms that provably recover the underlying signal.

In this paper, we introduce a gradient descent algorithm for empirical risk minimization under a generative network, given noisy compressive measurements of its output. We prove that if the network is random, the size of each layer grows appropriately, there are a sufficient number of compressive measurements, and the magnitude of the noise is sufficiently small, then the gradient descent algorithm converges to a

neighborhood of the global optimizer and the size of the neighborhood only depends on the magnitude of the noise. In particular, the gradient descent algorithm converges to the global minimizer for noiseless measurements. To the best of our knowledge, this is the first recovery guarantee for compressive sensing under a generative neural network model. Using numerical experiments, we empirically verify recovery up to the noise level, and in particular exact recovery in the noiseless case.

1.1 Relation to Previous Theoretical Work

A first theoretical analysis of compressive sensing under a generative prior appeared in [4]. In that work, the authors studied the task of recovering a signal near the range of a generative network by the same nonconvex empirical risk objective as in the present paper. They establish that if the number of measurements scales linearly in the latent dimensionality, then if one can solve to with an additive constant of global optimality the nonconvex empirical risk objective, then one recovers the signal to within the noise level, optimization error, and representational error of the network. Because the objective is nonconvex, and nonconvex problems are NP-hard in general, it is not clear that any particular computationally efficient optimization algorithm can actually find the global optimum. That is, it is possible that any particular numerically efficient optimization algorithm gets stuck in local minima. In the present paper, we provide a specific computationally efficient numerical algorithm and establish a recovery guarantee for compressive sensing under generative models that satisfy suitable architectural assumptions.

A recent paper by a subset of the authors [12] provides a global analysis of the nonconvex empirical risk objective below for expansive-Gaussian networks. That paper shows that, under appropriate conditions, there are descent directions, of the nonconvex objective, outside neighborhoods of the global optimizer and a negative multiple thereof in the latent code space. That work, however, does not provide an analysis of the behavior of the empirical risk objective within these two neighborhoods, a specific algorithm, a proof of convergence of an algorithm, or a principled reason why the negative multiple of the global optimizer would not be returned by a naively applied gradient scheme. Additionally, that work does not study noise tolerance. Each of these aspects require considerable technical advances, for example establishing a nontrivial convexity-like property near the global minimizer.

The paper [2] presents a simple layer-wise inversion process for neural networks. In the current setting, this result is not applicable because the final compressive layer can not be directly inverted without structural assumptions. Instead, in the present paper, we analyze the inversion of the compressive measurements and the generative network together.

2 Problem Statement

We consider a generator $G\colon \mathbb{R}^k \to \mathbb{R}^n$ with $k \ll n$, given by a d-layer network of the form

$$G(x) = \text{relu}(W_d \ldots \text{relu}(W_2\, \text{relu}(W_1 x)) \ldots),$$

where $\text{relu}(a) = \max(a, 0)$ applies entrywise, $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ are the weights of the network, and $n_0 = k$ and $n_d = n$ are, respectively, the dimensionality of the input and output of G. This model for G is a d-layer neural network with no bias terms. Let $\mathfrak{y}_* = G(x_*) \in \mathbb{R}^n$ be an image in the range of the generator G, and let $A \in \mathbb{R}^{m \times n}$ be a measurement matrix, where typically $m \ll n$.

Our goal is to estimate the image $\mathfrak{y}_*$ from noisy compressive measurements $y = AG(x_*) + e$, where A and G are known and $e \in \mathbb{R}^m$ is an unknown noise vector. To estimate this image, we first estimate its latent code, x, and then compute $G(x)$. In order to estimate x_*, we consider minimization of the empirical risk

$$\min_{x \in \mathbb{R}^k} \frac{1}{2} \|AG(x) - y\|^2. \tag{2.1}$$

Throughout this paper, $\|\cdot\|$ denotes 2-norm of a vector or a matrix. For notational convenience, we let $W_{+,x}$ denote the matrix obtained by zeroing out the rows of W that do not have a positive dot product with x, i.e.,

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

where $\text{diag}(Wx > 0)$ denotes the diagonal matrix whose (i, i)th entry is 1 if $(Wx)_i > 0$ and 0 otherwise. Furthermore, we define $W_{1,+,x} = (W_1)_{+,x} = \text{diag}(W_1 x > 0)W_1$ and

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

The matrix $W_{i,+,x}$ contains the rows of W_i that are active after taking a ReLU if the input to the network is x. Therefore, under the model for G, the empirical risk (2.1) becomes

$$f(x) = \frac{1}{2} \left\| A \left(\prod_{i=d}^{1} W_{i,+,x} \right) x - A \left(\prod_{i=d}^{1} W_{i,+,x_*} \right) x_* - e \right\|^2, \tag{2.2}$$

where $\prod_{i=d}^{1} W_{i,+,x} = W_{d,+,x} W_{d-1,+,x} \ldots W_{1,+,x}$ and likewise for $\prod_{i=d}^{1} W_{i,+,x_*}$.

3 Main Results: Two Algorithms and a Convergence Analysis

In this section, we propose two closely related algorithms for minimizing the empirical loss (2.1). The first algorithm is a subgradient descent method for which we prove

convergence. The second algorithm is a practical implementation that can be directly implemented with an explicit form of the gradient step that may or may not be within the subdifferential of the objective at some points.

3.1 A Provably Convergent Subgradient Descent Method

In order to state the first algorithm, Algorithm 1, we first introduce the notion of a subgradient. Since the cost function $f(x)$ is continuous, piecewise quadratic, and not differentiable everywhere, we use the notion of a generalized gradient, called the Clarke subdifferential or generalized subdifferential [6]. If a function f is Lipschitz from a Hilbert space $\mathcal{X}$ to $\mathbb{R}$, the Clarke generalized directional derivative of f at the point $x \in \mathcal{X}$ in the direction u, denoted by $f^o(x;u)$, is defined by $f^o(x;u) = \limsup_{y \to x, t \downarrow 0} \frac{f(y+tu)-f(y)}{t}$, and the generalized subdifferential of f at x, denoted by $\partial f(x)$, is defined by

$$\partial f(x) = \{v \in \mathbb{R}^k \mid \langle v, u \rangle \leq f^o(x;u), \forall u \in \mathcal{X}\}.$$

Any vector in $\partial f(x)$ is called a subgradient of f at x. Note that if f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$.

Algorithm 1 Provably convergent subgradient descent method

Input: Weights of the network W_i; noisy observation y; and step size $\nu > 0$;
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

$$\tilde{v}_{\tilde{x}_i} := \left(\prod_{i=d}^{1} W_{i,+,\tilde{x}_i}\right)^T A^T \left(A \left(\prod_{i=d}^{1} W_{i,+,\tilde{x}_i}\right) \tilde{x}_i - y\right);$$

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

We can now state Algorithm 1. It is a subgradient method and has an important twist. In lines 3–7, the algorithm checks whether negating the current iterate of the latent code causes a lower objective, and if so accepts that negation. The motivation for this step is as follows. The expectation of the empirical loss f is shown in Fig. 1. It contains a global minimum at x_*, a local maximum at 0, and a critical point at $-x_*\rho_d$, where $\rho_d \in (0,1)$, as established in [12]. The objective value in a neighborhood of the spurious critical point is higher than that of the negation of that neighborhood. Thus, while a simple gradient descent algorithm could in principle be attracted to $-x_*\rho_d$,

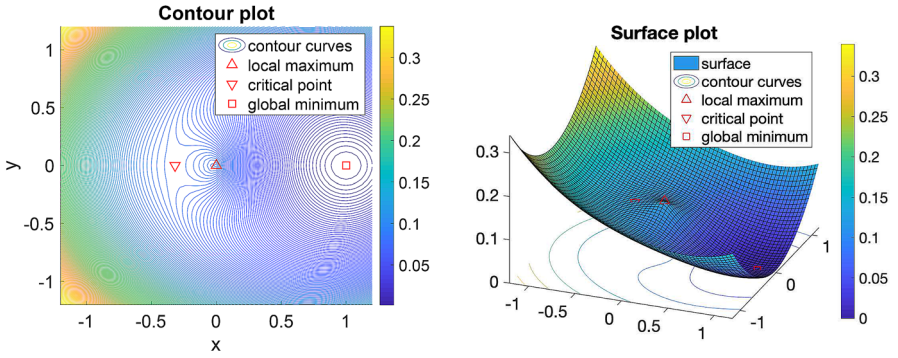

Fig. 1 The expectation of the empirical loss of f with $d = 2$, $k = 2$, and $x_* = (1, 0)^T$. It contains a global minimum at x_*, a local maximum at $(0, 0)^T$, and a critical point at $-\rho_d x_* = -x_*/\pi$

the check in lines 3 will be activated if any iterate is is sufficiently close to $-x_*\rho_d$, resulting in the next iterate jumping to the basin of attraction of the global minimizer x_*.

3.2 Convergence Analysis for Algorithm 1

In this section, we prove that under suitable conditions Algorithm 1 converges to the global minimizer x_* up to an error determined by the noise e. Consequently, the signal estimate $G(x_*)$ is also recovered up to an error determined by the noise. In the noiseless case (i.e., $e = 0$), Algorithm 1 converges to x^*, and $G(x^*)$ is recovered exactly, under suitable conditions. We will prove this in the case of deterministic assumptions on the generative model, and we will obtain as a corollary that the convergence proof holds for a suitable random model for G. Both the deterministic and probabilistic assumptions are the same as in [12], which did not contain a convergence analysis. The structure of this section is as follows: first we state the deterministic conditions on G, then we state a convergence guarantee under those conditions, then we state a probabilistic model for G, and finally we state a corollary for convergence under that probabilistic model.

3.2.1 Deterministic Conditions

We now state two sufficient deterministic conditions for convergence of Algorithm 1, both of which were introduced in [12]. First, we assume that the weights of the network, W_i, satisfy the Weight Distribution Condition (WDC) defined below. This condition states that the weights are roughly uniformly distributed over a sphere of an appropriate radius.

Definition 3.1 (*Weight Distribution Condition* (*WDC*)) A matrix $W \in \mathbb{R}^{n\times k}$ satisfies the *Weight Distribution Condition* with constant ϵ if for all nonzero $x, y \in \mathbb{R}^k$, it holds

that

$$\left\| \sum_{i=1}^{n} 1_{w_i \cdot x > 0} 1_{w_i \cdot y > 0} \cdot w_i w_i^T - Q_{x,y} \right\| \leq \epsilon, \text{ with } Q_{x,y} = \frac{\pi - \theta}{2\pi} I + \frac{\sin\theta}{2\pi} M_{\hat{x} \leftrightarrow \hat{y}}.$$

Here, $w_i^T \in \mathbb{R}^k$ is the ith row of W; $M_{\hat{x} \leftrightarrow \hat{y}} \in \mathbb{R}^{k \times k}$ is the matrix such that $\hat{x} \mapsto \hat{y}$, $\hat{y} \mapsto \hat{x}$, and $\hat{z} \mapsto 0$ for all $z \in \text{span}(\{x, y\})^{\perp}$; $\hat{x} = x/\|x\|$, and $\hat{y} = y/\|y\|$; $\theta = \angle(x, y)$; and 1_S is the indicator function on S.

Second, we assume that the measurement matrix A satisfies an isometry condition with respect to G, defined below.

Definition 3.2 (*Range Restricted Isometry Condition* (*RRIC*)) A matrix $A \in \mathbb{R}^{m \times n}$ satisfies the *Range Restricted Isometry Condition* with respect to G with constant ϵ if for all $x_1, x_2, x_3, x_4 \in \mathbb{R}^k$, it holds that

$$\begin{aligned} &|\langle A(G(x_1) - G(x_2)), A(G(x_3) - G(x_4))\rangle - \langle G(x_1) - G(x_2), G(x_3) - G(x_4)\rangle| \\ &\quad \leq \epsilon \|G(x_1) - G(x_2)\| \|G(x_3) - G(x_4)\|. \end{aligned}$$

3.2.2 Convergence Guarantee Under Deterministic Conditions

As our main theoretical result, we prove that the iterates generated by Algorithm 1 converge to x_* up to a term dependent on the noise level, provided that the deterministic WDC and RRIC conditions are met. The proof is given in Sect. 5.

Theorem 3.1 *Suppose the WDC and RRIC hold with $\epsilon \leq K_1/d^{90}$ and the noise e obeys $\|e\| \leq \frac{K_2\|x_*\|}{d^{42}2^{d/2}}$. Consider the iterates $\{x_i\}$ generated by Algorithm 1 with step size $\nu = K_3\frac{2^d}{d^2}$. Then there exists a number of iterations, denoted by N and upper bounded by $N \leq \frac{K_4 f(x_0) 2^d}{d^4 \epsilon \|x_*\|}$ such that*

$$\|x_N - x_*\| \leq K_5 d^9 \|x_*\| \sqrt{\epsilon} + K_6 d^6 2^{d/2} \|e\|. \tag{3.1}$$

In addition, for all $i \geq N$, we have

$$\|x_{i+1} - x_*\| \leq C^{i+1-N} \|x_N - x_*\| + K_7 2^{d/2} \|e\| \text{ and} \tag{3.2}$$

$$\|G(x_{i+1}) - G(x_*)\| \leq \frac{1.2}{2^{d/2}} C^{i+1-N} \|x_N - x_*\| + 1.2 K_7 \|e\|, \tag{3.3}$$

where $C = 1 - \frac{\nu}{2^d}\frac{7}{8} \in (0, 1)$. Here, $K_1, \ldots, K_7$ are universal positive constants.

Theorem 3.1 shows that after a certain number of iterations N, an iterate of Algorithm 1 is in a neighborhood of the true latent code x_*, and the size of this neighborhood depends on the parameter ϵ and the noise e (see (3.1)). Furthermore, by (3.2), the theorem guarantees that once the iterates are in this ball, they converge linearly to a smaller neighborhood of x_*, and the size of the neighborhood only depends on the noise term e.

If the noise term is zero, the algorithm converges linearly to x_*. Similarly, it follows from (3.3) that the recovered image $G(x_i)$ converges to $G(x_*)$ up to the noise level.

Note that the factors 2^d in the theorem are an artifact of the choice of scaling for the entries of W_i. Under the WDC, each $W_{i,+,x}$ has spectral norm at most approximately $1/2$, as the operation $\text{relu}(Wx)$ returns approximately half of the entries of Wx. Because of this, $G(x) = \left(\prod_{i=d}^{1} W_{i,+,x}\right)x$ scales like $2^{d/2}\|x\|$, and thus the noise e must scale like $2^{-d/2}$ and the step size ν must scale like 2^d. All of these scalings would be unity with respect to d under an alternate choice of the scaling of W_i.

3.2.3 Probabilistic Assumptions on Network Architecture and Measurements

We next provide a convergence guarantee for a model of a trained neural network G. As in [12], we consider an *expansive-Gaussian* network as a model for a trained network G, and we consider $\Omega(k)$ Gaussian measurements, as provided in the following assumptions:

(a) The network weights have i.i.d. $\mathcal{N}(0, 1/n_i)$ entries in the ith layer.
(b) The network is expansive in each layer, in that

$$n_i \geq c\epsilon^{-2}\log(1/\epsilon)n_{i-1}\log n_{i-1}, \tag{3.4}$$

where c is a universal constant and ϵ is sufficiently small.
(c) The measurement vectors have i.i.d. $\mathcal{N}(0, 1/m)$ entries.
(d) There are a sufficient number of measurements in that

$$m \geq c\epsilon^{-1}\log(1/\epsilon)dk\log\prod_{i=1}^{d} n_i, \tag{3.5}$$

where c is a universal constant and ϵ is sufficiently small.

As established in [12, Proposition 6], the probability that the WDC and RRIC hold with constant ϵ under the assumption above is at least

$$1 - \sum_{i=2}^{d} \tilde{c}n_i e^{-\gamma n_{i-2}} - \tilde{c}n_1 e^{-\gamma\epsilon^2\log(1/\epsilon)k} - \tilde{c}e^{-\gamma\epsilon m}, \tag{3.6}$$

where γ, and $\tilde{c}$ are universal constants.

The motivation for assumptions (a)–(d) is as follows. The Gaussian assumption of network weights is motivated by the observation that in some trained networks, such as AlexNet, the weights are approximately Gaussian [2]. Moreover, recent papers [1,7,19,26] assume that the weights are initialized during training by Gaussian distributions and establish that with sufficient overparameterization, the weights are only perturbed slightly during the training process. The expansiveness assumption is a natural condition given that the generator maps low-dimensional, highly-compressed representations to high-dimensional signals with substantial redundancy. The Gaussian assumption in the weights of the measurement matrix A has been well-accepted

and widely used, see e.g., [3,5]. Finally, the assumption on the number of measurements is optimal in k as $\Omega(k)$ measurements are needed to ensure uniqueness of recovering a point on a k-dimensional manifold under generic linear measurements. More discussion and justifications of these assumptions can be found in [12, Sect. D].

3.2.4 Convergence Guarantee Under Probabilistic Assumptions

Under assumptions (a)–(d), we can now state a convergence guarantee as a corollary. Combining Proposition 6 from the paper [12], with Theorem 3.1 yields the following. It states that if G is Gaussian and sufficiently expansive at each layer, and if the measurements are Gaussian, then under sample complexity $\Omega(k)$, the empirical risk optimization (2.1) can be provably optimized up to the noise level by the polynomial-time first order Algorithm 1 with high probability.

Corollary 3.1 *Consider an expansive-Gaussian generative neural network G that satisfies* (a) *and* (b)*, and let the measurements satisfy* (c) *and* (d)*. Suppose $\epsilon < K_1/d^{90}$ and $\|e\| \leq \frac{K_2\|x_*\|}{d^4 2^{d/2}}$. Then, at least with probability* (3.6)*, the iterates $\{x_i\}$ generated by Algorithm 1 with step size $\nu = K_3\frac{2^d}{d^2}$ satisfies the following: There exists a number of steps N upper bounded by $N \leq \frac{K_4 f(x_0)2^d}{d^4\epsilon\|x_*\|}$ such that*

$$\|x_N - x_*\| \leq K_5 d^9\|x_*\|\sqrt{\epsilon} + K_6 d^6 2^{d/2}\|e\|.$$

In addition, for all $i \geq N$, we have

$$\|x_{i+1} - x_*\| \leq C^{i+1-N}\|x_N - x_*\| + K_7 2^{d/2}\|e\|,$$
$$\|G(x_{i+1}) - G(x_*)\| \leq \frac{1.2}{2^{d/2}}C^{i+1-N}\|x_N - x_*\| + 1.2K_7\|e\|,$$

where $C = 1 - \frac{\nu}{2^d}\frac{7}{8} \in (0,1)$. Here, $\gamma, c, \tilde{c}, K_1, \ldots, K_7$ are universal positive constants.

As with Theorem 3.1, the factors of 2^d are artifacts of the choice of scaling of W_i. Had the entries of W_i been scaled like $\mathcal{N}(0, 2/n_i)$, these factors would not be present.

3.3 Practical Algorithm

The empirical risk objective is nondifferentiable on a set of measure zero. At points of nondifferentiability, Algorithn 1 requires selection of a subgradient $\partial f(\tilde{x}_i)$. Such a subgradient could be determined by computing $\nabla f(\tilde{x}_i + \delta w)$ for a random w and sufficiently small δ. This is because $f(x)$ is a piecewise quadratic function, and by [6, Theorem 9.6], we can express the sub-differential as

$$\partial f(x) = conv(v_1, v_2, \ldots, v_t), \tag{3.7}$$

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. Because this computation of a subgradient is not explicit, we propose another algorithm, Algorithm 2, where the step direction is simply chosen as $v_{\tilde{x}_i} = \tilde{v}_{\tilde{x}_i}$. In practice, it is extremely unlikely to have an iterate on which the function is not differentiable. In other words, Algorithm 1 reduces in practice to Algorithm 2. However, strictly speaking, the convergence analysis does not apply for Algorithm 2 because of the possibility that $\tilde{v}_{\tilde{x}_i}$ is not a subgradient at $\tilde{x}_i$.

Algorithm 2 Practical gradient descent method

Input: Weights of the network W_i; noisy observation y; and step size $\nu > 0$;
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 $\tilde{v}_{\tilde{x}_i} := \left(\prod_{i=d}^{1} W_{i,+,\tilde{x}_i}\right)^T A^T (A \left(\prod_{i=d}^{1} W_{i,+,\tilde{x}_i}\right) \tilde{x}_i - y)$;
9: $x_{i+1} = \tilde{x}_i - \nu \tilde{v}_{\tilde{x}_i}$;
10: **end for**

4 Experiments

In this section, we tested the performance of Algorithm 2 on synthetic data with various sizes of noise, and verified Theorem 3.1 by numerical results. Note that we do not observe that any entry in $W_i W_{i-1,+,x} \ldots W_{2,+,x} W_{1,+,x} x$, for any i, is zero in our experiments. Therefore, Algorithm 2 is equivalent to Algorithm 1 in this case.

The entries of A are drawn from $\mathcal{N}(0, 1/m)$ and the entries in W_i are drawn from $\mathcal{N}(0, 1/n_i)$. We consider a two-layer network with multiple numbers of input neurons k shown in Fig. 2. The numbers of neurons in the middle layer and output layer are fixed to be 250 and 600, respectively. The number of rows in the measurement matrix A is $m = 150$. The latent code x_* and the noise $\tilde{e}$ are drawn from the standard normal distribution. The noisy measurement y is set to be $y = AG(x_*) + \tau \tilde{e}/\|\tilde{e}\|$, and four values of τ are used such that the signal to noise ratio (SNR) values are 40, 80, 120 and inf, where SNR is defined to be $10 \log_{10} \left(\frac{\|AG(x_*)\|}{\|e\|}\right)$. The step size is chosen to be $2^d/d^2$, which is 1 since $d = 2$. Algorithm 2 stops when either the norm of $\tilde{v}$ is smaller than the machine epsilon or the number of iterations reaches 50000.

The Lasso model $\min_z \frac{\mu}{2}\|Az - y\|_2^2 + \|z\|_1$ for compressive sensing in [11] is used to test the empirical probability of successful recovery for noiseless problems. The number of nonzero values in z_* is k and the nonzero values are drawn from the standard normal distribution. The locations of the nonzero values are selected randomly. The vector y is set to be Az_*. The value of μ is $2000 + 1/\|A'y\|_\infty$, where

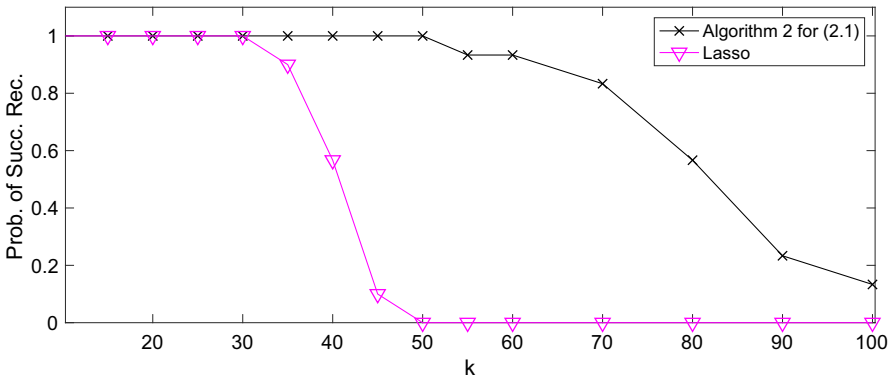

Fig. 2 Empirical probability of successful recovery for noiseless problems (from 30 random runs) versus k, where k is the number of input neurons of the generative model G for Algorithm 2 and k is the sparsity level for Lasso. In these experiments, the network has 2 layers, the middle layer has 250 neurons, and the output layer has 600 neurons, after which $m = 150$ random measurements are taken

$\|v\|_\infty = \max\{|v_i|\}$. A run is called successful if the relative error $\|z - z_*\|/\|z_*\|$ is smaller than 10^{-1}. We use the implementation in [11][1] for solving the Lasso model.

Figure 2 reports the empirical probability of successful recovery for noiseless problems. A run of Algorithm 2 is called successful if the relative error $\|x - x_*\|/\|x_*\|$ is smaller than 10^{-3}. We observe that Algorithm 2 is able to find the true code x_* when m is sufficiently large relative to the latent dimensionality k. This experiment shows that signal recovery by empirical risk optimization for compressive sensing under expansive-Gaussian generative priors succeeds in a much larger parameter range than that given by the theorem. In particular, the empirical dependence on d appears to be much milder in practice than what was assumed in the theorem. For comparison, we also plot the recovery error versus sparsity level k using Lasso. We observe that empirical risk optimization for compressive sensing with generative models is able to recover the true signals with smaller signal dimensionality than by solving Lasso under a sparsity prior.

Figure 3 shows graphs of the relative square errors versus the number of input neurons at different noise levels. The figure is consistent with the theoretical result in the sense that, fixing k, the relative error of the solution found by Algorithm 2 is proportional to the norm of the noise, formally stated in (3.2). Note that for noisy measurements, the relative square error decreases approximately linearly as the number k decreases. This result is better than what is predicted by the theorem because the theorem is proved in the case of arbitrary noise. In this case, the noise is random, and one expects superior performance for smaller values of k because only a fraction k/n of the noise energy projects onto the k-dimensional signal manifold in $\mathbb{R}^n$, i.e., $\|x_i - x_*\|^2$ is approximately proportional to k. We refer to [13] for more details in the case of random noise.

Figure 4 shows the relationship between the relative error $\|x_i - x_*\|/\|x_*\|$ and the number of iterations for four values of SNR. The number of input neurons is

[1] This implementation is available at https://www.caam.rice.edu/~optimization/L1/fpc/.

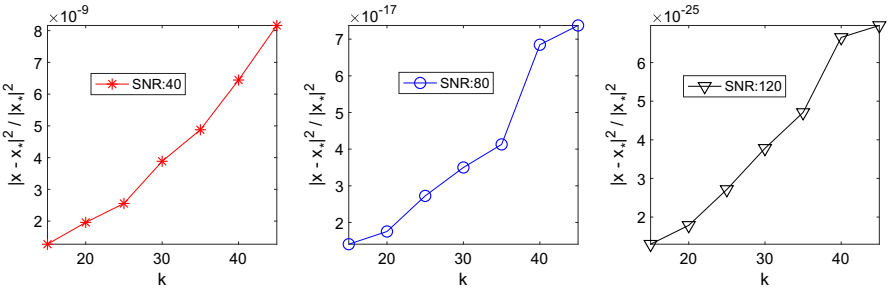

Fig. 3 The relative square error $\|x - x_*\|^2/\|x_*\|^2$ versus the number of input neurons k. The average of successful runs is reported

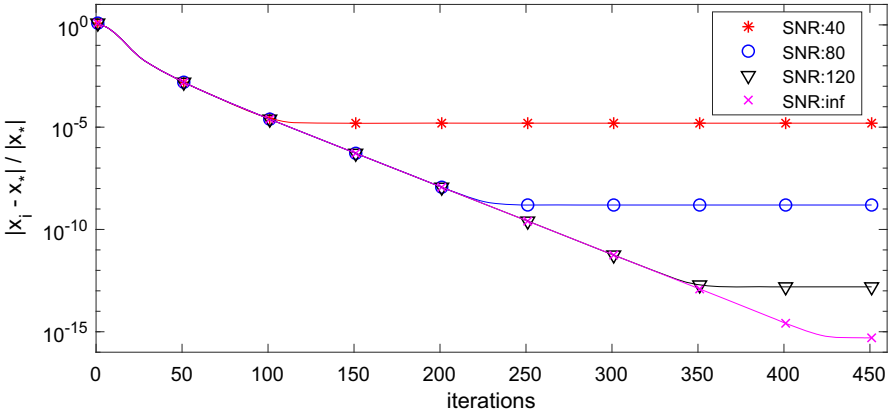

Fig. 4 The relative error $\|x - x_*\|/\|x_*\|$ versus the number of iterations. A typical result is reported

$k = 10$. Note that in the tests for the different values of SNR, all the other settings are identical, i.e., the initial iterate, latent code x_*, weights matrices A and W_i are the same. We observe that after approximately 20 iterations, Algorithm 2 converges linearly to a neighborhood of the true solution x_*, and the size of the neighborhood only depends on the magnitude of the noise. These results are consistent with Theorem 3.1. Additionally, this figure demonstrates that the relative error in the recovered latent code scales linearly with the magnitude of the noise, which is also consistent with the theorem.

5 Proof of Theorem 3.1

In this section, we prove our main result. Recall that the goal of Algorithm 1 is to minimize the cost function

$$f(x) = \frac{1}{2}\|AG(x) - y\|^2, \tag{5.1}$$

where $y = AG(x_*) + e$ and e is noise.

The proof relies on a concentration of measure argument which ensures that the cost function $f(x)$ and the step direction v_x concentrate around $f^E(x)$ and h_x, respectively, where $f^E(x)$ and h_x are defined later in Sect. 5.1. In particular, if the WDC and the RRIC hold with $\epsilon = 0$ and the noise e is 0, then $f(x) = f^E(x)$ and $v_x = h_x$. The idea of our convergence analysis is to prove properties of $f^E(x)$ and the direction h_x that are sufficient for a convergence analysis, if our method where to be run on $f^E(x)$ with step directions given by h_x, and then show that the actual cost function $f(x)$ and step direction are 'close enough' to establish convergence.

It is well known that if the gradient of a function is Lipschitz continuous, then a steepest descent method with a sufficient small step size converges to a stationary point from any starting point [25]. However, this result can not be used here since the gradient of the function (2.2) is not continuous. We overcome this technical difficulty by the following three steps, rigorously stated in Lemmas 5.1, 5.2, and 5.3, respectively.

1. The function h_x is Lipschitz continuous except in a ball around 0.
2. The (sub)-gradient of $f(x)$ is close to h_x.
3. The iterates generated by Algorithm 1 stay sufficiently far away from 0.

Those three steps are sufficient to show that the gradient of $f(x)$ is close to being Lipschitz continuous, and therefore the iterates from Algorithm 1 converge to a neighborhood of a *stationary point*. The size of the neighborhood depends on how close the (sub)-gradient of $f(x)$ is to h_x, and is controlled by the noise energy $\|e\|$ and the variable ϵ in the WDC and the RRIC. Of course, we also have to ensure that the algorithm not only converges to any of the three stationary points, but that it actually converges to a point close to x_*, for this we rely on the 'twist' of the algorithm in steps 1–3.

The remainder of the proof is organized as follows. We start by defining notation used throughout the proof (see Sect. 5.1). In Sect. 5.2 we introduced several technical results formalizing the steps 1–3 above, and in Sect. 5.3 we use those properties to formally prove Theorem 3.1.

5.1 Notation

Here, we define some useful quantities, in particular $f^E(x)$ and h_x, and introduce standard notation used throughout. We start with defining a function that is helpful for controlling how the operator $x \to W_{+,x}x$ distorts angles, and is defined as

$$g(\theta) = \cos^{-1}\left(\frac{1}{\pi}\left[(\pi - \theta)\cos\theta + \sin\theta\right]\right).$$

With this notation, define

$$h_{x,y} = \frac{1}{2^d}x - \tilde{h}_{x,y},$$

where

$$\tilde{h}_{x,y} = \frac{1}{2^d}\left(\prod_{i=0}^{d-1}\frac{\pi-\bar{\theta}_{i,x,y}}{\pi}\right)y + \frac{1}{2^d}\sum_{i=0}^{d-1}\frac{\sin\bar{\theta}_{i,x,y}}{\pi}\left(\prod_{j=i+1}^{d-1}\frac{\pi-\bar{\theta}_{j,x,y}}{\pi}\right)\|y\|\hat{x}.$$

Here, $\bar{\theta}_{0,x,y} = \angle(x, y)$ and $\bar{\theta}_{i,x,y} = g(\bar{\theta}_{i-1,x,y})$. Moreover, given a vector $z \in \mathbb{R}^t$, $\hat{z} = z/\|z\|$. For simplicity of notation, we use h_z and $\tilde{h}_z$ to denote h_{z,x_*} and $\tilde{h}_{z,x_*}$, respectively, i.e., we omit x_*. Next, define

$$f^E(x) = \frac{1}{2^{d+1}}x^Tx - x^T\tilde{h}_{x,x_*} + \frac{1}{2^{d+1}}x_*^Tx_*.$$

Moreover, let

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

where $\breve{\theta}_0 = \pi$, and $\breve{\theta}_i = g(\breve{\theta}_{i-1})$.

Next, define $\mathcal{B}(x, a) := \{y \in \mathbb{R}^k \mid \|y - x\| \le a\}$, let $u = v + cO_1(t)$ denote $\|u - v\| \le c|t|$, and $f(t) = O(g(t))$ denotes $\limsup_{t\to\infty}|f(t)|/|g(t)| < C$, where $C > 0$ is a constant. Moreover, $\|\cdot\|$ denotes the spectral norm.

Finally, define

$$\begin{aligned}
S_\epsilon &:= \{x \in \mathbb{R}^k \mid \|h_{x,x_*}\| \le \frac{1}{2^d}\epsilon\max(\|x\|, \|x_*\|)\},\\
S_\epsilon^+ &:= S_\epsilon \cap \mathcal{B}(x_*, 5000d^6\epsilon\|x_*\|), \text{ and}\\
S_\epsilon^- &:= S_\epsilon \cap \mathcal{B}(-\rho_d x_*, 500d^{11}\sqrt{\epsilon}\|x_*\|).
\end{aligned}$$

5.2 Preliminaries

In this section, we state formal results making steps 1–3 from the beginning of this section rigorous, and collect properties used later in the proof of Theorem 3.1.

We start by showing that the function h_x is Lipschitz continuous except in a ball around 0:

Lemma 5.1 *For all $x, y \neq 0$, it holds that*

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

In addition, if $x, y \notin \mathcal{B}(0, r\|x_*\|)$ for any $r > 0$, then $\|h_x - h_y\| \le \left(\frac{1}{2^d} + \frac{6d+4d^2}{\pi r 2^d}\right)\|x - y\|$.

The next lemma states that h_x and the sub-gradient v_x are close:

Lemma 5.2 *Suppose the WDC and RRIC hold with $\epsilon \leq 1/(16\pi d^2)^2$. Then for any $x \neq 0$ and any $v_x \in \partial f(x)$,*

$$\|v_x - h_x\| \leq a_1 \frac{d^3\sqrt{\epsilon}}{2^d} \max(\|x\|, \|x_*\|) + \frac{2}{2^{d/2}}\|e\|,$$

where a_1 is a universal constant.

Next, we ensure that after sufficiently many steps, the algorithm will be relatively far from the maximum around 0:

Lemma 5.3 *Suppose that WDC holds with $\epsilon < 1/(16\pi d^2)^2$ and $\|e\| \leq \frac{\|x_*\|}{8\pi 2^{d/2}}$. Moreover, suppose that the step size in Algorithm 1 satisfies $0 < \nu < \frac{a_2 2^d}{d^2}$, where a_2 is a universal constant. Then, after at most $T = (\frac{2^d}{2\nu})^2$ steps, we have that for all $i > T$ and for all $t \in [0, 1]$ that $t\tilde{x}_i + (1-t)x_{i+1} \notin \mathcal{B}(0, \frac{1}{64\pi}\|x_*\|)$.*

Recall that S_β is the set of points with the norm of h_x upper bounded by $\beta \max(\|x\|, \|x_*\|)$. The following lemma shows that this set is contained in balls around x_* and $\rho_d x_*$, thus outside those balls, the norm of h_x is lower bounded, which, together with Lemma 5.2, establishes that the sub-gradients are bounded away from zero. This in turn is important to show that outside those balls our gradient scheme makes progress.

Lemma 5.4 [13, Lemma 8] *For any $\beta \leq \frac{1}{64^2 d^{12}}$,*

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

Here, $\rho_d > 0$ obeys $\rho_d \to 1$ as $d \to \infty$.

It has been shown in [12] that the function $f^E(x)$ has three stationary points: one at $-\rho_d x_*$, one global minimizer at x_* and a local maximizer at 0. Therefore, Algorithm 1 could in principle be attracted to $-\rho_d x_*$. Lemma 5.5 guarantees that with the twist from Step 3 to Step 7, the iterates of Algorithm 1 converges to a neighborhood of x_*.

Lemma 5.5 *Suppose the WDC and RRIC hold with $\epsilon < 1/(16\pi d^2)^2$. Moreover, suppose the noise e satisfies $\|e\| \leq \frac{a_3\|x_*\|}{d^2 2^{d/2}}$, where a_3 is a universal constant. Then for any $\phi_d \in [\rho_d, 1]$, it holds that*

$$f(x) < f(y) \tag{5.2}$$

for all $x \in \mathcal{B}(\phi_d x_*, a_4 d^{-10}\|x_*\|)$ and $y \in \mathcal{B}(-\phi_d x_*, a_4 d^{-10}\|x_*\|)$, where $a_4 < 1$ is a universal constant.

Once an iterate is in a small neighborhood of x_*, Lemma 5.6 guarantees that the search directions of the iterates afterward point to x_* up to the noise e. Therefore, the iterates by Algorithm 1 converge to x_* up to the noise. In other words, the parameter ϵ in WDC and RRIC does not influence the size of the neighborhood that iterates converge to.

Lemma 5.6 *Suppose the WDC and RRIC hold with* $200d\sqrt{d\sqrt{\epsilon}} < 1$ *and* $x \in \mathcal{B}(x_*, d\sqrt{\epsilon}\|x_*\|)$. *Then for all* $x \neq 0$ *and for all* $v_x \in \partial f(x)$,

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

5.3 Proof of Theorem 3.1

The proof can be divided into three parts. We first show that the iterates $\{x_i\}$ converge to a neighborhoods of x_* and $-\rho_d x_*$, whose sizes depend on ϵ and the noise energy $\|e\|$. Second, we show that the iterates only converge to the neighborhood of x_* that depends both on ϵ as well as on the noise energy $\|e\|$. Lastly, we show that once an iterate is in the aforementioned neighborhood of x_*, the subsequent iterates converge to a neighborhood of x_* whose size only depends on the noise $\|e\|$, but not on ϵ.

1. *Convergence to a neighborhood of* x_* **or** $-\rho_d x_*$: We prove that if $\|h_{x_i}\|$ is sufficiently large, specifically if the iterate x_i is not in the set S_β with

$$\beta = 4a_1 d^3\sqrt{\epsilon} + 26\|e\|2^{d/2}/\|x_*\|,$$

then Algorithm 1 makes progress in the sense that $f(x_{i+1}) - f(x_i)$ is smaller than a certain negative value. Therefore, the iterates of Algorithm 1 converge to S_β.

Consider i such that $\tilde{x}_i \notin S_\beta$. Let $\eta_{\hat{x}_i} \in \partial f(\hat{x}_i)$ and define $\hat{x}_i = \tilde{x}_i - a\nu v_{\tilde{x}_i}$, where $a \in [0, 1]$. By the mean value theorem, for some $a \in [0, 1]$, we have

$$\begin{aligned} f(\tilde{x}_i - \nu v_{\tilde{x}_i}) - f(\tilde{x}_i) &= \langle \eta_{\hat{x}_i}, -\nu v_{\tilde{x}_i} \rangle \\ &= \langle \tilde{v}_{\tilde{x}_i}, -\nu v_{\tilde{x}_i} \rangle + \langle \eta_{\hat{x}_i} - v_{\tilde{x}_i}, -\nu v_{\tilde{x}_i} \rangle \\ &\leq -\nu\|v_{\tilde{x}_i}\|(\|v_{\tilde{x}_i}\| - \|\eta_{\hat{x}_i} - v_{\tilde{x}_i}\|). \end{aligned} \tag{5.3}$$

Next, we provide a lower and upper bound of the terms $\|v_{\tilde{x}_i}\|$ and $\|v_{\hat{x}_i} - v_{\tilde{x}_i}\|$, respectively, which appear on the right hand side of (5.3).

First, we have

$$\begin{aligned} \|v_{\tilde{x}_i}\| &\geq \|h_{\tilde{x}_i}\| - \|h_{\tilde{x}_i} - v_{\tilde{x}_i}\| \\ &\geq 2^{-d}\max(\|\tilde{x}_i\|, \|x_*\|)\left(\beta - a_1 d^3\sqrt{\epsilon} - 2\|e\|\frac{2^{d/2}}{\|x_*\|}\right) \\ &\geq 2^{-d}\max(\|\tilde{x}_i\|, \|x_*\|)\left(3a_1 d^3\sqrt{\epsilon} + 24\|e\|\frac{2^{d/2}}{\|x_*\|}\right) \qquad (5.4) \\ &\geq 2^{-d}\|x_*\|3a_1 d^3\sqrt{\epsilon}, \qquad (5.5) \end{aligned}$$

where the second inequality follows from the definition of S_β and Lemma 5.2, and the third inequality follows from the definition of β.

Second, by Lemmas 5.1 and 5.3, for all $a \in [0,1]$ and $i > T$ (T is defined in Lemma 5.3), we have

$$\|h_{\hat{x}_i} - h_{\tilde{x}_i}\| \le \frac{b_0 d^2}{2^d}\|\hat{x}_i - \tilde{x}_i\|, \tag{5.6}$$

where $\hat{x}_i = \tilde{x}_i - a\nu v_{\tilde{x}_i}$, and b_0 is a universal constant. Thus, for any $v_{\hat{x}_i} \in \partial f(\hat{x}_i)$,

$$\begin{aligned}
\|v_{\hat{x}_i} - v_{\tilde{x}_i}\| &\le \|v_{\hat{x}_i} - h_{\hat{x}_i}\| + \|h_{\hat{x}_i} - h_{\tilde{x}_i}\| + \|h_{\tilde{x}_i} - v_{\tilde{x}_i}\| \\
&\le a_1 \frac{d^3\sqrt{\epsilon}}{2^d}\max(\|\hat{x}_i\|, \|x_*\|) + \frac{2}{2^{d/2}}\|e\| + \frac{b_0 d^2}{2^d}\|\hat{x}_i - \tilde{x}_i\| \\
&\quad + a_1 \frac{d^3\sqrt{\epsilon}}{2^d}\max(\|\tilde{x}_i\|, \|x_*\|) + \frac{2}{2^{d/2}}\|e\| \\
&\le a_1 \frac{d^3\sqrt{\epsilon}}{2^d}\max(\|\tilde{x}_i\| + \nu\|v_{\tilde{x}_i}\|, \|x_*\|) + \frac{b_0 d^2}{2^d}\nu\|v_{\tilde{x}_i}\| \\
&\quad + a_1 \frac{d^3\sqrt{\epsilon}}{2^d}\max(\|\tilde{x}_i\|, \|x_*\|) + \frac{4}{2^{d/2}}\|e\| \\
&\le a_1 \frac{d^3\sqrt{\epsilon}}{2^d}\left(2 + \frac{\nu d a_6}{2^d}\right)\max(\|\tilde{x}_i\|, \|x_*\|) \\
&\quad + \frac{b_0 d^2}{2^d}\nu\|v_{\tilde{x}_i}\| + 4\frac{K_2/d^{42}}{2^d}\|x_*\|,
\end{aligned} \tag{5.7}$$

where the second inequality follows from Lemma 5.2 and (5.6), and the fourth inequality follows from Lemma A.1 and the assumption $\|e\| \le \frac{K_2\|x_*\|}{d^{42}2^{d/2}}$.

Combining (5.7) and (5.4), we get that

$$\|v_{\hat{x}_i} - v_{\tilde{x}_i}\| \le \left(\frac{5}{6} + \nu b_1 \frac{d^2}{2^d}\right)\|v_{\tilde{x}_i}\|,$$

with the appropriate constants chosen sufficiently small, where b_1 is a universal constant. Choosing $v_{\hat{x}_i} = \eta_{\hat{x}_i}$ yields

$$\|\eta_{\hat{x}_i} - v_{\tilde{x}_i}\| \le \left(\frac{5}{6} + \nu b_0 \frac{d^2}{2^d}\right)\|v_{\tilde{x}_i}\|. \tag{5.8}$$

Therefore, combining (5.3) and (5.8) yields

$$f(\tilde{x}_i - \nu v_{\tilde{x}_i}) - f(\tilde{x}_i) \le -\frac{1}{12}\nu\|v_{\tilde{x}_i}\|^2, \tag{5.9}$$

where we used that $\nu b_0 \frac{d^2}{2^d} \le 1/12$ by the assumption that the step size obeys $\nu = K_3 2^d/d^2$ and by taking K_3 appropriately small. Applying (5.5) to (5.9) yields

$$f(\tilde{x}_i - \nu v_{\tilde{x}_i}) - f(\tilde{x}_i) \le -\frac{1}{12}\nu \|v_{\tilde{x}_i}\|^2 \le -2^{-d} d^4 b_1 \epsilon \|x_*\|^2,$$

where b_1 is a universal constant and we used $\nu = K_3 \frac{2^d}{d^2}$. Therefore, there can be at most $\frac{f(x_0)2^d}{b_1 d^4 \epsilon \|x_*\|^2}$ iterations for which $\tilde{x}_i \notin S_\beta$. In other words, there exists $N \le \frac{f(x_0)2^d}{b_1 d^4 \epsilon \|x_*\|^2}$ such that $\tilde{x}_N \in S_\beta$.

2. *Convergence to a neighborhood of* x_*: Note that by the assumption $\|e\| \le \frac{K_2 \|x_*\|}{d^{42} 2^{d/2}}$ and $\epsilon \le K_1/d^{90}$, our choice of β obeys $\beta \le \frac{1}{64^2 d^{12}}$ for sufficiently small K_1, K_2, and thus the assumptions of Lemma 5.4 are met and we have

$$S_\beta \subset \mathcal{B}(x_*, r) \cup \mathcal{B}(-\rho_d x_*, \sqrt{r\|x_*\|} d^8). \tag{5.10}$$

Here, we defined the radius $r = K_5 d^9 \sqrt{\epsilon}\|x_*\| + K_6 d^6 \|e\| 2^{d/2}$, and K_5 and K_6 are universal constants and are used in (3.1).
By the assumption $\|e\| \le \frac{K_2\|x_*\|}{d^{42}2^{d/2}}$ and $\epsilon \le K_1/d^{90}$ and choosing K_1 and K_2 sufficiently small, we have $r \le a_4 d^{-10}\|x_*\|$ and $\sqrt{r\|x_*\|} d^8 \le a_4 d^{-10}\|x_*\|$. Note that the powers of d in the upper bounds of $\|e\|$ and ϵ, which are -42 and -90 respectively, are used to get $\sqrt{r\|x_*\|} d^8 \le a_4 d^{-10}\|x_*\|$. It follows from (5.10) that

$$S_\beta^+ \subset \mathcal{B}(x_*, a_4 d^{-10}\|x_*\|) \text{ and } S_\beta^- \subset \mathcal{B}(-\rho_d x_*, a_4 d^{-10}\|x_*\|).$$

Therefore, by Lemma 5.5, for any $x \in S_\beta^-$ and $y \in S_\beta^+$, it holds that $f(x) > f(y)$. Thus, if $\tilde{x}_N \in S_\beta$, then $\tilde{x}_N$ must be in S_β^+ due to the operations from Step 3 to Step 7.

We claim that if x_i is inside the ball $\mathcal{B}(x_*, r)$, then all iterates afterward stay in $\mathcal{B}(x_*, 2r)$. To see this, note that by Lemma A.1 and the choice of the step size, we have for any $v_{\tilde{x}_i} \in \partial f(\tilde{x}_i)$, $\nu\|v_{\tilde{x}_i}\| \le \frac{a_6}{d^2 2^d} \max(\|x\|, \|x_*\|)$.

3. *Convergence to* x_* *up to the noise* e: Next we show that for any $i \ge N$, it holds that $x_i \in \mathcal{B}(x_*, a_4 d^{-10}\|x_*\|)$, $\tilde{x}_i = x_i$, and

$$\|x_{i+1} - x_*\| \le b_2^{i+1-N}\|x_N - x_*\| + b_4 2^{d/2}\|e\|.$$

where a_4 is defined in Lemma 5.5, $b_2 = 1 - \frac{\nu}{2^d}\frac{7}{8}$ and b_4 is a universal constant.

Suppose $\tilde{x}_i \in \mathcal{B}(x_*, a_4 d^{-10}\|x_*\|)$. By the assumption $\epsilon \le K_1/d^{90}$ for sufficiently small K_1, the assumptions in Lemma 5.6 are met. Therefore,

$$\begin{aligned}\|x_{i+1} - x_*\| &= \|\tilde{x}_i - \nu v_{\tilde{x}_i} - x_*\| \\ &= \|\tilde{x}_i - x_* - \frac{\nu}{2^d}(\tilde{x}_i - x_*) - \nu v_{\tilde{x}_i} + \frac{\nu}{2^d}(\tilde{x}_i - x_*)\|\end{aligned}$$

$$\begin{aligned}
&\leq \left(1 - \frac{\nu}{2^d}\right) \|\tilde{x}_i - x_*\| + \nu \|v_{\tilde{x}_i} - \frac{1}{2^d}(\tilde{x}_i - x_*)\| \\
&\leq \left(1 - \frac{\nu}{2^d}\right) \|\tilde{x}_i - x_*\| + \nu \left(\frac{1}{8}\frac{1}{2^d}\|\tilde{x}_i - x_*\| + \frac{2}{2^{d/2}}\|e\|\right) \\
&= \left(1 - \frac{\nu}{2^d}\frac{7}{8}\right) \|\tilde{x}_i - x_*\| + \nu \frac{2}{2^{d/2}}\|e\|, \qquad (5.11)
\end{aligned}$$

where the second inequality holds by Lemma 5.6. By the assumptions $\tilde{x}_i \in \mathcal{B}(x_*, a_4 d^{-10}\|x_*\|)$, $\|e\| \leq \frac{K_2\|x_*\|}{d^{42}2^{d/2}}$, and using (5.11), we have $x_{i+1} \in \mathcal{B}(x_*, a_4 d^{-10}\|x_*\|)$. In addition, using Lemma 5.5 yields that $\tilde{x}_{i+1} = x_{i+1}$. Repeat the above steps yields that $x_i \in \mathcal{B}(x_*, a_4 d^{-10}\|x_*\|)$ and $\tilde{x}_i = x_i$ for all $i \geq N$.

Using (5.11) and $\nu = K_3 \frac{2^d}{d^2}$, we have

$$\|x_{i+1} - x_*\| \leq b_2 \|x_i - x_*\| + b_3 \frac{2^{d/2}}{d^2}\|e\|, \qquad (5.12)$$

where $b_2 = 1 - 7K_3/(8d^2)$ and b_3 is a universal constant. Repeatedly applying (5.12) yields

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

where the last inequality follows from the definition of b_2 and the step size $\nu = K_3 \frac{2^d}{d^2}$, and b_4 is a universal constant. This finishes the proof for (3.2). Inequality (3.3) follows from Lemma A.8.

This concludes the proof of our main result. In the remainder, we provide proofs of the lemmas above.

5.4 Proof of Lemma 5.1

It holds that

$$\|x - y\| \geq 2\sin(\theta_{x,y}/2)\min(\|x\|, \|y\|), \qquad \forall x, y \qquad (5.13)$$

$$\sin(\theta/2) \geq \theta/4, \qquad \forall \theta \in [0, \pi] \qquad (5.14)$$

$$\frac{d}{d\theta} g(\theta) \in [0, 1] \qquad \forall \theta \in [0, \pi] \qquad (5.15)$$

where $\theta_{x,y} = \angle(x, y)$.

For brevity of notation, let $\zeta_{j,z} = \prod_{i=j}^{d-1} \frac{\pi - \bar{\theta}_{i,z,x_*}}{\pi}$. Combining (5.13) and (5.14) gives $|\bar{\theta}_{0,x,x_*} - \bar{\theta}_{0,y,x_*}| \le 4 \max\left(\frac{1}{\|x\|}, \frac{1}{\|y\|}\right) \|x-y\|$. Inequality (5.15) implies $|\bar{\theta}_{i,x,x_*} - \bar{\theta}_{i,y,x_*}| \le |\bar{\theta}_{j,x,x_*} - \bar{\theta}_{j,y,x_*}|, \forall i \ge j$. It follows that

$$\begin{aligned}\|h_{x,x_*} - h_{y,x_*}\| \le \frac{1}{2^d}\|x-y\| + \frac{1}{2^d}\underbrace{\left|\zeta_{0,x} - \zeta_{0,y}\right|}_{T_1}\|x_*\| \\ + \frac{1}{2^d}\underbrace{\left|\sum_{i=0}^{d-1}\frac{\sin\bar{\theta}_{i,x,x_*}}{\pi}\zeta_{i+1,x}\hat{x} - \sum_{i=0}^{d-1}\frac{\sin\bar{\theta}_{i,y,x_*}}{\pi}\zeta_{i+1,y}\hat{y}\right|}_{T_2}\|x_*\|.\end{aligned} \tag{5.16}$$

We use the following result which is proven later in Lemma A.2:

$$T_1 \le \frac{d}{\pi}|\bar{\theta}_{0,x,x_*} - \bar{\theta}_{0,y,x_*}| \le \frac{4d}{\pi}\max\left(\frac{1}{\|x\|}, \frac{1}{\|y\|}\right)\|x-y\|. \tag{5.17}$$

Additionally, it holds that

$$\begin{aligned}T_2 &= \left|\sum_{i=0}^{d-1}\frac{\sin\bar{\theta}_{i,x,x_*}}{\pi}\zeta_{i+1,x}\hat{x} - \frac{\sin\bar{\theta}_{i,x,x_*}}{\pi}\zeta_{i+1,x}\hat{y} + \frac{\sin\bar{\theta}_{i,x,x_*}}{\pi}\zeta_{i+1,x}\hat{y}\right. \\ &\quad \left. - \sum_{i=0}^{d-1}\frac{\sin\bar{\theta}_{i,y,x_*}}{\pi}\zeta_{i+1,y}\hat{y}\right| \\ &\le \frac{d}{\pi}\|\hat{x}-\hat{y}\| + \underbrace{\left|\sum_{i=0}^{d-1}\frac{\sin\bar{\theta}_{i,x,x_*}}{\pi}\zeta_{i+1,x} - \sum_{i=0}^{d-1}\frac{\sin\bar{\theta}_{i,y,x_*}}{\pi}\zeta_{i+1,y}\right|}_{T_3}.\end{aligned} \tag{5.18}$$

We have

$$\begin{aligned}T_3 &\le \sum_{i=0}^{d-1}\left[\left|\frac{\sin\bar{\theta}_{i,x,x_*}}{\pi}\zeta_{i+1,x} - \frac{\sin\bar{\theta}_{i,x,x_*}}{\pi}\zeta_{i+1,y}\right|\right. \\ &\quad \left. + \left|\frac{\sin\bar{\theta}_{i,x,x_*}}{\pi}\zeta_{i+1,y} - \frac{\sin\bar{\theta}_{i,y,x_*}}{\pi}\zeta_{i+1,y}\right|\right] \\ &\le \sum_{i=0}^{d-1}\left[\frac{1}{\pi}\left(\frac{d-i-1}{\pi}\left|\bar{\theta}_{i-1,x,x_*} - \bar{\theta}_{i-1,y,x_*}\right|\right) + \frac{1}{\pi}|\sin\bar{\theta}_{i,x,x_*} - \sin\bar{\theta}_{i,y,x_*}|\right] \\ &\le \frac{d^2}{\pi}|\bar{\theta}_{0,x,x_*} - \bar{\theta}_{0,y,x_*}| \le \frac{4d^2}{\pi}\max\left(\frac{1}{\|x\|}, \frac{1}{\|y\|}\right)\|x-y\|.\end{aligned} \tag{5.19}$$

Using (5.13) and (5.14) and noting $\|\hat{x} - \hat{y}\| \leq \theta_{x,y,x_*}$ yield

$$\|\hat{x} - \hat{y}\| \leq \theta_{x,y,x_*} \leq 2 \max\left(\frac{1}{\|x\|}, \frac{1}{\|y\|}\right) \|x - y\|. \tag{5.20}$$

Finally, combining (5.16), (5.17), (5.18), (5.19) and (5.20) yields the result.

5.5 Proof of Lemma 5.2

Let $\bar{v}_x = \left(\prod_{i=d}^{1} W_{i,+,x}\right)^T A^T (A\left(\prod_{i=d}^{1} W_{i,+,x}\right) x - A\left(\prod_{i=d}^{1} W_{i,+,x_*}\right) x_*)$ and $q_x = \left(\prod_{i=d}^{1} W_{i,+,x}\right)^T A^T e$. Therefore, $\tilde{v}_x = \bar{v}_x - q_x$.

For any $x \neq 0$ and suppose $G(x)$ is differentiable at x, we have

$$\begin{aligned}
\|\tilde{v}_x - h_x\| &= \|\bar{v}_x + q_x - h_x\| \leq \|\bar{v}_x - h_x\| + \|q_x\| \\
&\leq b_0 \frac{d^3\sqrt{\epsilon}}{2^d} \max(\|x\|, \|x_*\|) + \|q_x\| \\
&\leq b_0 \frac{d^3\sqrt{\epsilon}}{2^d} \max(\|x\|, \|x_*\|) + \frac{2}{2^{d/2}} \|e\|,
\end{aligned} \tag{5.21}$$

where the second inequality follows from [12, (26)] and the third inequality follows from Lemma A.3 given later.

Since $f(x)$ is a piecewise quadratic function, by [6, Theorem 9.6], we have

$$\partial f(x) = conv(v_1, v_2, \ldots, v_t), \tag{5.22}$$

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. Therefore, for any $v \in \partial f(x)$, there exist $c_1, c_2, \ldots, c_t \geq 0$ such that $c_1 + c_2 + \ldots + c_t = 1$ and $v = c_1 v_1 + c_2 v_2 + \ldots + c_t v_t$. Note that for any v_i, there exists u_i so that $v_i = \lim_{\delta \to 0^+} \nabla f(x + \delta_i u_i)$, and f is differentiable at $f(x + \delta u_i)$ for sufficiently small δ.

The proof is concluded by appealing to the continuity of h_x with respect to nonzero x, inequality (5.21), and by noting that

$$\begin{aligned}
\|v_x - h_x\| &\leq \sum_i c_i \|v_i - h_x\| = \sum_i c_i \| \lim_{\delta_i \to 0^+} \nabla f(x + \delta_i u_i) - h_x\| \\
&= \sum_i c_i \lim_{\delta_i \to 0^+} \|\nabla f(x + \delta_i u_i) - h_{x+\delta_i u_i}\| \\
&= \sum_i c_i \lim_{\delta_i \to 0^+} \|\tilde{v}_{x+\delta_i u_i} - h_{x+\delta_i u_i}\| \\
&\leq b_0 \frac{d^3\sqrt{\epsilon}}{2^d} \max(\|x\|, \|x_*\|) + \frac{2}{2^{d/2}} \|e\|,
\end{aligned}$$

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

5.6 Proof of Lemma 5.3

First suppose that $\tilde{x}_i \in \mathcal{B}(0, \frac{1}{32\pi}\|x_*\|)$. We show that after a polynomial number of iterations N, we have that $\tilde{x}_{i+N} \notin \mathcal{B}(0, \frac{1}{32\pi}\|x_*\|)$. Below, we use that

$$\langle x, v_x\rangle < 0 \text{ and } \|v_x\| \geq \frac{1}{2^d 16\pi}\|x_*\| \text{ for all } x \in \mathcal{B}(0, \frac{1}{32\pi}\|x_*\|) \text{ and } v_x \in \partial f(x), \tag{5.23}$$

which will be proven later. It follows that for any $\tilde{x}_i \in \mathcal{B}(0, \frac{1}{32\pi}\|x_*\|)$, $\tilde{x}_i$ and the next iterate produced by the algorithm, $x_{i+1} = \tilde{x}_i - \nu v_{\tilde{x}_i}$, and the origin form an obtuse triangle. As a consequence,

$$\|\tilde{x}_{i+1}\|^2 = \|x_{i+1}\|^2 \geq \|\tilde{x}_i\|^2 + \nu^2\|v_{\tilde{x}_i}\|^2 \geq \|\tilde{x}_i\|^2 + \nu^2\frac{1}{(2^d 16\pi)^2}\|x_*\|^2, \tag{5.24}$$

where the last inequality follows from (5.23). Thus, the norm of the iterates $\tilde{x}_i$ will increase until after $\left(\frac{2^d}{2\nu}\right)^2$ iterations, we have $\tilde{x}_{i+N} \notin \mathcal{B}(0, \frac{1}{32\pi}\|x_*\|)$.

Consider $\tilde{x}_i \notin \mathcal{B}(0, \frac{1}{32\pi}\|x_*\|)$, and note that

$$\nu\|v_{\tilde{x}_i}\| \leq \nu\frac{da_6}{2^d}\max(\|\tilde{x}_i\|, \|x_*\|) \leq \nu\frac{32\pi a_6 d}{2^d}\|\tilde{x}_i\| \leq \frac{1}{2}\|\tilde{x}_i\|,$$

where the first inequality follows from Lemma A.1, the second inequality from $\|\tilde{x}_i\| \geq \frac{1}{32\pi}\|x_*\|$, and finally the last inequality from our assumption on the step size ν. Therefore, from $x_{i+1} = \tilde{x}_i - \nu v_{\tilde{x}_i}$, we have that $t\tilde{x}_i + (1-t)x_{i+1} \notin \mathcal{B}(0, \frac{1}{64\pi}\|x_*\|)$ for all $t \in [0, 1]$, which completes the proof.

It remains to prove (5.23). We start with proving $\langle x, \tilde{v}_x\rangle < 0$. For brevity of notation, let $\Lambda_z = \prod_{i=d}^{1} W_{i,+,z}$. We have

$$\begin{aligned}
x^T\tilde{v}_x &= \left\langle \Lambda_x^T A^T A\Lambda_x x - \Lambda_x^T A^T A\Lambda_{x_*}x_* + \Lambda_x^T A^T e, x\right\rangle \\
&\leq \left\langle \Lambda_x^T A^T A\Lambda_x x - \Lambda_x^T A^T A\Lambda_{x_*}x_* - \Lambda_x^T\Lambda_x x + \Lambda_x^T\Lambda_{x_*}x_*, x\right\rangle \\
&\quad + \left\langle \Lambda_x^T\Lambda_x x - \Lambda_x^T\Lambda_{x_*}x_* + \Lambda_x^T A^T e, x\right\rangle \\
&\leq \epsilon\|\Lambda_x x\|^2 + \epsilon\|\Lambda_x x\|\|\Lambda_{x_*}x_*\| + \left\langle \Lambda_x^T\Lambda_x x - \Lambda_x^T\Lambda_{x_*}x_* + \Lambda_x^T A^T e, x\right\rangle \\
&\leq \frac{13}{12}2^{-d}\|x\|^2 - \frac{1}{4\pi}\frac{1}{2^d}\|x\|\|x_*\| + \|x\|\frac{2}{2^{d/2}}\|e\| \\
&\leq \|x\|\left(\frac{13}{12}2^{-d}\|x\| + \frac{1/(8\pi)}{2^d}\|x_*\| - \frac{1}{4\pi}\frac{1}{2^d}\|x_*\|\right) \\
&\leq \|x\|\frac{1}{2^d}\left(2\|x\| - \frac{1}{8\pi}\|x_*\|\right) \leq -\frac{\|x\|}{16\pi 2^d}\|x_*\|.
\end{aligned}$$

The second inequality follows from RRIC, [12, (10)] that $\|\Lambda_x x\|^2 \le \frac{1+4\epsilon d}{2^d} \le \frac{13}{12}\frac{1}{2^d}$; the third inequality follows from [12, Lemma 8] that $\langle \Lambda_x x, \Lambda_{x_*} x_* \rangle \ge \frac{1}{4\pi}\frac{1}{2^d}\|x\|\|x_*\|$, and the fourth inequality follows from Lemma A.3.

If $G(x)$ is differentiable at x, then $v_x = \tilde{v}_x$ and $\langle x, v_x \rangle \le -\frac{\|x\|}{16\pi 2^d}\|x_*\| < 0$. If $G(x)$ is not differentiable at x, by Eq. (5.22), we have

$$x^T v_x = x^T(c_1 v_1 + c_2 v_2 + \cdots + c_t v_t) \le (c_1 + c_2 + \ldots + c_t)\left(-\frac{\|x\|}{16\pi 2^d}\|x_*\|\right)$$
$$= -\frac{\|x\|}{16\pi 2^d}\|x_*\| < 0,$$

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

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

5.7 Proof of Lemma 5.5

Consider the function

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

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

$$|\langle AG(x) - AG(x_*), e\rangle| \le \left|\left\langle A\prod_{i=d}^{1} W_{i,+,x}x, e\right\rangle\right| + \left|\left\langle A\prod_{i=d}^{1} W_{i,+,x_*}x_*, e\right\rangle\right|$$
$$= \left|\left\langle x, \left(\prod_{i=d}^{1} W_{i,+,x}\right)^T A^T e\right\rangle\right|$$
$$+ \left|\left\langle x_*, \left(\prod_{i=d}^{1} W_{i,+,x_*}\right)^T A^T e\right\rangle\right|$$
$$\le (\|x\| + \|x_*\|)\frac{2}{2^{d/2}}\|e\|$$
$$\le (\varphi\|x_*\| + \|x_*\|)\frac{2}{2^{d/2}}\|e\|,$$

where the second inequality holds by Lemma A.3, and the last inequality holds by our assumption on x. Thus, by Lemmas A.5 and A.6, we have

$$f_\eta(x) \le f_0^E(x) + |f_0(x) - f_0^E(x)| + |\langle AG(x) - AG(x_*), e\rangle|$$
$$\le \frac{1}{2^{d+1}}\left(\phi_d^2 - 2\phi_d + \frac{10}{a_8^3}d\varphi\right)\|x_*\|^2 + \frac{1}{2^{d+1}}\|x_*\|^2$$

$$+\frac{\epsilon(1+4\epsilon d)}{2^d}\|x\|^2+\frac{\epsilon(1+4\epsilon d)+48d^3\sqrt{\epsilon}}{2^{d+1}}\|x\|\|x_*\|+\frac{\epsilon(1+4\epsilon d)}{2^d}\|x_*\|^2$$
$$+(\varphi\|x_*\|+\|x_*\|)\frac{2}{2^{d/2}}\|e\|. \tag{5.25}$$

Additionally, for $x\in\mathcal{B}(\phi_d x_*,\varphi\|x_*\|)$, we have

$$\begin{aligned}(5.25)&\le\frac{1}{2^{d+1}}\left(\phi_d^2-2\phi_d+\frac{10}{a_8^3}d\varphi\right)\|x_*\|^2+\frac{1}{2^{d+1}}\|x_*\|^2\\&\quad+\frac{\epsilon(1+4\epsilon d)}{2^d}(\phi_d+\varphi)^2\|x_*\|^2+\frac{\epsilon(1+4\epsilon d)+48d^3\sqrt{\epsilon}}{2^{d+1}}(\phi_d+\varphi)\|x_*\|^2\\&\quad+\frac{\epsilon(1+4\epsilon d)}{2^d}\|x_*\|^2+(\varphi\|x_*\|+\|x_*\|)\frac{2}{2^{d/2}}\|e\|\\&\le\frac{\|x_*\|^2}{2^{d+1}}\left(1+\phi_d^2-2\phi_d+\frac{10}{a_8^3}d\epsilon+68d^2\sqrt{\epsilon}\right)+(\varphi\|x_*\|+\|x_*\|)\frac{2}{2^{d/2}}\|e\|.\end{aligned} \tag{5.26}$$

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

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

$$\begin{aligned}f_\eta(y)&\ge\mathbb{E}[f(y)]-|f(y)-\mathbb{E}[f(y)]|-|\langle AG(x)-AG(x_*),e\rangle|\\&\ge\frac{1}{2^{d+1}}\left(\phi_d^2-2\phi_d\rho_d-10d^3\varphi\right)\|x_*\|^2+\frac{1}{2^{d+1}}\|x_*\|^2\\&\quad-\left(\frac{\epsilon(1+4\epsilon d)}{2^d}\|y\|^2+\frac{\epsilon(1+4\epsilon d)+48d^3\sqrt{\epsilon}}{2^{d+1}}\|y\|\|x_*\|+\frac{\epsilon(1+4\epsilon d)}{2^d}\|x_*\|^2\right)\\&\quad-(\varphi\|x_*\|+\|x_*\|)\frac{2}{2^{d/2}}\|e\|\\&\ge\frac{\|x_*\|^2}{2^{d+1}}\left(1+\phi_d^2-2\phi_d\rho_d-10d^3\varphi-68d^2\sqrt{\epsilon}\right)-(\varphi\|x_*\|+\|x_*\|)\frac{2}{2^{d/2}}\|e\|.\end{aligned} \tag{5.27}$$

Using $\epsilon<\sqrt{\epsilon}$, $\rho_d\le 1$, $4\epsilon d<1$, $\varphi<1$, $\|e\|\le\frac{K_2\|x_*\|}{d^4 2^{d/2}}\le\frac{K_2\|x_*\|}{d^2 2^{d/2}}$ and assuming $\varphi=\epsilon$, the right side of (5.26) is smaller than the right side of (5.27) if

$$\varphi=\epsilon\le\left(\frac{(1-\rho_d)\phi_d-4K_2/d^2}{\left(125+\frac{5}{a_8^3}\right)d^3}\right)^2. \tag{5.28}$$

It follows from Lemma A.4 that $1-\rho_d\ge 1/(a_7(d+2)^2)$. Thus, it suffices to have $\varphi=\epsilon=\frac{a_4}{d^{10}}$ and $4K_2/d^2\le\frac{1}{2}\frac{1}{a_7(d+2)^2}\le 1-\rho_d$ for an appropriate universal constant K_2, and for an appropriate universal constant a_4.

5.8 Proof of Lemma 5.6

For brevity of notation, let $\Lambda_{j,z} = \prod_{i=j}^{1} W_{i,+,z}$. Suppose the function $G(x)$ is differentiable at x. Then the local linearity of G gives that $G(x+z) - G(x) = \Lambda_{j,x} z$ for any sufficiently small $z \in \mathbb{R}^k$. Using the RRIC, [12, (10)] and Lemma A.8, we have

$$\begin{aligned}
&|\langle A\Lambda_{j,x} z, A\Lambda_{j,x} x - A\Lambda_{j,x_*} x_*\rangle - \langle \Lambda_{j,x} z, \Lambda_{j,x} x - \Lambda_{j,x_*} x_*\rangle| \\
&\quad \le \epsilon \left\|\Lambda_{j,x} z\right\| \left\|\Lambda_{j,x} x - \Lambda_{j,x_*} x_*\right\| \le \epsilon \frac{1}{2^{\frac{d}{2}}} (1+2\epsilon d) \left\|\Lambda_{j,x} x - \Lambda_{j,x_*} x_*\right\| \|z\| \\
&\quad \le \epsilon \frac{1.2}{2^d} (1+2\epsilon d) \|x - x_*\| \|z\|.
\end{aligned}$$

Therefore, $\left\|\bar{v}_x - \Lambda_{j,x}^T (\Lambda_{j,x} x - \Lambda_{j,x_*} x_*)\right\| \le \epsilon \frac{1.2}{2^d}(1+2\epsilon d)\|x - x_*\| \le \frac{1}{16}\frac{1}{2^d}\|x - x_*\|$. Combining with Lemma A.9 yields that

$$\left\|\bar{v}_x - \frac{1}{2^d}(x - x_*)\right\| \le \frac{1}{2^d}\frac{1}{8}\|x - x_*\|.$$

It follows that

$$\|\tilde{v}_x - \frac{1}{2^d}(x - x_*)\| = \|\bar{v}_x + q_x - \frac{1}{2^d}(x - x_*)\| \le \frac{1}{2^d}\frac{1}{8}\|x - x_*\| + \frac{2}{2^{d/2}}\|e\|.$$

For any $x \neq 0$ and for any $v \in \partial f(x)$, by (5.22), there exist $c_1, c_2, \ldots, c_t \ge 0$ 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 $\|v - \frac{1}{2^d}(x - x_*)\| \le \sum_{j=1}^{t} c_j \|v_j - \frac{1}{2^d}(x - x_*)\| \le \frac{1}{2^d}\frac{1}{8}\|x - x_*\| + \frac{2}{2^{d/2}}\|e\|$.

Acknowledgements W.H. is partially supported by the Fundamental Research Funds for the Central Universities (No. 20720190060) and the National Natural Science Foundation of China (No. 12001455).P.H. is partially supported by NSF CAREER Award DMS-1848087 and NSF Award DMS-2022205. RH is partially supported by NSF Award IIS-1816986.

Appendix A: Supporting Lemmas

Lemma A.1 is used in proofs for Sect. 5.3 and Lemma 5.3.

Lemma A.1 *Suppose that the WDC and RRIC holds with $\epsilon < 1/(16\pi d^2)^2$ and that the noise e satisfies $\|e\| \le a_5 2^{-d/2}\|x_*\|$. Then, for all x and all $v_x \in \partial f(x)$,*

$$\|v_x\| \le \frac{a_6 d}{2^d} \max(\|x\|, \|x_*\|), \tag{A.1}$$

where a_5 and a_6 are universal constants.

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

$$\|v_x\| \le \|h_x\| + \|h_x - v_x\|$$

$$
\begin{aligned}
&\leq \left\| \frac{1}{2^d} x - \frac{1}{2^d} \zeta_0 x_* - \frac{1}{2^d} \sum_{i=0}^{d-1} \frac{\sin \bar{\theta}_{i,x}}{\pi} \zeta_{i+1} \frac{\|x_*\|}{\|x\|} x \right\| \\
&\quad + a_1 \frac{d^3 \sqrt{\epsilon}}{2^d} \max(\|x\|, \|x_*\|) + \frac{2}{2^{d/2}} \|e\| \\
&\leq \frac{1}{2^d} \|x\| + \left(\frac{1}{2^d} + \frac{d}{\pi 2^d} \right) \|x_*\| + a_1 \frac{d^3 \sqrt{\epsilon}}{2^d} \max(\|x\|, \|x_*\|) + \frac{2}{2^{d/2}} \|e\| \\
&\leq \frac{a_6 d}{2^d} \max(\|x\|, \|x_*\|),
\end{aligned}
$$

where the second inequality follows from the definition of h_x and Lemma 5.2, the third inequality uses $|\zeta_j| \leq 1$, and the last inequality uses the assumption $\|e\| \leq a_5 2^{-d/2} \|x_*\|$. □

Lemma A.2 is used in proofs for Lemma 5.1.

Lemma A.2 *Suppose $a_i, b_i \in [0, \pi]$ for $i = 1, \ldots, k$, and $|a_i - b_i| \leq |a_j - b_j|, \forall i \geq j$. Then it holds that*

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

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

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

□

Lemma A.3 is used in proofs for Lemmas 5.2, 5.3, and 5.5.

Lemma A.3 *Suppose the WDC and RRIC hold with $\epsilon \leq 1/(16\pi d^2)^2$. Then we have*

$$
\left| x^T q_x \right| \leq \frac{2}{2^{d/2}} \|e\| \|x\|,
$$

where $q_x = \left(\prod_{i=d}^{1} W_{i,+,x} \right)^T A^T e$. In addition, if x is differentiable at $G(x)$, then we have

$$
\|q_x\| \leq \frac{2}{2^{d/2}} \|e\|.
$$

Proof We have

$$|x^T q_x|^2 = |e^T AG(x)|^2 \le \|AG(x)\|^2\|e\|^2 \le (1+\epsilon)\|G(x)\|^2\|e\|^2$$
$$\le (1+\epsilon)\prod_{i=d}^{1}\|W_{i,+,x}\|^2\|e\|^2\|x\|^2 \le (1+\epsilon)(1+2\epsilon d)^2\frac{1}{2^d}\|e\|^2\|x\|^2,$$

where the second inequality follows from RRIC and the last inequality follows from [12, (10)]. Therefore, $\left|x^T q_x\right| \le \frac{2}{2^{d/2}}\|e\|\|x\|$.

Suppose G is differentiable at x. Then the local linearity of G implies that $G(x+z)-G(x) = \left(\prod_{i=d}^{1} W_{i,+,x}\right) z$ for any sufficiently small $z \in \mathbb{R}^k$. By the RRIC, we have

$$\left|\left\langle A\left(\prod_{i=d}^{1} W_{i,+,x}\right) z, A\left(\prod_{i=d}^{1} W_{i,+,x}\right) z\right\rangle - \left\langle \left(\prod_{i=d}^{1} W_{i,+,x}\right) z, \left(\prod_{i=d}^{1} W_{i,+,x}\right) z\right\rangle\right|$$
$$\le \epsilon\prod_{i=d}^{1}\|W_{i,+,x}\|^2\|z\|^2,$$

which implies

$$\left|\left\langle A\left(\prod_{i=d}^{1} W_{i,+,x}\right) z, A\left(\prod_{i=d}^{1} W_{i,+,x}\right) z\right\rangle\right| \le (1+\epsilon)\prod_{i=d}^{1}\|W_{i,+,x}\|^2\|z\|^2.$$

Therefore, we obtain

$$\left\|A\left(\prod_{i=d}^{1} W_{i,+,x}\right)\right\| \le \sqrt{1+\epsilon}\prod_{i=d}^{1}\|W_{i,+,x}\|.$$

Combining above inequality with $\prod_{i=d}^{1}\|W_{i,+,x}\| \le (1+2\epsilon d)/2^{d/2} \le 1.5/2^{d/2}$ given in [12, (10)] yields

$$\left\|A\left(\prod_{i=d}^{1} W_{i,+,x}\right)\right\| \le 1.5\sqrt{1+\epsilon}/2^{d/2} \le 2/2^{d/2},$$

where the second inequality follows from the assumption on ϵ. Therefore, we obtain

$$\|q_x\| = \left\|\left(\prod_{i=d}^{1} W_{i,+,x}\right)^T A^T e\right\| \le \left\|\left(\prod_{i=d}^{1} W_{i,+,x}\right)^T A^T\right\|\|e\| \le \frac{2}{2^{d/2}}\|e\|.$$

□

Lemma A.4 is used in proofs for Lemma 5.5.

Lemma A.4 *For all $d \geq 2$, that*

$$1/\left(a_7(d+2)^2\right) \leq 1-\rho_d \leq 250/(d+1),$$

and $a_8 = \min_{d\geq 2} \rho_d > 0$.

Proof It holds that

$$\log(1+x) \leq x \qquad \forall x \in [-0.5, 1] \tag{A.2}$$
$$\log(1-x) \geq -2x \qquad \forall x \in [0, 0.75] \tag{A.3}$$

where $\theta_{x,y} = \angle(x, y)$.

We recall the results in [12, (35), (36), and (49)]:

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

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

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

where the second and the fifth inequalities follow from (A.2) and (A.3) respectively. Since $\pi^3/(12(i+1)^3) \leq \check{\theta}_i^3/12 \leq \check{\theta}_i - \sin\check{\theta}_i \leq \check{\theta}_i^3/6 \leq 27\pi^3/(6(i+3)^3)$, we have that for all $d \geq 3$

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

and

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

where we use $\sum_{i=4}^{\infty} \frac{1}{i^2} \le \frac{\pi^2}{6}$ and $\sum_{i=1}^{n} i^3 = O(n^4)$. Since $\rho_d \ge 1 - 250/(d+1)$ and $\rho_d > 0$ for all $d \ge 2$, we have $\min_{d\ge 2} \rho_d > 0$. □

Lemma A.5 is used in proofs for Lemma 5.5.

Lemma A.5 *Fix* $0 < a_9 < \frac{1}{4d^2\pi}$. *For any* $\phi_d \in [\rho_d, 1]$, *it holds that*

$$f^E(x) < \frac{1}{2^{d+1}} \left(\phi_d^2 - 2\phi_d + \frac{10}{a_8^3} d a_9 \right) \|x_*\|^2 + \frac{\|x_*\|^2}{2^{d+1}}, \forall x \in \mathcal{B}(\phi_d x_*, a_9 \|x_*\|) \text{ and}$$

$$f^E(x) > \frac{1}{2^{d+1}} \left(\phi_d^2 - 2\phi_d \rho_d - 10 d^3 a_9 \right) \|x_*\|^2 + \frac{\|x_*\|^2}{2^{d+1}}, \forall x \in \mathcal{B}(-\phi_d x_*, a_9 \|x_*\|),$$

where a_8 *is defined in Lemma A.4.*

Proof If $x \in \mathcal{B}(\phi_d x_*, a_9\|x_*\|)$, then we have $0 \le \bar{\theta}_{0,x,x_*} \le \arcsin(a_9/\phi_d) \le \frac{\pi a_9}{2\phi_d}$, $0 \le \bar{\theta}_{0,x,x_*} \le \bar{\theta}_{i,x,x_*} \le \frac{\pi a_9}{2\phi_d}$, and $\phi_d\|x_*\| - a_9\|x_*\| \le \|x\| \le \phi_d\|x_*\| + a_9\|x_*\|$. Note that $\cos\theta \ge 1 - \frac{\theta^2}{2}, \forall \theta \in [0, \pi]$. We have

$$\begin{aligned}
f^E(x) - \frac{\|x_*\|^2}{2^{d+1}} &\le \frac{1}{2^{d+1}} \|x\|^2 - \frac{1}{2^d} \left(\prod_{i=0}^{d-1} \frac{\pi - \bar{\theta}_{i,x,x_*}}{\pi} \right) x_*^T x \\
&\le \frac{1}{2^{d+1}} (\phi_d + a_9)^2 \|x_*\|^2 - \frac{1}{2^d} \left(\prod_{i=0}^{d-1} \frac{\pi - \frac{\pi a_9}{2\phi_d}}{\pi} \right) \|x_*\| \|x\| \cos \bar{\theta}_{0,x,x_*} \\
&\le \frac{1}{2^{d+1}} (\phi_d + a_9)^2 \|x_*\|^2 - \frac{1}{2^d} \left(\prod_{i=0}^{d-1} \frac{\pi - \frac{\pi a_9}{2\phi_d}}{\pi} \right) (\phi_d - a_9) \|x_*\|^2 \left(1 - \frac{\pi^2 a_9^2}{8\phi_d^2} \right) \\
&\le \frac{1}{2^{d+1}} \left(\phi_d^2 + 2\phi_d a_9 + a_9^2 - 2 \left(1 - \frac{d a_9}{\phi_d} \right) (\phi_d - a_9) \left(1 - \frac{\pi^2 a_9^2}{8\phi_d^2} \right) \right) \|x_*\|^2 \\
&\le \frac{1}{2^{d+1}} \left(\phi_d^2 - 2\phi_d + \frac{10}{a_8^3} d a_9 \right) \|x_*\|^2,
\end{aligned}$$

where the last inequality is by Lemma A.4 and $a_9 < 1/(4\pi)$.

If $x \in \mathcal{B}(-\phi_d x_*, a_9\|x_*\|)$, then we have $0 \le \pi - \bar{\theta}_{0,x,x_*} \le \arcsin(a_9\pi) \le \frac{\pi^2}{2} a_9$, and $\phi_d\|x_*\| - a_9\|x_*\| \le \|x\| \le \phi_d\|x_*\| + a_9\|x_*\|$. It follows that

$$\begin{aligned}
f^E(x) - \frac{\|x_*\|^2}{2^{d+1}} &\ge \frac{1}{2^{d+1}} \|x\|^2 - \frac{1}{2^d} \sum_{i=0}^{d-1} \frac{\sin \bar{\theta}_{i,x,x_*}}{\pi} \left(\prod_{j=i+1}^{d-1} \frac{\pi - \bar{\theta}_{j,x,x_*}}{\pi} \right) \|x_*\| \|x\| \\
&\ge \frac{1}{2^{d+1}} \|x\|^2 - \frac{1}{2^d} \left(\rho_d + \frac{3d^3 a_9 \pi^2}{2} \right) \|x_*\| \|x\| \quad \text{(by [12, (40)])} \\
&\ge \frac{1}{2^{d+1}} (\phi_d - a_9)^2 \|x_*\|^2 - \frac{1}{2^d} \left(\rho_d + \frac{3d^3 a_9 \pi^2}{2} \right) (\phi_d + a_9) \|x_*\|^2
\end{aligned}$$

$$\geq \frac{1}{2^{d+1}}\left(\phi_d^2 - 2\phi_d\rho_d - 10d^3a_9\right)\|x_*\|^2.$$

□

Lemma A.6 is used in proofs for Lemma 5.5.

Lemma A.6 *If the WDC and RRIC hold with $\epsilon < 1/(16\pi d^2)^2$, then we have*

$$|f(x) - f^E(x)| \leq \frac{\epsilon(1+4\epsilon d)}{2^d}\|x\|^2 + \frac{\epsilon(1+4\epsilon d) + 48d^3\sqrt{\epsilon}}{2^{d+1}}\|x\|\|x_*\| + \frac{\epsilon(1+4\epsilon d)}{2^d}\|x_*\|^2.$$

Proof For brevity of notation, let $\Lambda_z = \prod_{i=d}^{1} W_{i,+,z}$. We have

$$\begin{aligned}
\left|f(x) - f^E(x)\right| &= \left|\frac{1}{2}x^T\left(\Lambda_x^T A^T A\Lambda_x - \Lambda_x^T\Lambda_x\right)x + \frac{1}{2}x^T\left(\Lambda_x^T\Lambda_x - \frac{I_k}{2^d}\right)x\right. \\
&\quad - x^T\left(\Lambda_x^T A^T A\Lambda_{x_*}x_* - \Lambda_x^T\Lambda_{x_*}x_*\right) - x^T\left(\Lambda_x^T\Lambda_{x_*}x_* - h_{x,x_*}\right) \\
&\quad + \frac{1}{2}x_*^T\left(\Lambda_{x_*}^T A^T A\Lambda_{x_*} - \Lambda_{x_*}^T\Lambda_{x_*}\right)x_* \\
&\quad \left. + \frac{1}{2}x_*^T\left(\Lambda_{x_*}^T\Lambda_{x_*} - \frac{I_k}{2^d}\right)x_*\right| \\
&\leq \frac{\epsilon}{2}\frac{1+4\epsilon d}{2^d}\|x\|^2 + \frac{\epsilon}{2}\frac{1+4\epsilon d}{2^d}\|x\|^2 + \frac{\epsilon}{2}\frac{1+4\epsilon d}{2^d}\|x\|\|x_*\| \\
&\quad + \frac{24d^3\sqrt{\epsilon}}{2^d}\|x\|\|x_*\| \\
&\quad + \frac{\epsilon}{2}\frac{1+4\epsilon d}{2^d}\|x_*\|^2 + \frac{\epsilon}{2}\frac{1+4\epsilon d}{2^d}\|x_*\|^2 \\
&= \frac{\epsilon(1+4\epsilon d)}{2^d}\|x\|^2 \\
&\quad + \frac{\epsilon(1+4\epsilon d) + 48d^3\sqrt{\epsilon}}{2^{d+1}}\|x\|\|x_*\| + \frac{\epsilon(1+4\epsilon d)}{2^d}\|x_*\|^2,
\end{aligned}$$

where the first inequality uses the WDC, the RRIC, and [12, Lemma 8]. □

Lemma A.7 is used in proofs for Lemma A.8.

Lemma A.7 *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*

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

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

Proof We have

$$\begin{aligned}\|W_{+,x}x - W_{+,y}y\| &\le \|W_{+,x}x - W_{+,x}y\| + \|W_{+,x}y - W_{+,y}y\| \\ &= \|W_{+,x}(x-y)\| + \|(W_{+,x} - W_{+,y})y\| \\ &\le \|W_{+,x}\|\|x-y\| + \|(W_{+,x} - W_{+,y})y\|. \end{aligned} \tag{A.4}$$

By WDC assumption, we have

$$\begin{aligned}\|W_{+,x}^T(W_{+,x} - W_{+,y})\| \le \left\|W_{+,x}^T W_{+,x} - I/2\right\| + \left\|W_{+,x}^T W_{+,y} - Q_{x,y}\right\| \\ + \left\|Q_{x,y} - I/2\right\| \le 2\epsilon + \theta. \end{aligned} \tag{A.5}$$

We also have

$$\begin{aligned}&\|(W_{+,x} - W_{+,y})y\|^2 \\ &= \sum_{i=1}^n (1_{w_i\cdot x>0} - 1_{w_i\cdot y>0})^2 (w_i\cdot y)^2 \\ &\le \sum_{i=1}^n (1_{w_i\cdot x>0} - 1_{w_i\cdot y>0})^2 ((w_i\cdot x)^2 + (w_i\cdot y)^2 - 2(w_i\cdot x)(w_i\cdot y)) \\ &= \sum_{i=1}^n (1_{w_i\cdot x>0} - 1_{w_i\cdot y>0})^2 (w_i\cdot (x-y))^2 \\ &= \sum_{i=1}^n 1_{w_i\cdot x>0} 1_{w_i\cdot y\le 0} (w_i\cdot (x-y))^2 + \sum_{i=1}^n 1_{w_i\cdot x\le 0} 1_{w_i\cdot y>0} (w_i\cdot (x-y))^2 \\ &= (x-y)^T W_{+,x}^T (W_{+,x} - W_{+,y})(x-y) + (x-y)^T W_{+,y}^T (W_{+,y} - W_{+,x})(x-y) \\ &\le 2(2\epsilon + \theta)\|x-y\|^2. \quad \text{(by (A.5))} \end{aligned} \tag{A.6}$$

Combining (A.4), (A.6), and $\|W_{i,+,x}\|^2 \le 1/2 + \epsilon$ given in [12, (9)] yields the result. □

Lemma A.8 is used in proofs for Lemma 5.6 and Lemma A.9.

Lemma A.8 *Suppose* $x \in \mathcal{B}(x_*, d\sqrt{\epsilon}\|x_*\|)$, *and the WDC holds with* $\epsilon < 1/(200)^4/d^6$. *Then it holds that*

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

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 $x \in \mathcal{B}(x_*, d\sqrt{\epsilon}\|x_*\|)$, we have

$$\bar{\theta}_i \le \bar{\theta}_0 \le 2d\sqrt{\epsilon}. \tag{A.7}$$

By [12, (14)], we also have $|\theta_i - \bar{\theta}_i| \le 4i\sqrt{\epsilon} \le 4d\sqrt{\epsilon}$. It follows that

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

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

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

where the second inequality is from that $(1+x)^d \le 1+2dx$ if $0 < xd < 1$. Combining the above inequality with Lemma A.7 yields

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

□

Lemma A.9 is used in proofs for Lemma 5.6.

Lemma A.9 *Suppose* $x \in \mathcal{B}(x_*, d\sqrt{\epsilon}\|x_*\|)$, *and the WDC holds with* $\epsilon < 1/(200)^4/d^6$. *Then it holds that*

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

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

$$\begin{aligned} &\Lambda_{d,1,x}^T \left(\Lambda_{d,1,x} x - \Lambda_{d,1,x_*} x_*\right) \\ &= \Lambda_{d,1,x}^T \left[\Lambda_{d,1,x} x - \sum_{j=1}^{d} \left(\Lambda_{d,j,x} \Lambda_{j-1,1,x_*} x_*\right)\right. \\ &\quad \left. + \sum_{j=1}^{d} \left(\Lambda_{d,j,x} \Lambda_{j-1,1,x_*} x_*\right) - \Lambda_{d,1,x_*} x_*\right] \end{aligned}$$

$$= \underbrace{\Lambda_{d,1,x}^T \Lambda_{d,1,x}(x - x_*)}_{T_1} + \underbrace{\Lambda_{d,1,x}^T \sum_{j=1}^{d} \Lambda_{d,j+1,x} \left(W_{j,+,x} - W_{j,+,x_*}\right) \Lambda_{j-1,1,x_*} x_*}_{T_2}. \tag{A.9}$$

For T_1, we have

$$T_1 = \frac{1}{2^d}(x - x_*) + \frac{4d}{2^d}\|x - x_*\| O_1(\epsilon). \quad ([12, (10)]) \tag{A.10}$$

For T_2, we have

$$\begin{aligned} T_2 &= O_1(1) \sum_{j=1}^{d} \left(\frac{1}{2^{d-\frac{j}{2}}} + \frac{(4d-2j)\epsilon}{2^{d-\frac{j}{2}}} \right) \left\| (W_{j,+,x} - W_{j,+,x_*}) \Lambda_{j-1,1,x_*} x_* \right\| \\ &= O_1(1) \sum_{j=1}^{d} \left(\frac{1}{2^{d-\frac{j}{2}}} + \frac{(4d-2j)\epsilon}{2^{d-\frac{j}{2}}} \right) \left\| (\Lambda_{j-1,1,x} x - \Lambda_{j-1,1,x_*} x_*) \right\| \\ &\quad \sqrt{2(\theta_{i,x,x_*} + 2\epsilon)} \\ &= O_1(1) \sum_{j=1}^{d} \left(\frac{1}{2^{d-\frac{j}{2}}} + \frac{(4d-2j)\epsilon}{2^{d-\frac{j}{2}}} \right) \frac{1.2}{2^{\frac{j}{2}}} \|x - x_*\| \frac{1}{30\sqrt{2d}} \\ &= \frac{1}{16} \frac{1}{2^d} \|x - x_*\| O_1(1), \end{aligned} \tag{A.11}$$

where the first equation is by [12, (10)]; the second equation is by (A.6); the third equation is by Lemma A.8 and (A.8). The result follows from (A.9), (A.10) and (A.11). □

References

1. Allen-Zhu, Z., Li, Y., Song, Z.: A convergence theory for deep learning via over-parameterization. In: Proceedings of the 36th International Conference on Machine Learning, vol. 97, pp. 242–252. PMLR, 09–15 (2019)
2. Arora, S., Liang, Y., Ma, T.: Why are deep nets reversible: a simple theory, with implications for training. Preprint (2015). arXiv:1511.05653
3. Blanchard, J.D., Cartis, C., Tanner, J.: Compressed sensing: How sharp is the restricted isometry property? SIAM Rev. Soc. Ind. Appl. Math. **53**(1), 105–125 (2011)
4. Bora, A., Jalal, A., Price, E., Dimakis, A.G.: Compressed sensing using generative models. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 537–546. PMLR, 06–11 (2017)
5. Candes, E.J., Tao, T.: Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theory **52**(12), 5406–5425 (2006)
6. Clason, C.: Nonsmooth analysis and optimization. Preprint (2017). arXiv:1708.04180
7. Du, S.S., Zhai, X., Poczos, B., Singh, A.: Gradient descent provably optimizes over-parameterized neural network. In: Proceedings of the 7nd International Conference on Learning Representations (2019)
8. Eldar, Y.C., Kutyniok, G.: Compressed Sensing: Theory and Applications. Cambridge University Press, Cambridge (2012)
9. Foucart, S., Rauhut, H.: A Mathematical Introduction to Compressive Sensing. Birkhäuser/Springer, Boston (2013)

10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
11. Hale, E.T., Yin, W., Zhang, Y.: Fixed-point continuation for ℓ_1-minimization: methodology and convergence. SIAM J. Optim. **19**(3), 1107–1130 (2008)
12. Hand, P., Voroninski, V.: Global guarantees for enforcing deep generative priors by empirical risk. IEEE Trans. Inf. Theory **66**(1), 401–418 (2019)
13. Heckel, R., Huang, W., Hand, P., Voroninski, V.: Deep denoising: rate-optimal recovery of structured signals with a deep prior. Inf. Inference (2020. accepted)
14. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision, pp. 694–711. Springer, Cham (2016)
15. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation (2018). arXiv:1710.10196
16. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1×1 convolutions. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 10236–10245 (2018)
17. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: Proceedings of the 2nd International Conference on Learning Representations (2014)
18. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690 (2017)
19. Li, Y., Liang, Y.: Learning overparameterized neural networks via stochastic gradient descent on structured data. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 8168–8177 (2018)
20. Mao, X.-J., Shen, C., Yang, Y.-B.: Image restoration using convolutional auto-encoders with symmetric skip connections. Preprint (2016). arXiv:1606.08921
21. Mardani, M., Gong, E., Cheng, J.Y., Vasanawala, S.S., Zaharchuk, G., Xing, L., Pauly, J.M.: Deep generative adversarial neural networks for compressive sensing MRI. IEEE Trans. Med. Imaging **38**(1), 167–179 (2019)
22. Mardani, M., Monajemi, H., Papyan, V., Vasanawala, S., Donoho, D., Pauly, J.: Recurrent generative adversarial networks for proximal learning and automated compressive image recovery. Preprint (2017). arXiv:1711.10046
23. Mousavi, A., Baraniuk, R.G.: Learning to invert: Signal recovery via deep convolutional networks. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2272–2276 (2017)
24. Mousavi, A., Patel, A.B., Baraniuk, R.G.: A deep learning approach to structured signal recovery. In: 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 1336–1343 (2015)
25. Nocedal, J., Wright, S.J.: Numerical Optimization, 2nd edn. Springer, Cham (2006)
26. Oymak, S., Soltanolkotabi, M.: Toward moderate overparameterization: global convergence guarantees for training shallow neural networks. IEEE J. Sel. Areas Inf. Theory **1**(1), 84–105 (2020)
27. Rippel, O., Bourdev, L.: Real-time adaptive image compression. In: International Conference on Machine Learning, pp. 2922–2930. PMLR (2017)
28. Sønderby, C.K., Caballero, J., Theis, L., Shi, W., Huszár, F.: Amortised MAP inference for image super-resolution. In: Proceedings of the 5th International Conference on Learning Representations (2017)
29. Van Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: International Conference on Machine Learning, pp. 1747–1756. PMLR (2016)
30. Yeh, R.A., Chen, C., Lim, T.Y., Schwing, A.G., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with deep generative models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5485–5493 (2017)
31. Zhu, J.-Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: European Conference on Computer Vision, pp. 597–613. Springer, Cham (2016)