Compressive phase retrieval: Optimal sample complexity with deep generative priors

Paul Hand¹ | Oscar Leong² | Vladislav Voroninski³

¹Department of Mathematics and Khoury College of Computer Sciences, Northeastern University, Boston, Massachusetts, USA
²Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California, USA
³Helm.ai, Menlo Park, California, USA

Abstract
Advances in compressive sensing (CS) provided reconstruction algorithms of sparse signals from linear measurements with optimal sample complexity, but natural extensions of this methodology to nonlinear inverse problems have been met with potentially fundamental sample complexity bottlenecks. In particular, tractable algorithms for compressive phase retrieval with sparsity priors have not been able to achieve optimal sample complexity. This has created an open problem in compressive phase retrieval: under generic, phaseless linear measurements, are there tractable reconstruction algorithms that succeed with optimal sample complexity? Meanwhile, progress in machine learning has led to the development of new data-driven signal priors in the form of generative models, which can outperform sparsity priors with significantly fewer measurements. In this work, we resolve the open problem in compressive phase retrieval and demonstrate that generative priors can lead to a fundamental advance by permitting optimal sample complexity by a tractable algorithm. We additionally provide empirics showing that exploiting generative priors in phase retrieval can significantly outperform sparsity priors. These results provide support for generative priors as a new paradigm for signal recovery in a variety of contexts, both empirically and theoretically. The strengths of this paradigm are that (1) generative priors can represent some classes of natural signals more concisely than sparsity priors, (2) generative
priors allow for direct optimization over the natural signal manifold, which is intractable under sparsity priors, and (3) the resulting non-convex optimization problems with generative priors can admit benign optimization landscapes at optimal sample complexity, perhaps surprisingly, even in cases of nonlinear measurements.

1 | INTRODUCTION

The study of inverse problems pervades virtually all of the natural sciences including biological and astronomical imaging, x-ray crystallography, oil exploration, and shape optimization and reconstruction. An object of interest is observed via some forward mapping process, and the task is to recover the object, often subject to ill-posedness and noise. In order to increase fidelity of the estimate or decrease the number of required measurements, one can enforce structural assumptions or priors on the signal, a practice dating as far back as Tikhonov regularization [71] and the Nyquist sampling theorem [57]. A canonical example of an ill-posed inverse problem in the field of imaging is compressive sensing (CS), in which one aims to recover a signal from undersampled linear measurements. By exploiting the sparsity of natural images in the wavelet domain as a structural prior, CS has led to a number of practical developments across the imaging sciences, such as speeding up some forms of MRI imaging by an order of magnitude [25].

In terms of theory, advances in CS have provided reconstruction algorithms using sparsity priors with information theoretically optimal sample complexity [17, 24]. A seminal result in the field states that if given \( m < n \) undersampled linear measurements \( b^* = Ay^* \) where \( A \in \mathbb{R}^{m \times n} \) has independent and identically distributed (i.i.d.) Gaussian entries and \( y^* \in \mathbb{R}^n \) is an \( s \)-sparse signal, recovery is guaranteed with high probability when \( m = O(s \log n) \) by solving the following convex program:

\[
\min_{y \in \mathbb{R}^n} \|y\|_1 \quad \text{s.t.} \quad Ay = b^*.
\]

The success of CS has popularized the notion of signal sparsity throughout the imaging sciences, resulting in sparsity becoming a common choice as a structural prior.

Sparsity-based priors when applied to nonlinear inverse problems such as phase retrieval have been met with potentially fundamental sample complexity bottlenecks. In phase retrieval, a signal \( y^* \in \mathbb{R}^n \) or \( \mathbb{C}^n \) is to be estimated from observations \( |\langle a_i, y^* \rangle|^2 \), \( i = 1, 2, \ldots, m \). Compressive phase retrieval considers the case \( m < n \), which requires structural priors to enable recovery. While an \( s \)-sparse signal is information theoretically recoverable from \( O(s \log n) \) generic phaseless measurements, compressive phase retrieval algorithms have not achieved sample complexity below \( O(s^2 \log n) \) [43]. In fact, convex algorithms such as PhaseLift [15, 18], provably fail below \( O(s^2 \log n) \) measurements under natural extensions to incorporate sparsity [52, 59]. While signal recovery is possible from \( O(s \log n) \) measurements under certain non-generic measurement designs (see Section 1.1), this gap continues to hold for sparsity priors under generic, phaseless linear measurements. This has created an open problem in compressive phase retrieval to find a computationally efficient algorithm to reconstruct signals from generic, phaseless linear measurements with optimal sample complexity with respect to the
signal’s intrinsic dimensionality. Furthermore, there is evidence to support that these sample complexity limitations may be fundamental for sparse phase retrieval. In the closely related sparse Principal Component Analysis (PCA) problem, a reduction from planted clique was found, and it is widely conjectured to be NP-hard [9, 12]. These observations open the question of whether other signal priors may successfully achieve sample-optimal reconstruction algorithms.

Simultaneously, there has been tremendous progress on priors in the form of generative models given by a deep neural network, which in some cases significantly outperforms sparsity priors at CS. These generative models, such as Generative Adversarial Networks [33] and Variational Autoencoders (VAEs) [49], learn an explicit mapping from a low-dimensional latent space $\mathbb{R}^k$ to an approximation of the natural image manifold in $\mathbb{R}^n$ and can be trained on datasets of various natural signal classes to create realistic, yet synthetic samples of human faces [47], MRIs [66], cells [58], human fingerprints [27], and more. Enforcing a generative prior in CS tasks by directly optimizing over the latent space has been shown to outperform sparsity-based methods such as Lasso by 5–10× fewer measurements [13] in some cases. Moreover, while the optimization problem posed over latent space is non-convex, [38] showed that when the number of measurements $m$ is proportional to $k$ up to log factors, the empirical risk minimization problem under a suitable random generator model exhibits favorable global geometry in the sense that there are no spurious local minima away from small neighborhoods of the true solution and a negative multiple thereof.

The above empirical and theoretical evidence indicates that generative neural networks can potentially succeed as structural priors in nonlinear inverse problems where previous methods exploiting sparsity have thus far been met with likely fundamental bottlenecks.

In this work, we resolve the open problem in compressive phase retrieval by presenting a computationally efficient algorithm that achieves optimal sample complexity with generic Gaussian measurements under a generative prior. In particular, we consider a deep generative prior for compressive phase retrieval by supposing that the desired signal lives in the range of a feed-forward neural network with Rectified Linear Unit (ReLU) activation functions and latent code dimensionality $k$. We establish the sufficiency of two deterministic conditions on the weights of the generative model and the measurement matrix to guarantee that the signal can be recovered by a subgradient descent algorithm. Moreover, we show that these conditions are satisfied with high probability for Gaussian weights and generic Gaussian measurements as soon as $m$ is proportional to $k$, up to log factors, which is information theoretically optimal in $k$. The present work is a significant extension of prior work by the authors [36], which provided encouraging evidence for the existence of a provably convergent subgradient descent scheme via a global landscape analysis. In addition to our theoretical results, we empirically establish that exploiting generative models in phase retrieval tasks can significantly outperform sparsity-based methods.

Subsequent to the publication of preliminary versions of the results of this paper [36], generative priors have also been shown to break through sample complexity barriers in PCA. In particular, all known algorithms to achieve optimal statistical sample complexity in sparse PCA are computationally intractable and all known polynomial time algorithms exhibit a sub-optimal quadratic sample complexity on the sparsity of the true signal [23, 50]. Gaps of this nature have also been observed in a number of related problems [22, 63]. However, with respect to PCA, recent work in both the asymptotic [5] and non-asymptotic regimes [20] have shown that the low rank matrix recovery problem with generative priors does not exhibit a computational-to-statistical gap. Along these lines, subsequent work in the asymptotic regime for compressive phase retrieval
under a generative prior [4] demonstrated that the problem becomes tractable in the information-theoretic limit. These results offer further evidence of the benefit of generative priors in inverse problems.

The results in the present work provide empirical and theoretical support to the notion that deep generative priors offer a new paradigm for signal recovery that offers fundamental advances. In this paradigm, a model of a natural signal class is learned from data in the form of a generative model. The generative model directly parameterizes a low-dimensional signal manifold, and recovering a signal subject to noisy measurements can be posed as a direct optimization problem whose search space is restricted to the range of the generative model. This paradigm has several strengths in comparison to sparsity priors. First, generative models may provide better compression of natural signals than sparsity priors. Precisely, the dimensionality of the manifold modeling the natural signal class under a generative prior may be lower than the sparsity level of the same signals. Second, generative priors allow for direct optimization over the natural signal manifold. In contrast, sparsity priors give rise to combinatorial optimization problems which can not be directly solved. Tractable convex relaxations have not been successful in important nonlinear settings. Third, the non-convex optimization problems under generative priors can admit benign optimization landscapes at optimal sample complexity even in the case of nonlinear measurements. This fundamental advance has so far not been realized by sparsity priors.

1.1 Related work

Phase retrieval
Some of the earliest methods to solve phase retrieval tasks are the non-convex alternating minimization Gerchberg-Saxton [30] and Fienup [29] algorithms. Recently, a variety of methods have been introduced that enjoy theoretical guarantees. Convex methods, such as the seminal lifting-based approach PhaseLift [18], can achieve optimal sample complexity for unstructured signals [15]. Further recovery guarantees have been extended to non-convex formulations such as Wirtinger Flow [16, 67, 69] and its non-smooth variant Amplitude Flow [26, 67, 81]. Other approaches include those based on linear programming [7, 31], Phasecut [76], AltMinPhase [45], and alternating projection methods [75].

Since the success of exploiting sparsity in linear compressed sensing, many works have attempted to leverage similar techniques to solve the phase retrieval problem in the compressive setting $m < n$. When the $n$-dimensional signal is $s$-sparse, the information theoretic lower bound of $m = O(s \log n)$ measurements was shown to be required for the injectivity of phaseless Gaussian measurements [74]. However, attempts at achieving this optimal sample complexity via a polynomial time algorithm have proven quite difficult and, in some cases, impossible. For example, the natural $\ell_1$-penalized variant of Phaselift was shown to be able to recover an $s$-sparse signal with $O(s^2 \log n)$ generic Gaussian measurements, but this bound was also proven to be tight [52, 59]. Moreover, there are a number of results that show, if one were able to construct a sufficiently accurate initializer of the true solution, then recovery from $O(s \log n)$ Gaussian measurements is possible by a variety of methods [37, 67, 77]. Known initialization schemes to accomplish this, however, require $O(s^2 \log n)$ measurements [14]. For a more complete discussion of prior methodologies for phase retrieval, we refer the reader to ref. [28].
Some existing works in compressive phase retrieval establish optimal sample complexity recovery guarantees under non-generic measurements [43]. For example, [6] showed that assuming the measurement vectors were chosen from an incoherent subspace, then recovery is possible with $O(s \log \frac{n}{s})$ measurements. When the measurement matrix is a product of matrices amenable for dense phase retrieval and CS algorithms, then $O(s \log n)$ measurements suffices as well [42]. Lastly, using the notion of polarization, [8] showed that $O(s \log n)$ measurements also suffices for recovery when the measurement vectors have an associated graph with sufficient connectivity properties. However, these results would be difficult to generalize to the experimental setting as their measurement design architectures are often unrealistic, with generic Gaussian measurements offering a closer model to the goal of Fourier diffraction measurements.

In addition to standard sparsity-based models, there have also been methods that utilize more complex and structured sparsity-based models for phase retrieval, such as group sparsity and non-local self-similarity as signal priors. For example, [54] incorporates a BM3D-based denoiser in a general message passing framework for compressive phase retrieval. Such a prior is based on the exploitation of group sparsity for signals. Such methods have shown to succeed with fewer measurements than standard sparsity priors. Much less theory is known, however, in regards to sample complexity guarantees for structured sparsity-based models.

We further highlight recent empirical work in using deep learning to solve phase retrieval problems. In ref. [55], recovery is posed as an optimization problem via the regularization-by-denoising framework where a deep convolutional neural network is used as a denoiser. Convolutional neural networks have also been used in an end-to-end fashion to solve various phase retrieval tasks, such as Fourier ptychography [46], holography [34], and holographic microscopy [80]. As many traditional iterative procedures for phase retrieval are sensitive to initialization, other works have used data-driven approaches to yield good initializers for such methods [62, 70]. Lastly, we highlight empirical work with promising results on using generative models as priors for phase retrieval [41, 65] along with extensions that incorporate measurement information to improve performance [72]. The results in ref. [65], for example, showcase that generative priors can also outperform structured sparsity-based methods in compressive regimes.

Signal recovery with generative priors

In ref. [13], the authors studied enforcing a generative prior in the linear CS regime. In particular, given $m$ linear measurements $Ay_*$ where $y_* \in \mathbb{R}^n$, the authors modeled natural signals as being in the range of a trained generative model $G : \mathbb{R}^k \to \mathbb{R}^n$ where $k \ll n$. To solve the inverse problem, they proposed to find a latent code $x_* \in \mathbb{R}^k$ such that $G(x_*) \approx y_*$ by solving the following least squares objective

$$\min_{x \in \mathbb{R}^k} \frac{1}{2} \| AG(x) - Ay_* \|^2. \tag{1.1}$$

They provided empirical evidence showing that 5–10× fewer measurements were needed to achieve comparable reconstruction errors, compared to standard sparsity-based approaches such as Lasso in some parameter regimes. Based on the success of generative priors in CS, a number of followup works have considered a similar setup in a variety of inverse problems, ranging from compressed sensing [32, 37, 40, 64, 68], denoising [1, 39, 56], phase retrieval [36, 65], low-
rank matrix recovery [5, 20], one-bit CS [53, 61], blind deconvolution [3, 35], and more. This framework, in the case of CS, enjoys multiple theoretical analyses. A subset of the authors in ref. [37] presented the first global landscape analysis of the empirical risk minimization problem and showed that, in fact, when the network is sufficiently expansive with Gaussian weights and the number of measurements is proportional to \( k \), there exists a descent direction everywhere outside of potentially two small neighborhoods of the minimizer and true solution. Followup work later established convergence guarantees of first order methods in compressed sensing [40] and denoising [39] under similar statistical assumptions on the generator. In the present work, we consider precisely the same random model in the context of compressive phase retrieval.

1.2 Compressive phase retrieval with generative priors

The compressive phase retrieval problem is as follows. We consider the real-valued version out of simplicity. Consider a signal \( y^* \in \mathbb{R}^n \). Given \( m \) phaseless linear measurements of the form

\[
b^* = | Ay^* | + \eta,
\]

where \( m < n \), \( A \in \mathbb{R}^{m \times n} \) is a known linear operator, and \( \eta \in \mathbb{R}^m \) denotes measurement noise, the goal is to recover \( y^* \) from knowledge of \( b^* \) and \( A \). As \( m < n \), additional structure must be exploited to accurately estimate \( y^* \). In this work, we assume that \( y^* \) belongs in or near the range of a trained generative model \( G : \mathbb{R}^k \rightarrow \mathbb{R}^n \). That is, \( y^* \approx G(x^*) \) for some latent code \( x^* \). In order to recover an estimate of a signal \( y^* \), it suffices to recover \( x^* \) and then compute \( G(x^*) \). We propose to solve the following nonlinear least squares problem:

\[
\min_{x \in \mathbb{R}^k} f(x) := \frac{1}{2} \left\| |Ag(x)| - b^* \right\|^2 .
\]

This formulation attempts to find the signal in the range of the generative model \( G \) that is most consistent with provided measurements in a particular sense. It is motivated both by the non-convex generative modeling formulation for compressed sensing in ref. [13] with an Amplitude Flow [81] perspective from phase retrieval and was originally introduced by the present authors in ref. [36]. As the underlying optimization problem is posed over an explicitly parameterized \( k \)-dimensional manifold where \( k \ll n \), compressive phase retrieval may be possible from \( m = \Omega(k) \ll n \) measurements.

In this paper, we prove that (1.3) can be solved with sample complexity proportional to \( k \), under an appropriate model for \( G \) and a generic measurement model. This theoretical result is in stark contrast to algorithms for compressive phase retrieval based on sparsity priors, where no known tractable algorithm achieves information theoretically optimal sample complexity under a generic measurement model. This result extends the work of ref. [39, 40] which established similar algorithmic guarantees for recovery with generative priors in the linear measurement regimes of denoising and compressed sensing. Additionally, we provide experimental results that (1.3) can outperform sparsity-based compressive phase retrieval algorithms in the presence of a trained \( G \) from standardly available datasets.

The formulation assumes that the generative model \( G \) is already known. In practice, it typically is a neural network whose parameters (weights) are learned from a large collection of
training images belonging to a particular natural signal class. The field of generative modeling has demonstrated multiple types of neural networks which can be effectively trained (e.g., VAEs [49] and Generative Adversarial Networks [33]). The dimensionality $k$ of the latent codes is fixed at training time and its value is selected in order to balance multiple effects; for example, the range of $G$ should be large enough to approximately include all of the desired signal class, and the image representations should be as concise as possible. A particular image of interest may not be exactly in the range of a trained model $G$ because the model has representation error, but this error is expected to become smaller as techniques for training generative models improve.

1.3 Deterministic and probabilistic models for generative priors

In order to establish recovery guarantees for phase retrieval with generative priors, we will assume a neural network architecture and a model for the weights of the network once trained. Our intention is to analyze a model which is realistic enough to describe trained nets, yet tractable enough to permit rigorous analysis of sample complexity for a convergent optimization algorithm. To achieve both of these objectives we consider the following models. We assume that the generative model $G : \mathbb{R}^k \to \mathbb{R}^n$ is given by a $d$-layer feedforward neural network with ReLU activation functions and no bias terms. Specifically, we assume that

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

(1.4)

where $\text{relu}(x) := \max(x, 0)$ acts entrywise and each $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ for $i \in [d]$ with $k = n_0 < n_1 < n_2 < \cdots < n_d = n$. Each matrix $W_i$ corresponds to the neural network weights of the $i$-th layer, and the $j$-th row of $W_i$ are the weights of the $j$-th neuron in the $i$-th layer.

We will assume an expansive-Gaussian probabilistic model for the weights of $G$. That is, $n_i$ increases sufficiently with $i$, and the weights within each layer are i.i.d. Gaussians. This model was introduced by a subset of the authors in ref. [38]. We additionally assume a Gaussian model of the measurement matrix $A$. The justification for these assumptions is as follows. The Gaussianity of $A$ ensures that measurements are suitably generic, and, indeed, achieving optimal sample complexity in sparse phase retrieval has not been attained for this measurement model. Regarding the expansivity assumption, we note that generative models with low dimensional latent spaces are inherently expansive when considered as a whole. In a sense, the network and each layer therein could be viewed as adding redundancy to a more compact representation, though in practice some successful network architectures do not have strict layerwise expansivity. Regarding the Gaussianicity model of neural network weights, it has been shown that neural networks, such as AlexNet, trained on real data have resulting weights with statistics similar to Gaussians [2].

We emphasize that the use of generative models as priors in regularizing inverse problems is nascent, and we use this model because it balances mathematical tractability with authenticity toward applications.

In order to establish a recovery guarantee for this random model, we establish it for models satisfying deterministic conditions on $G$ and $A$. Then we show that an appropriate expansive-Gaussian model satisfies these deterministic conditions with high probability. The first deterministic condition we consider roughly states that the neural network weights are approximately distributed uniformly on a sphere of a particular radius. For $W \in \mathbb{R}^{n \times k}$ and $x \in \mathbb{R}^k$, define
$W_{+,x} := \text{diag}(Wx > 0)W$ where the $i$-th diagonal entry of $\text{diag}(v > 0)$ is 1 if $v_i > 0$ and 0 otherwise. Note that $W_{+,x} x = \text{relu}(Wx)$. The condition is stated as follows and was introduced in ref. [38]:

**Definition 1.1** (Weight Distribution Condition [38]). We say that $W \in \mathbb{R}^{n \times k}$ satisfies the **Weight Distribution Condition (WDC)** with constant $\varepsilon > 0$ if for all nonzero $x, z \in \mathbb{R}^k$,

$$\left\| W_{+,x}^T W_{+,z} - Q_{x,z} \right\| \leq \varepsilon$$

where

$$Q_{x,z} := \frac{\pi - \theta_{x,z}}{2\pi} I_k + \frac{\sin \theta_{x,z}}{2\pi} M_{\hat{x} \leftrightarrow \hat{z}}. \quad (1.5)$$

Here $\theta_{x,z} = \angle(x, z)$, $\hat{x} = x/\|x\|$, $\hat{z} = z/\|z\|$, $I_k$ is the $k \times k$ identity matrix, and $M_{\hat{x} \leftrightarrow \hat{z}}$ is the matrix that sends $\hat{x} \mapsto \hat{z}$, $\hat{z} \mapsto \hat{x}$, and $u \mapsto 0$ for any $u \in \text{span}\{x, z\}^\perp$.

A specific formula for $M_{\hat{x} \leftrightarrow \hat{z}}$ is given as follows. Consider a rotation matrix $R$ that sends $\hat{x} \mapsto e_1$ and $\hat{z} \mapsto \cos \theta_{x,z} e_1 + \sin \theta_{x,z} e_2$ where $\theta_{x,z} = \angle(x, z)$. Then

$$M_{\hat{x} \leftrightarrow \hat{z}} = R^T \begin{bmatrix} \cos \theta_{x,z} & \sin \theta_{x,z} & 0 \\ \sin \theta_{x,z} & -\cos \theta_{x,z} & 0 \\ 0 & 0 & 0_{k-2} \end{bmatrix} R,$$

where $0_{k-2}$ is the $k-2 \times k-2$ matrix of zeros. Note that if $\theta_{x,z} = 0$ or $\pi$, $M_{\hat{x} \leftrightarrow \hat{z}} = \hat{x} \hat{x}^T$ or $-\hat{x} \hat{x}^T$, respectively.

The second deterministic condition provides an RIP-like property for the measurement matrix $A$ when acting on pairs of secant directions within the range of $G$. For $A \in \mathbb{R}^{m \times n}$ and $y \in \mathbb{R}^n$, define $A_y := \text{diag}(\text{sgn}(Ay))A$ where $\text{sgn}$ acts entrywise, and $\text{sgn}(0) = 0$. Note that $A_y y = |Ay|$. The condition is stated as follows and was introduced in a conference version of this work [36]:

**Definition 1.2** (Range Restricted Concentration Property). We say that $A \in \mathbb{R}^{m \times n}$ satisfies the **Range Restricted Concentration Property (RRCP)** with respect to $G$ with constant $\varepsilon > 0$ if for all $x, z, x_1, x_2, x_3, x_4 \in \mathbb{R}^k$:

$$|\langle (A_{G(x)}^TA_{G(z)} - \Phi_{G(x),G(z)})(G(x_1) - G(x_2)), G(x_3) - G(x_4) \rangle| \leq L \varepsilon \|G(x_1) - G(x_2)\| \|G(x_3) - G(x_4)\|$$

where

$$\Phi_{y,w} := \begin{cases} \frac{\pi - 2\theta_{y,w}}{\pi} I_n + \frac{2\sin \theta_{y,w}}{\pi} M_{\hat{y} \leftrightarrow \hat{w}} & \text{if } y \neq 0, w \neq 0, \\ 0 & \text{otherwise}. \end{cases} \quad (1.6)$$

Here, $L$ is a universal constant and can be taken to be 33.
Algorithm 1 Deep Phase Retrieval (DPR) Subgradient method

Require: Weights $W_i$, measurement matrix $A$, measurements $b_* = |Ay_*| + \eta$, & step size $\alpha > 0$

1: Choose an arbitrary initial point $x_0 \in \mathbb{R}^k \setminus \{0\}$
2: for $t = 0, 1, 2, \ldots$ do
3: if $f(-x_t) < f(x_t)$ then
4: $\bar{x}_t \leftarrow -x_t$;
5: else
6: $\bar{x}_t \leftarrow x_t$;
7: end if
8: Compute $v_{\bar{x}_t} \in \partial f(\bar{x}_t)$;
9: $x_{t+1} = \bar{x}_t - \alpha v_{\bar{x}_t}$;
10: end for

1.4 Algorithm

We provide a subgradient algorithm for optimizing (1.3) under noisy measurements. We show that this algorithm converges up to the noise level when the WDC and RRCP properties are met. In order to state the algorithm, we need some notation. For a locally Lipschitz function $f : \mathcal{X} \to \mathbb{R}$ from a Hilbert space $\mathcal{X}$ to $\mathbb{R}$, the Clarke generalized directional derivative of $f$ at $x \in \mathcal{X}$ in the direction $u$, is defined by

$$f^0(x; u) := \limsup_{z \to x, t \downarrow 0} \frac{f(z + tu) - f(z)}{t}.$$ 

Then the generalized subdifferential of $f$ at $x$ is defined by

$$\partial f(x) = \{v \in \mathbb{R}^k : \langle v, u \rangle \leq f^0(x; u), \forall u \in \mathcal{X}\}.$$ 

Any element $v_x \in \partial f(x)$ is called a subgradient of $f$ at $x$. When $f$ is differentiable at $x$, $\partial f(x) = \{\nabla f(x)\}$.

We now introduce a subgradient descent scheme, given by Algorithm 1, whose intuition is as follows. In expectation, the optimization landscape is characterized by Figure 1. There exists two critical points away from the origin: the true minimizer and a negative multiple thereof. Moreover, the value of the objective function is higher near the negative multiple than near the global minimizer. At each iterate, we check the objective function value at the current latent code and its negative, choosing the point with smaller objective function value as our new iterate; see Steps 3–7. We then perform subgradient descent.

1.5 Main results

In this section, we outline our main results in both the probabilistic and deterministic settings. In particular, in Theorem 1.3 we show that if the weights of $G$ satisfy the WDC and the measurement matrix $A$ satisfies the RRCP, then the iterates of Algorithm 1 converge to the true solution up to
the noise level in the measurements. Then, in Corollary 1.6 we show that the same conclusion holds when $W_i$ and $A$ are Gaussian with high probability as soon as $m = \Omega(d k \log(n_1 n_2 \ldots n_d))$.

We consider the possibly noisy measurements (1.2) and assume that the signal $y_\star$ is in the range of $G$ with latent code $x_\star$; that is, $y_\star = G(x_\star)$. The following Theorem states that if the two deterministic conditions are satisfied with a sufficiently small parameter $\varepsilon$ and the noise is sufficiently small, then the iterates of Algorithm 1 will converge to $x_\star$ up to the noise level.

**Theorem 1.3** (Deterministic Convergence Guarantee). Let $d \geq 2$ and fix $0 < \varepsilon < c_1 \frac{1}{d^{1/2}}$. Suppose the noise satisfies $\|\eta\| \leq c_2 \frac{\|x_\star\|_2}{d^{1/2}}$. Suppose each $W_i$ of $G$ satisfies the WDC with constant $\varepsilon$, and suppose $A$ satisfies the RRCP with respect to $G$ with constant $\varepsilon$. Then the iterates $\{x_t\}_{t \geq 0}$ generated by Algorithm 1 with step size $\alpha \leq c_3 \frac{\varepsilon d}{d^2}$ obey the following:
(1) there exists an $N \in \mathbb{N}$ satisfying $N \leq C_4 \frac{f(x_0)(2d)}{d^4 \pi^2 \|x_0\|^2}$ such that

$$
\|x_N - x_*\| \leq C_5 d^{12} \sqrt{\varepsilon} \|x_0\| + C_6 d^{9} \|\eta\|; \quad (1.7)
$$

(2) for all $t \geq N$, we have

$$
\|x_{t+1} - x_*\| \leq \tau^{t+1-N} \|x_N - x_*\| + \partial \frac{2d}{d^2} \|\eta\|, \quad \text{and} \quad (1.8)
$$

$$
\|G(x_{t+1}) - G(x_*)\| \leq \frac{1.2}{2d^2} \tau^{t+1-N} \|x_N - x_*\| + \frac{1.2}{d^2} \partial \|\eta\| \quad (1.9)
$$

where $\tau := 1 - \frac{7}{8} \frac{\varepsilon}{2^d} \in (0, 1)$ and $\partial := \frac{2c_3}{1-\varepsilon}$.

Here $c_1, c_2, c_3, C_4, C_5,$ and $C_6$ are positive universal constants.

This result asserts that the iterates of Algorithm 1 will eventually be in a small neighborhood of the true solution whose size depends on $\varepsilon$ and $\|\eta\|$ after $N = O(\varepsilon^{-1})$ iterations. Furthermore, once in this neighborhood, the iterates will continue to converge linearly to the true solution up to the noise level. If no noise is present, then the true signal will be recovered. Note that the $2^d$ factors in the theorem are an artifact of the problem scaling. Roughly, the weights $W_i$ have spectral norm approximately 1, and subsequent application of a ReLU will effectively zero out roughly half of the rows of $W_i$. The resulting rows of $W_i$ will have spectral norm of roughly 1/2. Hence $G(x)$ scales like $2^{-d/2} \|x\|$, $f(x)$ scales like $2^{-d} \|x\|^2$, and any subgradient $v_x$ scales like $2^d$. We also assume the noise scales like $2^{-d/2}$ to ensure it is on the order of the measurements. Doubling the variance of each entry of $W_i$ would eliminate these factors, but we consider the unscaled version because it is more convenient in the analysis.

The main deterministic result required that the weight ensembles and measurement matrix to satisfy the WDC and RRCP, respectively. We now analyze under what regimes do Gaussian models satisfy such conditions. For the generator weights, we appeal to a result that shows that expansive (tall) Gaussian matrices satisfy the WDC with high probability.

**Lemma 1.4** (Lemma 11 in ref. [38]). Fix $0 < \varepsilon < 1$ and suppose $W \in \mathbb{R}^{n \times k}$ has i.i.d. $\mathcal{N}(0, 1/n)$ entries. Then if $n \geq C_7 k \log k$, then with probability at least $1 - 8n \exp(-\gamma \varepsilon k)$, $W$ satisfies the WDC with constant $\varepsilon$. Here $C_7$ and $\gamma$ depend polynomially on $\varepsilon^{-1}$.

In this work, we establish that Gaussian matrices $A$ satisfies the RRCP with respect to an expansive-Gaussian $G$ with high probability if they are sufficiently tall. This result is proven in Section 3:

**Lemma 1.5.** Fix $0 < \varepsilon < 1$ and suppose $A \in \mathbb{R}^{m \times n}$ has i.i.d. $\mathcal{N}(0, 1/m)$ entries. Let $G$ be a generative model of the form (1.4) where each $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ has i.i.d. $\mathcal{N}(0, 1/n_i)$ entries. If $m \geq C_8 k d \log(n_1 n_2 \ldots n_d)$, then with probability $1 - \gamma m^{4k} \exp(-c_{10} m)$, $A$ satisfies the RRCP with respect to $G$ with constant $\varepsilon$. Here $\gamma$ is a universal constant and $C_8$ and $c_{10}$ depend polynomially on $\varepsilon^{-1}$. 
Hence for Gaussian measurements and weight ensembles, we can combine Lemma 1.4 and 1.5 with Theorem 1.3 to obtain the following Corollary:

**Corollary 1.6 (Probabilistic Convergence Guarantee).** Fix $0 < \epsilon < c_1 \frac{1}{d^{102}}$ and suppose the noise satisfies $\|\eta\| \leq c_2 \frac{\|x_s\|}{d^{2/d^2}}$ for some universal constants $c_1$ and $c_2$. Suppose $G$ is such that $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ has i.i.d. $\mathcal{N}(0, 1/n_i)$ entries for $i = 1, \ldots, d$. Suppose that $A \in \mathbb{R}^{m \times n_d}$ has i.i.d. $\mathcal{N}(0, 1/m)$ entries independent from $\{W_i\}$. Then if $m \geq C_\epsilon dk \log(n_1 n_2 \ldots n_d)$ and $n_i \geq C_\epsilon n_{i-1} \log n_{i-1}$ for $i = 1, \ldots, d$, then with probability at least $1 - \sum_{i=1}^d \gamma n_i \exp(-c_\epsilon n_{i-1}) - \gamma m^{4k} \exp(-c_\epsilon m)$, the same conclusion as Theorem 1.3 holds. Here $C_\epsilon$ depends polynomially on $\epsilon^{-1}$, $c_\epsilon$ depends on $\epsilon$, and $\gamma$ is a universal constant.

To the authors’ knowledge, this is the first result establishing provable signal recovery with a computationally efficient algorithm for undersampled generic phaseless linear measurements with optimal sample complexity. This sample complexity in our result scales with $k$, which can not be improved. We made no attempt to obtain tight bounds on $d$, except to ensure that all dependencies on $d$ are polynomial. We remind the reader that any $2^d$ terms that appear are due to the problem scaling. We further note that subsequent developments since the original release of ref. [36] relaxed the logarithmic growth factor on the sizes of each layer of the generative model [21].

Lastly, we note that this result for compressive phase retrieval under optimal sample complexity implies recovery for linear CS under optimal sample complexity. As such, this work subsumes the work of a subset of the authors in ref. [38]. This generalization of compressed sensing to compressive phase retrieval is conspicuously absent for structural priors based on sparsity, as the best known computationally efficient algorithms for sparsity priors require sample complexity that is quadratic in the sparsity level.

### 1.6 Experiments on MNIST

In this section, we compare the generative modeling approach for compressive phase retrieval with three sparse phase retrieval algorithms: the sparse truncated amplitude flow algorithm (SPARTA) [77], Thresholded Wirtinger Flow (TWF) [14], and the compressive phase retrieval with alternating minimization algorithm (CoPRAM) [44]. For the generative modeling approach, we used a modified version of Algorithm 1 as we empirically found the negation step (Steps 3–4) only occurred at the first iterate. Hence we ran two gradient descents, one starting from a random initial iterate $x_0$ and another starting from its negation $-x_0$. We report results for the most successful reconstruction. Gradient descent was performed using the Adam optimizer [48]. For the remainder of this section, we will refer to the generative modeling approach as DPR.

In each task, the goal is to recover an image $y_*$ given $|Ay_*|$ where $A \in \mathbb{R}^{mn \times n}$ has i.i.d. $\mathcal{N}(0, 1/m)$ entries. The images were from the Modified National Institute of Standards and Technology (MNIST) dataset [51]. This dataset consists of 60,000 $28 \times 28$ images of handwritten digits. The generative model was a pretrained VAE from ref. [13]. The encoder network is of size $784 \rightarrow 500 \rightarrow 500 \rightarrow 20$ while the generator network $G$ is of size $20 \rightarrow 500 \rightarrow 500 \rightarrow 784$. The latent code space dimension is $k = 20$.

For the sparse phase retrieval methods, we performed sparse recovery in the Daubechies-4 Wavelet domain. We zero-padded the images to be of size $32 \times 32$. The resulting images generated by our algorithm were also uniformly padded with zeros around the border to obtain
32 × 32 images. For SPARTA and CoPRAM, we ran each algorithm with sparsity parameters ranging from 2 to 212 in increments of 15, choosing the best reconstruction in terms of lowest reconstruction error.

We aimed to reconstruct 10 images from the MNIST test set. We allowed five random restarts for each algorithm and chose the result with the least $\ell_2$ reconstruction error per pixel. We also report the Structural Similarity Index Measure (SSIM) \[78\] for each reconstruction. The results in Figure 2 demonstrate the success of our algorithm with very few measurements. For 200 measurements, we can achieve accurate recovery with a mean SSIM value of over 0.9 while other algorithms require 1000 measurements or more. In terms of reconstruction error, our algorithm exhibits recovery with 200 measurements comparable to the alternatives requiring 750 measurements or more, which is where they begin to succeed.

We note that while our algorithm succeeds with fewer measurements than the other methods, our performance, as measured by per-pixel reconstruction error, saturates as the number of measurements increases since our reconstruction accuracy is ultimately bounded by the generative model’s representational error. As generative models improve, their representational errors will decrease. Nonetheless, as can be seen in the reconstructed digits in Figure 3, the recoveries are semantically correct (the correct digit is legibly recovered) even though the reconstruction error does not decay to zero. In applications, such as MRI and molecular structure estimation via X-ray crystallography, semantic error measures would be a more informative estimates of recovery performance than per-pixel error measures.

1.6.1 Comparisons with dictionary learning

We additionally compared generative priors to sparse phase retrieval methods using a learned basis to sparsify MNIST. More specifically, let $Y \in \mathbb{R}^{n \times N}$ denote the training set of $N = 60000$ MNIST digits. We learned a dictionary $D_* \in \mathbb{R}^{n \times p}$ with $p = 300$ atoms by solving the following optimization problem:

$$
\min_{D \in D_p} \frac{1}{N} \sum_{i=1}^{N} \min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \| Y_i - D \alpha \|^2 + \lambda \| \alpha \|_1
$$

Here, $D_p$ denotes the set of $n \times p$ dictionaries with unit norm columns. We solved the problem in an alternating fashion, where for the inner sparse coding problem, we used the Fast Iterative Soft Thresholding Algorithm \[11\] with $\lambda = 0.2$ and projected gradient descent for the outer optimization over $D \in D_p$.

Once this dictionary $D_*$ is learned, we consider solving phase retrieval problems $|Ay_*|$ where a new digit $y_*$ from the test set is to be recovered. In order to incorporate the dictionary into the recovery algorithm for sparse phase retrieval, we employ a sparse phase retrieval algorithm to recover the sparse latent code $\alpha_*$ so that $y_* \approx D_* \alpha_*$. We compared deep generative priors to SPARTA and CoPRAM augmented with dictionary learning. These methods are denoted by SPARTA-DL and CoPRAM-DL. We ran each algorithm with sparsity parameters ranging from 5 to 300 in increments of 20, choosing the best reconstruction in terms of lowest reconstruction error.

In this experiment, we trained two VAEs with different latent dimensions, one with $k = 20$ and the other with $k = 50$. Similarly to the previous experiments, we report the average $\ell_2$ error
and SSIM each algorithm. Results are shown in Figure 4. Overall, DPR continues to outperform such sparsity-based methods incorporating a learned dictionary with the best performing method being DPR with a generator of latent dimension $k = 50$. The performance of such methods for $m \geq 200$ measurements is on par with the generative prior-based approach with a latent
dimension of \( k = 20 \) in terms of \( \ell_2 \) reconstruction error, but in terms of a perceptual metric, DPR outperforms both methods. For severely undersampled regimes, DPR outperforms both methods in both reconstruction error and SSIM.

2 PROOF OF CONVERGENCE RESULT IN DETERMINISTIC SETTING

In this section, we will formally prove Theorem 1.3. Section 2.1 outlines the notation we will use throughout the proofs. Section 2.2 provides a high-level sketch of our proof and outlines its central arguments while Section 2.3 discusses preliminary results that are used throughout the proofs. Then Section 2.4 presents the proof of Theorem 1.3 which is broken down into four central results. Finally, Section 2.5 presents supplementary results and their proofs that aid in establishing Theorem 1.3.
FIGURE 4  Experimental comparisons on solving sparse phase retrieval problems with methods incorporating a learned dictionary and deep generative priors. (Top) $\ell_2$ reconstruction error and (bottom) SSIM averaged over 10 images from the test set for varying numbers of measurements. SSIM, Structural Similarity Index Measure.

2.1  Notation

Let $(\cdot)^T$ denote the real transpose. Let $[n] = \{1, \ldots, n\}$. Let $B(x, r)$ denote the closed Euclidean ball centered at $x$ with radius $r$. Let $\| \cdot \|$ denote the $\ell_2$ norm for vectors and spectral norm for matrices. For any non-zero $y \in \mathbb{R}^n$, let $\hat{y} = y/\|y\|$. For non-zero $q, y \in \mathbb{R}^n$, let $\theta_{q,y} = \angle(q, y)$. Let
relu(x) := max(x, 0). Define sgn : ℝ → ℝ to be sgn(x) = x/|x| for non-zero x ∈ ℝ and sgn(0) = 0 otherwise. Let 1(E) be the indicator function on the event E. For a vector v ∈ ℝ^n, diag(sgn(v)) is sgn(v_i) in the i-th diagonal entry and diag(v > 0) is 1 in the i-th diagonal entry if v_i > 0 and 0 otherwise. Let Π_i=1^d W_i = W_d W_{d-1} ... W_1. For any x ∈ ℝ^k and i ∈ [d], define W_{i,+} := diag(W_{i-1,+} ... W_{2,+} W_{1,+} x > 0) W_i. Let Λ_x := Π_i=1^d W_{i,+} x = Λ_x x. Note that we have the following string of equalities: G(x) = Π_i=1^d W_{i,+} x = Λ_x x = x_d.

Let I_n be the n × n identity matrix. Let  denote the unit sphere in ℝ^k. We write γ = Ω(𝛿) when γ ≥ C 𝛿 for some positive constant C. Similarly, we write γ = O(𝛿) when γ ≤ C 𝛿 for some positive constant C. When we say that a constant depends polynomially on ε^{-1}, this means that it is at least Cε^{-k} for some positive C and positive integer k. Positive numerical constants will be denoted using C or K with various subscripts. In general, numerical constants larger than one will be denoted by capital letters and constants smaller than one with lower case letters. For notational convenience, we write a = b + O_1(ε) if ‖a − b‖ ≤ ε where ‖·‖ denotes |·| for scalars, ℓ_2 norm for vectors, and spectral norm for matrices.

2.2 Sketch of proof for Theorem 1.3

Theorem 1.3 is proven by showing that for all x ∈ ℝ^k, any subgradient v_x ∈ ∂ f(x) is approximated by h_x ∈ ℝ^k which has an analytical expression and that does not vanish outside of neighborhoods of the true solution x_* and a negative multiple −ρ_d x_* for some ρ_d ∈ (0, 1). Thus any v_x ∈ ∂ f(x) is bounded away from zero for x outside of these two neighborhoods, leading to convergence towards one of these regions. Then we ensure that the negation step of our algorithm (Steps 3–7) will update any iterate near −ρ_d x_* to eventually be in a neighborhood of x_* Finally, we ensure convergence to x_* up to the noise level by showing that the objective function exhibits a convexity-like property in a neighborhood of x_*.

To provide our sketch, we define some quantities. Define the function g : [0, 2π] → ℝ by

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

(2.1)

For any x ∈ ℝ^k \ {0}, let h_x ∈ ℝ^k be defined as

\[ h_x := \frac{‖x‖}{2^d} \left(\frac{\pi - 2\bar{\theta}_{d,x}}{\pi}\right) \left(\prod_{i=0}^{d-1} \frac{\pi - \bar{\theta}_{i,x}}{\pi}\right) \hat{x} + \frac{1}{2^d} \left[‖x‖ - ‖\hat{x}‖\right] \left(\frac{2 \sin \bar{\theta}_{d,x}}{\pi} + \left(\frac{\pi - 2\bar{\theta}_{d,x}}{\pi}\right) \sum_{i=0}^{d-1} \frac{\sin \bar{\theta}_{i,x}}{\pi} \left(\prod_{j=i+1}^{d-1} \frac{\pi - \bar{\theta}_{j,x}}{\pi}\right)\right] \hat{x} \]

where \( \bar{\theta}_{0,x} = \angle(x, x_*) \) and \( \bar{\theta}_{i,x} = g(\bar{\theta}_{i-1,x}) \) for i ∈ [d]. We further define

\[ \rho_d := \frac{2 \sin \bar{\theta}_d}{\pi} + \left(\frac{\pi - 2\bar{\theta}_d}{\pi}\right) \sum_{i=0}^{d-1} \frac{\sin \bar{\theta}_i}{\pi} \left(\prod_{j=i+1}^{d-1} \frac{\pi - \bar{\theta}_j}{\pi}\right) \]
where $\tilde{\theta}_0 = \pi$ and $\tilde{\theta}_i = g(\tilde{\theta}_{i-1})$ for $i \in [d]$. For a parameter $\beta > 0$, define

$$S_\beta := \left\{ x \in \mathbb{R}^k \setminus \{0\} : \|h_x\| \leq \frac{1}{2d}\beta \max(\|x\|, \|x_\ast\|) \right\}. \quad (2.2)$$

A direct analysis in Lemma 2.6 shows that for appropriate values of $\beta$, $S_\beta$ is contained in the union of neighborhoods of $x_\ast$ and $-\rho_d x_\ast$:

$$S_\beta \subset B(x_\ast, 70000\pi^2 d^9 \beta \|x_\ast\|) \cup B(-\rho_d x_\ast, 77422\pi^2 d^{12} \sqrt{\beta} \|x_\ast\|).$$

Set

$$S_\beta^+ := S_\beta \cap B(x_\ast, 70000\pi^2 d^9 \beta \|x_\ast\|)$$

and

$$S_\beta^- := S_\beta \cap B(-\rho_d x_\ast, 77422\pi^2 d^{12} \sqrt{\beta} \|x_\ast\|).$$

A sketch of our proof is outlined as follows:

- First, we establish that all subgradients are bounded away from zero for iterates outside of $S_\beta$. Specifically, we show that when the WDC and RRCP are satisfied with constant $\varepsilon$, any $v_{x_i} \in \partial f(x_i)$ satisfies $v_{x_i} \approx h_{x_i}$ and $h_{x_i}$ is bounded away from 0 by the definition of $S_\beta$. Thus $\|v_{x_i}\|$ must be bounded away from zero for points $x_i \not\in S_\beta$. See Section 2.4.1.

- Next, we establish convergence to $S_\beta$. In particular, we show that the previous result implies that subgradient descent at each iteration makes progress in the sense that for each non-zero $x_i \not\in S_\beta$

$$f(x_{i+1}) - f(x_i) \leq -C\varepsilon$$

for some $C > 0$. Thus after $\Omega(\varepsilon^{-1})$ iterations, the iterates will eventually belong to $S_\beta$. See Section 2.4.2.

- Third, we show that the negation step of our algorithm ensures that the iterates converge to $S_\beta^+$. Specifically, we prove that for points $x \approx x_\ast$ and $z \approx -\rho_d x_\ast$, $f(x) < f(z)$. Thus if an iterate $x_i \in S_\beta^-$, $f(-x_i) < f(x_i)$ so the negation step of our algorithm (Steps 3–7) and strict decreasing property of the objective from the previous step ensure the iterates eventually enter $S_\beta^+$. See Section 2.4.3.

- Finally, we establish convergence to $x_\ast$ up to the noise level. Specifically, we prove that once in $S_\beta^+$, a convexity-like property near $x_\ast$ implies that the iterates converge to $x_\ast$ up to the noise level in the measurements. See Section 2.4.4.

### 2.3 Preliminaries for proofs

We will make use of the following fact concerning the Clarke subdifferential of the objective function $f$. Since $f$ is piecewise quadratic, Theorem 9.6 from ref. [19] asserts that for any $x \in \mathbb{R}^k$, the
Clarke subdifferential $\partial f(x)$ can be written equivalently as
\[
\partial f(x) = \text{conv}(v_1, v_2, \ldots, v_s) = \left\{ \sum_{\ell=1}^{s} c_{\ell} v_{\ell} : \sum_{\ell=1}^{s} c_{\ell} = 1 \text{ and } c_{\ell} \geq 0 \text{ for } \ell \in [s] \right\} \tag{2.3}
\]
where $\text{conv}(\cdot)$ denotes the convex hull, $s$ is the number of quadratic functions adjoint to $x$, and $v_{\ell}$ is the gradient of the $\ell$-th quadratic function of $f$ at $x$. Moreover, for each $v_{\ell}$, there exists a direction $w_{\ell}$ and a sufficiently small $\delta_{\ell} > 0$ such that $f$ is differentiable at $x + \delta_{\ell} w_{\ell}$ and $v_{\ell} = \lim_{\delta_{\ell} \to 0^+} \nabla f(x + \delta_{\ell} w_{\ell})$.

2.4 Proof of Theorem 1.3

We now set out to prove Theorem 1.3. In Sections 2.4.1–2.4.4, we establish four main lemmas, each of which pertain to one of the items in the sketch of our proof from Section 2.2. Theorem 1.3 is then proven in Section 2.4.5. Prior to beginning the proof, we state the necessary assumptions we will make:

**Assumptions A.** We assume the following hold for some numerical constants $c_1$, $c_2$, and $c_3$:

A1. $0 < \varepsilon < c_1 d^{-102}$,
A2. the noise $\eta$ satisfies $\|\eta\| \leq c_2 \|x^*\| \frac{d}{2^{d/2}4^8}$, and
A3. the step size $\alpha > 0$ satisfies $\alpha \leq c_3 \frac{2^d}{d^2}$.

We note that Proposition 2.12 shows that after a polynomial number of steps, the iterates of our algorithm stay outside of a ball of the origin. Hence we assume throughout that the norm of our iterates are bounded away from zero. This result is proven in Section 2.5.1.

2.4.1 Uniform control over subgradients

We first show that the descent direction does not vanish for points outside of $S_\beta$. The main idea of this result is that for points $x$ such that $\|h_x\|$ is sufficiently bounded away from zero, any $v_x \in \partial f(x)$ is also bounded away from zero.

To prove this, we require the following three lemmas. The first gives a simple upper bound on the norm of our descent direction.

**Lemma 2.1.** Fix $\varepsilon > 0$ such that $Kd^3 \sqrt{\varepsilon} \leq 1$ where $K$ is a universal constant. Suppose $A \in \mathbb{R}^{m \times n_d}$ satisfies the RRCP with respect to $G$ with constant $\varepsilon$ and $G$ is such that each $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ satisfies the WDC with constant $\varepsilon$ for $i \in [d]$. Then for any $x \in \mathbb{R}^k \setminus \{0\}$ and $v_x \in \partial f(x)$,
\[
\|v_x\| \leq \frac{Cd}{2d} \max(\|x\|, \|x^*\|) + \frac{2}{2^{d/2}} \|\eta\|
\]
where $C$ is a numerical constant.
The second shows that $h_x$ is Lipschitz with respect to $x$ for points away from the origin.

**Lemma 2.2.** For all $x, z \neq 0$, we have that

$$
\|h_x - h_z\| \leq \left(\frac{2d^2 + (10\pi + 8)d + 20\pi}{\pi^2 2^d}\right) \max\left(\frac{1}{\|x\|}, \frac{1}{\|z\|}\right) + \frac{1}{2^d} \|x - z\|.
$$

In particular, if $x, z \notin B(0, r \|x_*\|)$ for some $r > 0$, then

$$
\|h_x - h_z\| \leq \left(\frac{2d^2 + (10\pi + 8)d + 20\pi}{r \pi^2 2^d} + \frac{1}{2^d}\right) \|x - z\|.
$$

The third states that for any non-zero $x \in \mathbb{R}^k$ and any $v_x \in \partial f(x)$, $h_x$ approximates $v_x$ well.

**Lemma 2.3.** Fix $\epsilon > 0$ such that $\epsilon < d^{-d}(1/16\pi)^2$ If $A \in \mathbb{R}^{m \times n_d}$ satisfies the RRCP with respect to $G$ with constant $\epsilon$ and $G$ is such that each $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ satisfies the WDC with constant $\epsilon$ for $i \in [d]$, then for any $x \neq 0$ and $v_x \in \partial f(x)$

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

where $K$ is a universal constant.

Each of these results are proven in Section 2.5.2. We are now ready to state and prove the main result of this section.

**Lemma 2.4.** Suppose Assumptions A1–A3 are satisfied and set $\beta := 4Kd^3 \sqrt{\epsilon} + 11\|\eta\|2^{d/2} / \|x_*\|$ where $K$ is a numerical constant. Let $A \in \mathbb{R}^{m \times n}$ satisfy the RRCP with respect to $G$ with constant $\epsilon$. Let $G$ be such that $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ satisfy the WDC with constant $\epsilon$ for all $i \in [d]$. Suppose that $x \notin S_\beta$ and $x \notin B(0, c_0 \|x_*\|)$ for some numerical constant $c_0$. Then for any $v_x \in \partial f(x)$, we have

$$
\frac{1}{3} \|v_x\| \geq \frac{Kd^3 \sqrt{\epsilon}}{2^d} \|x_*\|. \quad (2.4)
$$

Moreover, we have that for any $\lambda \in [0, 1]$,

$$
\|v_x - v_{\tilde{x}}\| \leq \frac{5}{6} \|v_x\| \quad (2.5)
$$

where $\tilde{x} = x - \lambda \alpha v_x$, $v_x \in \partial f(x)$ and $v_{\tilde{x}} \in \partial f(\tilde{x})$.

**Proof of Lemma 2.4.** By Lemma 2.2, we have that $h_x$ is Lipschitz for all $x \notin B(0, c_0 \|x_*\|)$, that is, there exists a numerical constant $L_{c_0}$ such that for any $x, z \notin B(0, c_0 \|x_*\|)$

$$
\|h_x - h_z\| \leq \frac{L_{c_0}d^2}{2^d} \|x - z\|.
$$
Moreover, Lemma 2.3 implies for any $x \neq 0$, we have

$$
\|v_x - h_x\| \leq \frac{Kd^3 \sqrt{\epsilon}}{2^d} \max(\|x\|, \|x_s\|) + \frac{2}{2^{d/2}} \|\eta\|
$$

for some numerical constant $K$. Hence

$$
\|v_x - \bar{v}_x\| \leq \|v_x - h_x\| + \|h_x - \bar{h}_x\| + \|h_x - v_x\|
$$

$$
\leq \frac{Kd^3 \sqrt{\epsilon}}{2^d} \max(\|\bar{x}\|, \|x_s\|) + \frac{Lc_0 d^2}{2^d} \|\bar{x} - x\| \tag{2.6}
$$

$$
+ \frac{Kd^3 \sqrt{\epsilon}}{2^d} \max(\|x\|, \|x_s\|) + \frac{4}{2^{d/2}} \|\eta\|
$$

$$
\leq \frac{Kd^3 \sqrt{\epsilon}}{2^d} \max(\|x - \lambda \alpha v_x\|, \|x_s\|) + \alpha \frac{Lc_0 d^2}{2^d} \|v_x\| \tag{2.7}
$$

$$
+ \frac{Kd^3 \sqrt{\epsilon}}{2^d} \max(\|x\|, \|x_s\|) + \frac{4}{2^{d/2}} \|\eta\|
$$

$$
\leq \frac{Kd^3 \sqrt{\epsilon}}{2^d} \max(\|x\| + \alpha \|v_x\|, \|x_s\|) + \alpha \frac{Lc_0 d^2}{2^d} \|v_x\| \tag{2.8}
$$

$$
+ \frac{Kd^3 \sqrt{\epsilon}}{2^d} \max(\|x\|, \|x_s\|) + \frac{4}{2^{d/2}} \|\eta\|
$$

$$
\leq \frac{Kd^3 \sqrt{\epsilon}}{2^d} \left(2 + \alpha \frac{Cd}{2^d}\right) \max(\|x\|, \|x_s\|) + \alpha \frac{Lc_0 d^2}{2^d} \|v_x\| + \frac{6}{2^{d/2}} \|\eta\| \tag{2.9}
$$

where we used Lemma 2.2 and Lemma 2.3 in the second inequality, the definition of $\bar{x}$ in the third inequality and Lemma 2.1 in the last inequality for some numerical constant $C$. Now, we lower bound $\|v_x\|$. Since $x \notin S_\beta$ we have that

$$
\|v_x\| \geq \|h_x\| - \|h_x - v_x\|
$$

$$
\geq \frac{1}{2^d} \max(\|x\|, \|x_s\|) \left(\beta - Kd^3 \sqrt{\epsilon} - 2\|\eta\| \frac{2^{d/2}}{\|x_s\|}\right)
$$

$$
= \frac{1}{2^d} \max(\|x\|, \|x_s\|) \left(3Kd^3 \sqrt{\epsilon} + 9\|\eta\| \frac{2^{d/2}}{\|x_s\|}\right) \tag{2.10}
$$

$$
\geq \frac{3Kd^3 \sqrt{\epsilon}}{2^d} \max(\|x\|, \|x_s\|) \tag{2.11}
$$
where we used the definition of $\beta$ and Lemma 2.3 in the second inequality. Note that this proves (2.4). Applying (2.10) to Equation (2.9), we attain

$$
\|\bar{v}_x - v_x\| \leq \frac{2}{3} \|v_x\| + \alpha \frac{Cd}{2^d} \cdot \frac{Kd^3 \sqrt{\epsilon}}{2^d} \max(\|x\|, \|x_\star\|) + \alpha \frac{L_{c_0} d^2}{2^d} \|v_x\|
$$

$$
\leq \frac{1}{3} \left(2 + \alpha \frac{Cd}{2^d}\right) \|v_x\| + \alpha \frac{L_{c_0} d^2}{2^d} \|v_x\|
$$

$$
\leq \left(\frac{2}{3} + \frac{\alpha}{3} \cdot \frac{\tilde{C}d^2}{2^d}\right) \|v_x\|
$$

$$
\leq \frac{5}{6} \|v_x\|
$$

where $\tilde{C} = C + L_{c_0}$. In the first inequality, we used (2.10). In the second inequality, we used (2.11). The last inequality follows by choosing $c_3$ in the upper bound $\alpha \leq c_3 \frac{d^2}{2^d}$ small enough so that $\alpha \frac{Cd^2}{3 \cdot 2^d} \leq \frac{1}{6}$. \[ \Box \]

2.4.2 Convergence to neighborhoods of $x_\star$ and $-\rho_d x_\star$

Using Lemma 2.4, we can now show that the iterates of our algorithm make sufficient progress at each step so that they eventually are in $S_\beta$ after a polynomial number of iterations.

**Lemma 2.5.** Suppose Assumptions A1–A3 are satisfied and set $\beta := 4Kd^3 \sqrt{\epsilon} + 11\|\eta\| 2^{d/2}/\|x_\star\|$ where $K$ is a numerical constant. Let $A \in \mathbb{R}^{m \times n}$ satisfy the RRCP with respect to $G$ with constant $\epsilon$. Let $G$ be such that $W_i \in \mathbb{R}^{n_i \times n_i - 1}$ satisfy the WDC with constant $\epsilon$ for all $i \in [d]$. For $x_t \not\in S_\beta$, we have

$$
f(x_{t+1}) - f(x_t) \leq -\alpha \frac{9K^2d^6\epsilon}{6(2^d)} \|x_\star\|^2.
$$

Moreover, there exists an $N \leq \frac{6f(x_0)(2^d)}{9K^2d^6\alpha \epsilon \|x_\star\|^2}$ such that $x_N \in S_\beta$ where $x_0$ is the initial iterate of our algorithm and $\alpha > 0$ is the step size.

**Proof of Lemma 2.5.** Recall that by Proposition 2.12, we may assume $x_t \not\in B(0, c_0 \|x_\star\|)$ where $c_0$ is a constant. We first consider the case when $\bar{x}_t = -x_t$. Then we must have that $f(\bar{x}_t) < f(x_t)$. Hence for any $v_{\bar{x}_t} \in \partial f(\bar{x}_t)$, we have

$$
f(x_{t+1}) - f(x_t) = f(x_{t+1}) - f(\bar{x}_t) + f(\bar{x}_t) - f(x_t) < f(\bar{x}_t) - \alpha v_{\bar{x}_t} - f(\bar{x}_t)
$$

where we used $f(\bar{x}_t) < f(x_t)$ and the definition of $x_{t+1}$ in the first inequality. Thus observe that it suffices to establish the inequality for $f(\bar{x}_t - \alpha v_{\bar{x}_t}) - f(\bar{x}_t)$ since this will also establish the case when $\bar{x}_t = x_t$.

Now, choose $v_{\bar{x}_t} \in \partial f(\bar{x}_t)$. By the generalized mean value theorem for the Clarke subdifferential (Theorem 8.13 in ref. [19]), there exists a $\lambda \in [0, 1]$ and $\tilde{v}_{\bar{x}_t} \in \partial f(\bar{x}_t)$ where $\tilde{x}_t = \bar{x}_t - \lambda \alpha v_{\bar{x}_t}$...
such that we have

\[
\begin{align*}
  f(\bar{x}_t - \alpha v_{\bar{x}_t}) - f(\bar{x}_t) &= \langle \bar{v}_{\bar{x}_t}, - \alpha v_{\bar{x}_t} \rangle \\
  &= \langle v_{\bar{x}_t}, - \alpha v_{\bar{x}_t} \rangle + \langle \bar{v}_{\bar{x}_t} - v_{\bar{x}_t}, - \alpha v_{\bar{x}_t} \rangle \\
  &\leq -\alpha \|v_{\bar{x}_t}\|^2 + \alpha \|v_{\bar{x}_t}\||\bar{v}_{\bar{x}_t} - v_{\bar{x}_t}\| \\
  &= -\alpha \|v_{\bar{x}_t}\|\left(\|v_{\bar{x}_t}\| - \|\bar{v}_{\bar{x}_t} - v_{\bar{x}_t}\|\right)
\end{align*}
\]

where we used the mean value theorem in the first equality.

We can now use our result from Section 2.4.1 to bound \(\|\bar{v}_{\bar{x}_t} - v_{\bar{x}_t}\|\) and \(\|v_{\bar{x}_t}\|\) from above and below, respectively. Observe that by Lemma 2.4, we have

\[
\begin{align*}
f(\bar{x}_t - \alpha v_{\bar{x}_t}) - f(\bar{x}_t) &\leq -\alpha \|v_{\bar{x}_t}\|\left(\|v_{\bar{x}_t}\| - \|\bar{v}_{\bar{x}_t} - v_{\bar{x}_t}\|\right) \\
&\leq -\alpha \left(1 - \frac{5}{6}\right)\|v_{\bar{x}_t}\|^2 \\
&= -\frac{1}{6}\alpha\|v_{\bar{x}_t}\|^2
\end{align*}
\]

where we used (2.5) in the second inequality. But by our lower bound on \(\|v_{\bar{x}_t}\|\), we have

\[
\begin{align*}
f(\bar{x}_t - \alpha v_{\bar{x}_t}) - f(\bar{x}_t) &\leq -\frac{1}{6}\alpha\|v_{\bar{x}_t}\|^2 \leq -\alpha \frac{9K^2d^6\epsilon}{6(2^2d)}\|x_0\|^2
\end{align*}
\]

where we used (2.4) in the second inequality. Hence there are at most \(\frac{6f(x_0)(2^4d)}{9K^2d^6\alpha\|x_0\|^2}\) iterations for which \(x_t \notin S_\beta\) where \(x_0\) is the initial iterate of our algorithm. Thus there exists a natural number \(N\) such that \(N \leq \frac{6f(x_0)(2^4d)}{9K^2d^6\alpha\|x_0\|^2}\) and \(x_N \in S_\beta\). □

### 2.4.3 Convergence to neighborhood of \(x_\ast\)

We now show that if any iterate is in \(S_\beta\), then the negation step of the algorithm (Steps 3–7) ensures that our iterates will now be in a neighborhood of \(x_\ast\) as opposed to a neighborhood of \(-\rho_d x_\ast\). We will use the following result that \(S_\beta\) is contained in the union of neighborhoods of the true solution and a negative multiple thereof if \(\beta\) is sufficiently small:

**Lemma 2.6.** If \(0 < 24\pi^2d^{6}\sqrt{\beta} \leq 1\), then

\[
S_\beta \subset B(x_\ast, 70000\pi^2d^9\beta\|x_\ast\|) \cup B(-\rho_d x_\ast, 77422\pi^2d^{12}\sqrt{\beta}\|x_\ast\|).
\]

We also need the following lemma which shows that the objective function is smaller near \(x_\ast\) than near \(-\rho_d x_\ast\):

**Lemma 2.7.** Fix \(0 < \epsilon < 1/(16\pi^2d^2)\) and suppose Assumption A3 is satisfied. Suppose that \(A \in \mathbb{R}^{m \times n_d}\) satisfies the RRCP with respect to \(G\) with constant \(\epsilon\) and \(G\) is such that each \(W_i \in \mathbb{R}^{n_i \times n_i-1}\)
satisfies the WDC with constant \( \varepsilon \). Then for any \( \phi_d \in [\rho_d, 1] \), we have that
\[
f(x) < f(z) \tag{2.12}
\]
for all \( x \in B(\phi_d x_* , r_1 d^{-12}\| x_* \|) \) and \( z \in B(-\phi_d x_* , r_1 d^{-12}\| x_* \|) \) where \( r_1 \) is a universal constant.

These results are proven in Section 2.5.3. The main result of this section is as follows.

**Lemma 2.8.** Suppose Assumptions A1–A3 are satisfied and set \( \beta := 4Kd^3\sqrt{\varepsilon} + 11\| \eta \|^{2d/2}/\| x_* \| \) where \( K \) is a numerical constant. Let \( A \in \mathbb{R}^{m \times n} \) satisfy the RRCP with respect to \( G \) with constant \( \varepsilon \). Let \( G \) be such that \( W_i \in \mathbb{R}^{n_i \times n_i-1} \) satisfy the WDC with constant \( \varepsilon \) for all \( i \in [d] \). If \( x_t \in S_\beta \), then there exists an \( N \leq \frac{6f(x_0)(2d)}{9K^2d^6\varepsilon\| x_* \|^2} \) such that \( x_N \in S_\beta^+ \), that is,
\[
\| x_N - x_* \| \leq C_5d^{12}\sqrt{\varepsilon}\| x_* \| + C_6d^92^{d/2}\| \eta \|
\]
where \( C_5 \) and \( C_6 \) are numerical constants.

**Proof of Lemma 2.8.** Suppose \( x_t \in S_\beta \). We require \( \beta \) to satisfy the assumption of Lemma 2.6 and for \( S_\beta \) to be contained in the balls of radius \( r_1 d^{-12}\| x_* \| \) from Lemma 2.7. Recall that by assumption \( 0 < \varepsilon < c_1d^{-102} \) and \( \| \eta \| \leq \frac{c_2\| x_* \|}{2d/2d^{48}} \) for some constants \( c_1 \) and \( c_2 \). Choosing \( c_1 \) and \( c_2 \) sufficiently small enough, we can have that
\[
\beta = 4Kd^3\sqrt{\varepsilon} + \frac{11\| \eta \|^{2d/2}/\| x_* \|}{\| x_* \|- c_1\frac{2d}{2d^{48}}}
\]

Hence \( \beta \) satisfies the assumptions of Lemma 2.6 and \( 77422\pi^2d^{12}\sqrt{\beta}\| x_* \| \leq r_1 d^{-12}\| x_* \| \). Note that this implies \( S_\beta^+ \subset B(x_* , r_1 d^{-12}\| x_* \|) \) while \( S_\beta^- \subset B(-\rho_d x_* , r_1 d^{-12}\| x_* \|) \). Therefore, we can apply Equation (2.12) in Lemma 2.7 so that for any \( z \in S_\beta^- \) and \( x \in S_\beta^+ \), \( f(x) < f(z) \).

We now show that the iterates eventually enter \( S_\beta^+ \). Since \( x_t \in S_\beta \), either \( x_t \in S_\beta^+ \) or \( x_t \in S_\beta^- \). If \( x_t \in S_\beta^+ \), then \( f(x_t) < f(-x_t) \) so \( \tilde{x}_t = x_t \in S_\beta^+ \). If \( x_t \in S_\beta^- \), then by Lemma 2.7 we have \( f(-x_t) < f(x_t) \) so \( \tilde{x}_t = -x_t \). If \( \tilde{x}_t \in S_\beta^+ \), then we are done. Otherwise, if \( \tilde{x}_t \notin S_\beta^+ \), then we can apply Lemma 2.5 to conclude there is an \( N \leq \frac{6f(x_0)(2d)}{9K^2d^6\varepsilon\| x_* \|^2} \) such that \( x_N \in S_\beta^- \) and for each \( k = t, t+1, ..., N-1 \), \( f(x_{k+1}) < f(x_k) \). Thus since \( f(x_N) < f(\tilde{x}_t) \) and \( f(\tilde{x}_t) < f(z) \) for any \( z \in S_\beta^- \), by Lemma 2.7, we must have that \( x_N \in S_\beta^- \). In either case, we must have that there is an \( N \leq \frac{6f(x_0)(2d)}{9K^2d^6\varepsilon\| x_* \|^2} \) such that \( x_N \in S_\beta^+ \). By the definition of \( S_\beta^+ \), this establishes the inequality
\[
\| x_N - x_* \| \leq C_5d^{12}\sqrt{\varepsilon}\| x_* \| + C_6d^92^{d/2}\| \eta \|
\]
for some numerical constants \( C_5 \) and \( C_6 \).
2.4.4 | Convergence to $x_*$ up to noise

Finally, we show that once in a neighborhood of $x_*$, the iterates of our algorithm will converge to $x_*$ up to the noise level in the measurements. We will use the following convexity-like property around the minimizer:

**Lemma 2.9.** Fix $0 < \varepsilon < 1/(200^4 d^6)$. Suppose that $A \in \mathbb{R}^{m \times n_d}$ satisfies the RRCP with respect to $G$ with constant $\varepsilon$ and $G$ is such that each $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ satisfies the WDC with constant $\varepsilon$ for $i \in [d]$. Then for all $x \in B(x_*, d \sqrt{\varepsilon ||x_*||})$ and any $v_x \in \partial f(x)$, we have

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

We now prove the following lemma.

**Lemma 2.10.** Suppose Assumptions A1–A3 are satisfied and set $\beta := 4 K d^3 \sqrt{\varepsilon} + 11 \|\eta\| 2d/2 \Vert x_* \Vert$ where $K$ is a numerical constant. Let $A \in \mathbb{R}^{m \times n}$ satisfy the RRCP with respect to $G$ with constant $\varepsilon$. Let $G$ be such that $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ satisfy the WDC with constant $\varepsilon$ for all $i \in [d]$. Suppose $x_N \in S_\beta^+$ for some $N \in \mathbb{N}$. Then for all $t \geq N$, we have that $\tilde{x}_t \in B(x_*, r_1 d^{-12} ||x_*||)$, $\tilde{x}_t = x_t$, and

$$||x_{t+1} - x_*|| \leq \tau^{t+1-N} ||x_N - x_*|| + \tilde{\sigma} \frac{2d/2}{d^2} \|\eta\|$$

where $\tau := 1 - \frac{7}{8} \frac{\alpha}{2d} \in (0, 1)$ and $\tilde{\sigma} := \frac{2c_3}{1 - \tau}$.

**Proof of Lemma 2.10.** Suppose $t = N$ so we have that $\tilde{x}_t = x_t \in S_\beta^+ \subset B(x_*, r_1 d^{-12} ||x_*||)$. As shown in Lemma 2.8, this inclusion holds by our assumptions on $\varepsilon$ and $\eta$. By Assumption A1, the requirements of Lemma 2.9 are met. Observe that for any $v_{\tilde{x}_t} \in \partial f(\tilde{x}_t)$, we have

$$||x_{t+1} - x_*|| = ||\tilde{x}_{t} - \alpha v_{\tilde{x}_t} - x_* + \frac{\alpha}{2d} (\tilde{x}_t - x_*) - \frac{\alpha}{2d} (\tilde{x}_t - x_*)||$$

$$\leq \left( 1 - \frac{\alpha}{2d} \right) ||\tilde{x}_t - x_*|| + \alpha \left\| v_{\tilde{x}_t} - \frac{1}{2d} (\tilde{x}_t - x_*) \right\|$$

$$\leq \left( 1 - \frac{\alpha}{2d} \right) ||\tilde{x}_t - x_*|| + \left( \frac{\alpha}{8} \right) \frac{1}{2d} ||\tilde{x}_t - x_*|| + \alpha \left( \frac{2}{2d/2} \|\eta\| \right)$$

$$= \left( 1 - \frac{7}{8} \frac{\alpha}{2d} \right) ||\tilde{x}_t - x_*|| + \alpha \frac{2}{2d/2} \|\eta\| \tag{2.13}$$

where we used Lemma 2.9 in the second inequality. Using $\alpha \leq c_3 \frac{2d}{d^2}$ and $\|\eta\| \leq \frac{c_2}{2d^4} \|x_*\|$ for sufficiently small constants $c_2$ and $c_3$, we have that if $\tilde{x}_t \in B(x_*, r_1 d^{-12} ||x_*||)$, then $x_{t+1} \in B(x_*, r_1 d^{-12} ||x_*||)$ so the iterates stay within a small ball around the minimizer. Hence Lemma 2.7 yields $\tilde{x}_{t+1} = x_{t+1}$. Repeatedly applying the above logic shows that for all $t \geq N$, $x_t \in B(x_*, r_1 d^{-12} ||x_*||)$ and $\tilde{x}_t = x_t$. 
Finally, using $\alpha \leq c_3 \frac{2^d}{d^2}$ in the second half of Equation (2.13) yields

$$\|x_{t+1} - x_*\| \leq \left(1 - \frac{7\alpha}{8d}\right)\|x_t - x_*\| + 2c_3 \frac{2^{d/2}}{d^2} \|\eta\| =: \tau \|x_t - x_*\| + 2c_3 \frac{2^{d/2}}{d^2} \|\eta\|$$

where $\tau := 1 - \frac{7\alpha}{8d}$. Choosing $c_3$ so that $c_3 < \frac{8}{7}$ implies $\tau \in (0, 1)$. Starting at $t = N$ and repeatedly applying this inequality, we attain

$$\|x_{t+1} - x_*\| \leq \tau^{t+1-N} \|x_N - x_*\| + (\tau^{t-N} + \tau^{t-N-1} + \cdots + 1)2c_3 \frac{2^{d/2}}{d^2} \|\eta\|$$

$$\leq \tau^{t+1-N} \|x_N - x_*\| + \frac{2c_3}{1 - \tau} \frac{2^{d/2}}{d^2} \|\eta\|$$

$$= : \tau^{t+1-N} \|x_N - x_*\| + \vartheta \frac{2^{d/2}}{d^2} \|\eta\|$$

where $\vartheta := \frac{2c_3}{1 - \tau}$. This completes the proof. \hfill $\square$

2.4.5 Final proof of Theorem 1.3

With all of the necessary lemmas proven, we bring them together to prove Theorem 1.3.

Proof of Theorem 1.3. Set $\beta := 4K\alpha^3 \sqrt{\xi} + 11\|\eta\|2^{d/2}/\|x_*\|$ where $K$ is a numerical constant. By Proposition 2.12, we may assume that our initial iterate $x_0 \notin B(0, c_0 \|x_*\|)$ for some numerical constant $c_0$. Then by Lemma 2.5, there is an $N \in \mathbb{N}$ such that $N \leq \frac{6\|x_0\|2^{d/2}}{9K^2d^6\alpha \|x_*\|^2}$ and $x_N \in S_\beta$. Then Lemma 2.8 implies there is an $N \leq 2N$ such that $x_N \in S^+_\beta$ which establishes inequality (1.7). Finally, Lemma 2.10 establishes inequality (1.8) for any $t \geq N$. Inequality (1.9) follows by using (1.8) and the following result with $j = d$ which established Lipschitz continuity of $G$ for $x$ within a neighborhood of $x_*$:

Lemma 2.11 (Lemma A.8 in ref. [40]). Suppose $0 < \varepsilon < 1/(200^4d^6)$, $x \in B(x_*, d \sqrt{\xi\|x_*\|})$, and $G$ is such that each $W_i \in \mathbb{R}_{n_i \times n_{i-1}}$ satisfies the WDC with constant $\varepsilon$ for $i \in [d]$. Then we have that for all $j \in [d]$,

$$\left\|\Pi_{i=j}^1 W_{i,+x} x - \Pi_{i=j}^1 W_{i,+x} x_*\right\| \leq \frac{1.2}{2^{d/2}} \|x - x_*\|.$$

\hfill $\square$

2.5 Supplementary results

In the following sections, we provide proofs for auxiliary results that were in used in the four main lemmas used to establish Theorem 1.3. Section 2.5.1 focuses on proving that after a polynomial number of iterations, the iterates of our algorithm are all bounded away from zero.
Section 2.5.2 establishes supplementary results about controlling subgradients in Section 2.4.1. Then Section 2.5.3 establishes results concerning the zeros of $h_x$ and properties of the objective function used in Section 2.4.3. Lastly, Section 2.5.4 focuses on establishing the convexity-like property near the minimizer which is formalized in Section 2.4.4.

2.5.1 Iterates are eventually bounded away from zero

We focus on proving the following proposition:

**Proposition 2.12.** Fix $\varepsilon > 0$ such that $Kd^3 \sqrt{\varepsilon} \leq 1$ where $K$ is a universal constant. Suppose that $A \in \mathbb{R}^{m \times nd}$ satisfies the RRCP with respect to $G$ with constant $\varepsilon$ and $G$ is such that each $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ satisfies the WDC with constant $\varepsilon$ for $i \in [d]$. Suppose that the step size $\alpha$ and noise $\eta$ satisfy $0 < \alpha < \frac{2d}{104\pi(Cd+2c_2)}$ and $\|\eta\| \leq \frac{c_2\|x_*\|}{2d^2/2}$ where $C$ and $c_2$ are numerical constants. If $x_t \in B(0, \frac{1}{52\pi} \|x_*\|)$, then after at most $N_0 = \lceil \frac{1}{52\pi} \|x_*\| \rceil$ iterations, we have that for all $t > N_0$ and $\lambda \in [0, 1]$, $\lambda \bar{x}_t + (1 - \lambda)x_{t+1} \not\in B(0, \frac{1}{104\pi} \|x_*\|)$.

This result asserts that if an iterate of our algorithm lies within a ball of the origin, then after a polynomial number of steps, it will leave this region. To prove it, we require the following lemma that establishes certain properties of any subgradient $v_x \in \partial f(x)$ for points $x$ near the origin:

**Lemma 2.13.** Fix $\varepsilon > 0$ such that $Kd^3 \sqrt{\varepsilon} \leq 1$ where $K$ is a universal constant. Suppose that $A \in \mathbb{R}^{m \times nd}$ satisfies the RRCP with respect to $G$ with constant $\varepsilon$ and $G$ is such that each $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ satisfies the WDC with constant $\varepsilon$ for $i \in [d]$. Then for all $x \in B(0, \frac{1}{52\pi} \|x_*\|)$ and any $v_x \in \partial f(x)$, we have that

$$\langle x, v_x \rangle < 0 \text{ and } \|v_x\| \geq \frac{1}{2d^2/4\pi} \|x_*\|.$$ 

We are now ready to proceed with a proof of Proposition 2.12.

**Proof of Proposition 2.12.** Suppose that $x_t \in B(0, \frac{1}{52\pi} \|x_*\|)$. By Lemma 2.13, we have that $\bar{x}_t$ and the next iterate $x_{t+1} = \bar{x}_t - \alpha v_{\bar{x}_t}$ form an obtuse triangle for any $v_{\bar{x}_t} \in \partial f(\bar{x}_t)$. Thus

$$\|\bar{x}_{t+1}\|^2 = \|x_{t+1}\|^2 \geq \|\bar{x}_t\|^2 + \alpha^2 \|v_{\bar{x}_t}\|^2$$ 

$$\geq \|\bar{x}_t\|^2 + \alpha^2 \frac{1}{2d^2/4\pi^2} \|x_*\|^2$$

where the last inequality follows from Lemma 2.13. Thus the norm of the iterates will increase until after $N_0 = \lceil \frac{1}{52\pi} \|x_*\| \rceil$ iterations we have $x_{t+N_0} \not\in B(0, \frac{1}{52\pi} \|x_*\|)$.

Now consider $x_t \not\in B(0, \frac{1}{52\pi} \|x_*\|)$. We will show that for any $\lambda \in [0, 1]$, $\lambda \bar{x}_t + (1 - \lambda)x_{t+1} \not\in B(0, \frac{1}{104\pi} \|x_*\|)$. Note that for $x_t \not\in B(0, \frac{1}{52\pi} \|x_*\|)$, we have $\|\bar{x}_t\| = \|x_t\| \geq \frac{1}{52\pi} \|x_*\|$. Then observe
that for any $v_{\bar{x}_i} \in \partial f(\bar{x}_i)$, we have
\[
\alpha \| v_{\bar{x}_i} \| \leq \alpha \frac{1}{2^d} \max(\| \bar{x}_i \|, \| x_s \|) \left( Cd + \frac{2}{\| x_s \|} 2^{d/2} \| \eta \| \right)
\]
\[
\leq \alpha \frac{1}{2^d} \max(\| \bar{x}_i \|, \| x_s \|)(Cd + 2c_2)
\]
\[
\leq \frac{\alpha}{2^d} 52\pi \| \bar{x}_i \|(Cd + 2c_2)
\]
\[
\leq \frac{1}{2} \| \bar{x}_i \|
\]
where the first inequality follows by Lemma 2.1, the second by the assumption on the noise energy $\| \eta \| \leq \frac{c_2 \| x_s \|}{2^{d/2}}$, the third due to $x_i \notin B(0, \frac{1}{52\pi} \| x_s \|)$, and the last inequality follows by the assumption on $\alpha$. Thus since $x_{i+1} = \bar{x}_i - \alpha v_{\bar{x}_i}$, we have that $\lambda \bar{x}_i + (1 - \lambda)x_{i+1} \notin B(0, \frac{1}{104\pi} \| x_s \|)$ for any $\lambda \in [0, 1]$.

We now focus on proving Lemma 2.13. To show this, we first require the following angle concentration property of the map $y \mapsto A_y y$ for $y$ in the range of $G$.

**Lemma 2.14.** Fix $0 < \varepsilon < 1/(4L)$ where $L$ is the universal constant specified in the RRCP. Let $A \in \mathbb{R}^{m \times n}$ satisfy the RRCP with respect to $G$ with constant $\varepsilon$. Let $G$ be such that $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ satisfy the WDC with constant $\varepsilon$ for all $i \in [d]$. Then for all $x, z \in \mathbb{R}^k \setminus \{0\}$, the angle $\theta_1 := \angle(A_G(x)G(x), A_G(z)G(z))$ is well-defined and
\[
| \cos \theta_1 - \cos \varphi(\theta_d) | \leq 4L \varepsilon
\]
where $\theta_d = \angle(G(x), G(z))$ and $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ defined by
\[
\varphi(\theta) := \cos^{-1}\left( \frac{(\pi - 2\theta) \cos \theta + 2 \sin \theta}{\pi} \right).
\]

**Proof of Lemma 2.14.** Fix $x, z \in \mathbb{R}^k \setminus \{0\}$. We use the shorthand notation $\Lambda_x := \Pi_{i=d}^{1} W_{i,x,x}$ and $x_d := \Lambda_x x$. Note that the WDC implies that for sufficiently small $\varepsilon$, we have that $\Lambda_x x, \Lambda_x z \neq 0$. Hence we may assume, without loss of generality, that $\| \Lambda_x x \| = \| \Lambda_x z \| = 1$. Now define the following quantities:
\[
\delta_1 := \langle \Lambda_x x, (A_{x_d}^T A_{x_d} - \Phi_{x_d,z_d}) \Lambda_x z \rangle,
\]
\[
\delta_2 := \langle \Lambda_x x, (A_{x_d}^T A_{x_d} - I) \Lambda_x x \rangle,
\]
\[
\delta_3 := \langle \Lambda_x z, (A_{z_d}^T A_{z_d} - I) \Lambda_x z \rangle.
\]
Observe that by the RRCP, we have that $\max_{i=1,2,3} |\delta_i| \leq L \varepsilon$. Hence if $0 < \varepsilon < 1/L$,
\[
0 < 1 - L \varepsilon \leq \| A_{x_d} \Lambda_x x \|^2
\]
so \(\|A_{zd}A_{x}\| \neq 0\). The same conclusion holds for \(\|A_{zd}A_{z}\|\) so \(\theta_1\) is well-defined. Furthermore, note that

\[
\cos \theta_1 = \frac{\langle A_{x}x, A^{T}_{zd}A_{zd}A_{z}z \rangle}{\|A_{zd}A_{x}\|\|A_{zd}A_{z}\|}
\]

\[
= \frac{\langle A_{x}x, A^{T}_{zd}A_{zd}A_{z}z \rangle}{\sqrt{\langle A_{zd}A_{x}, A_{zd}A_{x} \rangle \langle A_{zd}A_{z}, A_{zd}A_{z} \rangle}}
\]

\[
= \frac{\langle A_{x}x, \Phi_{xd,zd}A_{z}z \rangle + \delta_1}{\sqrt{(\langle A_{zd}, A_{x}x \rangle + \delta_2)((\langle A_{zd}, A_{z}z \rangle + \delta_3)}}
\]

\[
= \frac{\langle A_{x}x, \Phi_{xd,zd}A_{z}z \rangle + \delta_1}{\sqrt{(1 + \delta_2)(1 + \delta_3)}}.
\]

Thus

\[
\left| \cos \theta_1 - \langle A_{x}x, \Phi_{xd,zd}A_{z}z \rangle \right| = \left| \frac{\langle A_{x}x, \Phi_{xd,zd}A_{z}z \rangle + \delta_1}{\sqrt{(1 + \delta_2)(1 + \delta_3)}} - \langle A_{x}x, \Phi_{xd,zd}A_{z}z \rangle \right|
\]

\[
\leq \left| \langle A_{x}x, \Phi_{xd,zd}A_{z}z \rangle \right| \left| 1 - \frac{1}{\sqrt{(1 + \delta_2)(1 + \delta_3)}} \right|
\]

\[
+ \frac{|\delta_1|}{\sqrt{(1 + \delta_2)(1 + \delta_3)}}
\]

\[
\leq 2 \left| 1 - \frac{1}{1 - L\varepsilon} \right| + \frac{L\varepsilon}{1 - L\varepsilon}
\]

\[
\leq \frac{3L\varepsilon}{1 - L\varepsilon} \leq 4L\varepsilon
\]

where we used \(\|\Phi_{xd,zd}\| \leq 2\) in the third inequality and \(L\varepsilon < 1/4\) in the last inequality. The proof concludes by noting that \(\langle A_{x}x, \Phi_{xd,zd}A_{z}z \rangle = \frac{1}{\pi}[(\pi - 2\theta_d)\cos \theta_d + 2 \sin \theta_d]\). \(\square\)

We also require upper bounds on quantities that will be useful throughout the remaining proofs.

**Lemma 2.15.** Fix \(0 < \varepsilon < 1/(48d)\). Let \(A \in \mathbb{R}^{m \times n}\) satisfy the RRCP with respect to \(G\) with constant \(\varepsilon\). Let \(G\) be such that \(W_i \in \mathbb{R}^{n_i \times n_{i-1}}\) satisfy the WDC with constant \(\varepsilon\) for all \(i \in [d]\). Then for any \(x \in \mathbb{R}^k\), we have

\[
\|A_{x}\|^2 \leq \frac{13}{12}2^{-d},
\]

\[
\|A_{zd}A_{x}\|^2 \leq (1 + L\varepsilon)\|A_{x}\|^2
\]
Proof of Lemma 2.15. For Equation (2.14), note that the WDC implies that \( \|W_{i,+}\|^2 \leq \frac{1}{2} + \varepsilon \) for each \( i \in [d] \) so
\[
\|\Lambda_x\|^2 \leq \prod_{i=1}^{d} \|W_{i,+}\|^2 \leq \left( \frac{1}{2} + \varepsilon \right)^d = \frac{1}{2^d} \left( 1 + 2\varepsilon \right)^d = \frac{1}{2^d} e^{d \log(1+2\varepsilon)} \leq \frac{1 + 4\varepsilon d}{2} \leq \frac{13}{12} \frac{2^d}{d}
\]
where we used the fact that \( \log(1 + u) \leq u \) and \( e^u \leq 1 + 2u \) for \( u < 1 \) while the last inequality follows by our assumption on \( \varepsilon: \varepsilon < 1/(48d) \).

For Equation (2.15), observe that by the RRCP and the local linearity of \( G \), we have that for sufficiently small \( z \in \mathbb{R}^k \),
\[
\left| \left\langle A_{xd} \Lambda x z, A_{xd} \Lambda x z \right\rangle - \left\langle \Lambda x z, \Lambda x z \right\rangle \right| \leq L \varepsilon \|\Lambda x\|^2 \|z\|^2
\]
which implies that
\[
\left| \left\langle A_{xd} \Lambda x z, A_{xd} \Lambda x z \right\rangle \right| \leq (1 + L \varepsilon) \|\Lambda x\|^2 \|z\|^2.
\]
Since this holds for any \( z \in \mathbb{R}^k \), we conclude \( \|A_{xd} \Lambda x\|^2 \leq (1 + L \varepsilon) \|\Lambda x\|^2 \). \( \square \)

Now we set out to prove Lemma 2.13.

Proof of Lemma 2.13. Suppose \( f \) is differentiable at \( x \) so that \( v_x \) is precisely the gradient of \( f \). We first show that \( \langle x, v_x \rangle < 0 \). Note that
\[
\langle x, v_x \rangle = (I) - (II) - (III)
\]
We will bound the first and third term from above and the second from below. We first focus on the second term as its proof will give us a result for the first term.

(II): For the second term, note that we can write it as
\[
\left\langle A_{xd} \Lambda x_{*d} \Lambda x_{*d}, x \right\rangle = \cos(\angle(A_{xd} x_d, A_{x_{*d}d} x_{*d})) \|A_{xd} x_d\| \|A_{x_{*d}d} x_{*d}\|.
\]
By Lemma 2.14, we have that
\[
\cos(\phi(\theta_d)) - 4L \varepsilon \leq \cos(\angle(A_{xd} x_d, A_{x_{*d}d} x_{*d})) \leq \cos(\phi(\theta_d)) + 4L \varepsilon
\]
where \( \theta_d = \angle(x_d, x_{*d}) \). Thus
\[
\left\langle A_{xd} \Lambda x_{*d} \Lambda x_{*d}, x \right\rangle \geq (\cos(\phi(\theta_d)) - 4L \varepsilon) \|A_{xd} x_d\| \|A_{x_{*d}d} x_{*d}\|. \tag{2.16}
\]
However, note that
\[
\cos(\phi(\theta)) = \frac{(\pi - 2\theta) \cos \theta + 2 \sin \theta}{\pi} \geq \frac{2}{\pi} \forall \theta \in [0, \pi]. \tag{2.17}
\]
Hence if \( \varepsilon < 1/(4L\pi) \), applying (2.17) to (2.16) we have that

\[
\langle \Lambda_T^T A_{x_d}^T A_{x_s,d} A_{x_s,x}, x \rangle \geq \frac{1}{\pi} \| A_{x_d} x_d \| \| A_{x_s,d} x_{s,d} \|. \tag{2.18}
\]

We now bound \( \| A_{x_d} x_d \| \): observe that by the RRCP,

\[
| \langle (A_{x_d}^T A_{x_d} - I)x_d, x_d \rangle \rangle \leq \varepsilon \| x_d \|^2 \implies (1 - \varepsilon) \| x_d \|^2 \leq \| A_{x_d} x_d \|^2 \leq (1 + \varepsilon) \| x_d \|^2
\]

which gives

\[
\sqrt{1 - \varepsilon} \| x_d \| \leq \| A_{x_d} x_d \| \leq \sqrt{1 + \varepsilon} \| x_d \|.
\]

By Equation (11) of ref. [38], we have that

\[
\left( \frac{1}{2} - \varepsilon \right)^{d/2} \| x \| \leq \| x_d \| \leq \left( \frac{1}{2} + \varepsilon \right)^{d/2} \| x \|.
\]

Hence we attain

\[
\sqrt{1 - \varepsilon} \left( \frac{1}{2} - \varepsilon \right)^{d/2} \| x \| \leq \| A_{x_d} x_d \| \leq \sqrt{1 + \varepsilon} \left( \frac{1}{2} + \varepsilon \right)^{d/2} \| x \|. \tag{2.19}
\]

Analogous bounds hold for \( \| A_{x_s,d} x_{s,d} \| \). Applying (2.19) to Equation (2.18), we conclude that

\[
\langle \Lambda_T^T A_{x_d}^T A_{x_s,d} A_{x_s,x}, x \rangle \geq \frac{1}{\pi} \| A_{x_d} x_d \| \| A_{x_s,d} x_{s,d} \|
\]

\[
\geq \frac{1}{\pi} (1 - \varepsilon) \left( \frac{1}{2} - \varepsilon \right)^d \| x \| \| x_s \|.
\]

If \( 2\varepsilon < 2/3 \), we further have \( (1/2 - \varepsilon)^d \geq (1 - 2\varepsilon)^2/2^d \geq 1/3(1/2^d) \). Then if \( \varepsilon \) is chosen such that \( 1 - L\varepsilon \geq 1/2 \), then we get

\[
\langle \Lambda_T^T A_{x_d}^T A_{x_s,d} A_{x_s,x}, x \rangle \geq \frac{1}{6\pi} \frac{1}{2^d} \| x \| \| x_s \|. \tag{2.20}
\]

This concludes the bound of the second term. We then proceed to bounding (I) and (III).

(\text{I}) Observe that by Equation (2.19) and our choice of \( \varepsilon \), we get

\[
\langle \Lambda_T^T A_{x_d}^T A_{x_d} A_{x,x}, x \rangle = \| A_{x_d} x_d \|^2 \leq (1 + L\varepsilon) \left( \frac{1}{2} + \varepsilon \right)^d \| x \|^2 \leq 2 \cdot \frac{13}{12} \frac{1}{2^d} \| x \|^2 = \frac{13}{6} \frac{1}{2^d} \| x \|^2.
\]

(\text{III}) Observe that

\[
\| A_{x_d} \Lambda_x \| \leq \sqrt{1 + L\varepsilon} \| \Lambda_x \| \leq \sqrt{\frac{13}{12} (1 + L\varepsilon)} \frac{1}{2^{d/2}} \leq \frac{2}{2^{d/2}} \tag{2.21}
\]
where we used (2.15) in the first inequality, (2.14) in the second inequality and our assumption on \(\varepsilon\) in the last inequality. Thus we attain

\[
|x, \Lambda^T \Lambda x, \eta| \leq \|x\| \|A_{x,d} \Lambda x\| \|\eta\| \leq \frac{2}{2^{d/2}} \|\eta\| \|x\| \leq \frac{2c_2}{2^d} \|x\| \|x_s\| \leq \frac{1}{2^{d/2}} \|x\| \|x_s\|
\]

where the third inequality follows by \(\|\eta\| \leq c_2 2^{d/2} \|x_s\|\) and the last inequality is due to \(c_2 < \frac{1}{2^{4\pi}}\).

Using our results for (I), (II), and (III), we conclude that

\[
\langle x, v_x \rangle = \langle \Lambda^T A_{x,d}^T A_{x,d} \Lambda x, x \rangle - \langle \Lambda^T A_{x,d}^T A_{x,s,d} \Lambda x, x_s, x \rangle - \langle \Lambda^T A_{x,d}^T \eta, x \rangle \\
\leq \frac{1}{2^d} \|x\| \left( \frac{13}{6} \|x\| + \frac{1}{12\pi} \|x_s\| - \frac{1}{6\pi} \|x_s\| \right) \\
\leq \frac{1}{2^d} \|x\| \left( \frac{13}{6} \|x\| - \frac{1}{12\pi} \|x_s\| \right).
\]

Thus if \(\|x\| < \frac{1}{52\pi} \|x_s\|\), that is, \(x \in B(0, \frac{1}{52\pi} \|x_s\|)\), then

\[
\langle x, v_x \rangle \leq -\frac{1}{2^d} \frac{1}{24\pi} \|x\| \|x_s\| < 0.
\]

(2.22)

Lastly, observe that this gives

\[
\left\langle \frac{-x}{\|x\|}, v_x \right\rangle \geq \frac{1}{2^d} \frac{1}{24\pi} \|x_s\|.
\]

But by the Cauchy-Schwarz inequality, \(\left\langle \frac{-x}{\|x\|}, v_x \right\rangle \leq \|v_x\|\) so we obtain

\[
\|v_x\| \geq \frac{1}{2^d} \frac{1}{24\pi} \|x_s\|.
\]

(2.23)

When \(f\) is not differentiable at \(x\), we have that by Equation (2.3), we can write \(v_x = \sum_{\ell=1}^s c_\ell v_\ell\) where \(c_\ell \geq 0\) for \(\ell \in [s]\) and \(\sum_{\ell=1}^s c_\ell = 1\). Applying our result for differentiable points \(x\), we have that

\[
\langle x, v_x \rangle = \sum_{\ell=1}^s c_\ell \langle x, v_\ell \rangle \leq -\frac{1}{2^d} \frac{1}{24\pi} \|x\| \|x_s\| \sum_{\ell=1}^s c_\ell = -\frac{1}{2^d} \frac{1}{24\pi} \|x\| \|x_s\| < 0.
\]

For the lower bound on the norm of \(v_x\), note that by the Cauchy-Schwarz inequality and Equation (2.22), we have that

\[
\|v_x\| = \max_{\|u\|=1} \langle v_x, u \rangle \geq \left\langle v_x, \frac{-x}{\|x\|} \right\rangle = \sum_{\ell=1}^s c_\ell \left\langle v_\ell, \frac{-x}{\|x\|} \right\rangle \geq \frac{1}{2^d} \frac{1}{24\pi} \|x_s\| \sum_{\ell=1}^s c_\ell \\
= \frac{1}{2^d} \frac{1}{24\pi} \|x_s\|
\]

as desired.
2.5.2 | Proofs for Section 2.4.1

In this section, we focus on results that aided in establishing Lemma 2.4 in Section 2.4.1. The first result concerns a bound on the norm of our descent direction (Lemma 2.1). The second is that $h_x$ is Lipschitz with respect to $x \in \mathbb{R}^k$ outside of a ball of the origin (Lemma 2.2) and the third is that for all $x \in \mathbb{R}^k$, $h_x$ approximates any $v_x \in \partial f(x)$ (Lemma 2.3). Prior to beginning the proof of Lemma 2.1, we outline some notation. For $x \neq 0$, set $\psi_{d,x} := \frac{\pi - 2\bar{\theta}_{d,x}}{\pi}$, and $\zeta_{j+1,x} := \prod_{i=j}^{d-1} \frac{\pi - \bar{\theta}_{j+1,x}}{\sin \frac{\theta_{i,x}}{\pi} \zeta_{i+1,x}}$. Based on this notation, $h_x$ can be written as

$$h_x = \frac{1}{2^d} \left[ \psi_{d,x} \|x^*\| \hat{x}^* + \left( \|x\| - \|x^*\| \left( \frac{2 \sin \bar{\theta}_{d,x}}{\pi} + \psi_{d,x} \sum_{i=0}^{d-1} \frac{\sin \bar{\theta}_{i,x} \zeta_{i+1,x}}{\pi} \right) \right) \hat{x} \right].$$

In the remaining proofs, a number of results concerning properties of $\bar{\theta}_{i,x}$ and $\bar{\theta}_i$ will be useful. The following lemma records these results:

**Lemma 2.16** Bounds from Lemma 10 in ref. [38]. For $x \neq 0$, let $\bar{\theta}_{0,x} := \angle(x, x^*)$ and $\bar{\theta}_{i,x} := g(\bar{\theta}_{i-1,x})$ for $i \in [d]$. Let $\bar{\theta}_0 := \pi$ and $\bar{\theta}_i = g(\bar{\theta}_{i-1})$ for $i \in [d]$. Then the following all hold:

$$\left| \prod_{i=0}^{d-1} \frac{\pi - \bar{\theta}_{i,x}}{\pi} \right| \leq 1,$$  \hspace{1cm} (2.24)

$$\prod_{i=0}^{d-1} \frac{\pi - \bar{\theta}_{i,x}}{\pi} \geq \frac{\pi - \bar{\theta}_{0,x}}{\pi d^3},$$  \hspace{1cm} (2.25)

$$\left| \sum_{i=0}^{d-1} \frac{\sin \bar{\theta}_{i,x}}{\pi} \left( \prod_{j=i+1}^{d-1} \frac{\pi - \bar{\theta}_{j,x}}{\pi} \right) \right| \leq \frac{d}{\pi} \sin \bar{\theta}_{0,x},$$  \hspace{1cm} (2.26)

$$\bar{\theta}_{0,x} = \pi + O_1(\delta) \implies \bar{\theta}_{i,x} = \bar{\theta}_i + O_1(i\delta),$$  \hspace{1cm} (2.27)

$$\bar{\theta}_{0,x} = \pi + O_1(\delta) \implies \left| \prod_{i=0}^{d-1} \frac{\pi - \bar{\theta}_{i,x}}{\pi} \right| \leq \frac{\delta}{\pi},$$  \hspace{1cm} (2.28)

$$\left| \frac{\pi - 2\bar{\theta}_{i,x}}{\pi} \right| \leq 1 \forall i \geq 0,$$  \hspace{1cm} (2.29)

$$\bar{\theta}_{d,x} \leq \cos^{-1} \left( \frac{1}{\pi} \right) \forall d \geq 2,$$  \hspace{1cm} (2.30)

$$\bar{\theta}_i \leq \frac{3\pi}{i + 3} \forall i \geq 0,$$  \hspace{1cm} (2.31)
\[ \tilde{\theta}_i \geq \frac{\pi}{i + 1} \forall i \geq 0. \] (2.32)

We first focus on proving Lemma 2.1.

**Proof of Lemma 2.1.** Suppose \( f \) is differentiable at \( x \). By (2.24) and (2.29), we have that
\[
\max(|\psi_{d,x}|, |\zeta_{i,x}|) \leq 1 \quad \text{for any } i = 0, \ldots, d.
\]
Hence we have the bound
\[
\|h_x\| \leq \frac{1}{2^d}
\left|
\psi_{d,x}||\zeta_{0,x}||x_*\| + \|x\| + \frac{2|\sin \tilde{\theta}_{d,x}|}{\pi}\|x_*\| + \|\psi_{d,x}\| \sum_{i=0}^{d-1} \frac{|\sin \tilde{\theta}_{i,x}|}{\pi}|\zeta_{i+1,x}||x_*\|
\right|
\leq \frac{4 + d/\pi}{2^d} \max(\|x\|, \|x_*\|).
\] (2.33)

Combining Equation (2.33) and Lemma 2.3, we attain
\[
\|v_x\| \leq \|h_x\| + \|h_x - v_x\|
\leq \frac{4 + d/\pi}{2^d} \max(\|x\|, \|x_*\|) + \frac{Kd^3\sqrt{\varepsilon}}{2^d} \max(\|x\|, \|x_*\|) + \frac{2}{2^{d/2}} \|\eta\|
\leq \frac{Cd}{2^d} \max(\|x\|, \|x_*\|) + \frac{2}{2^{d/2}} \|\eta\|
\] (2.34)

where in the last inequality we used \( Kd^3\sqrt{\varepsilon} \leq 1 \) and set \( C = 5 + 1/\pi \).

When \( f \) is not differentiable at \( x \), we have that by Equation (2.3), we can write \( v_x = \sum_{\ell = 1}^{s} c_{\ell} v_{\ell} \) where each \( c_{\ell} \geq 0 \) and \( \sum_{\ell = 1}^{s} c_{\ell} = 1 \). Applying (2.34) for differentiable points, we have that
\[
\|v_x\| \leq \sum_{\ell = 1}^{s} c_{\ell} \|v_{\ell}\|
\leq \sum_{\ell = 1}^{s} c_{\ell} \left( \frac{Cd}{2^d} \max(\|x\|, \|x_*\|) + \frac{2}{2^{d/2}} \|\eta\| \right)
= \frac{Cd}{2^d} \max(\|x\|, \|x_*\|) + \frac{2}{2^{d/2}} \|\eta\|.
\]

We now show that \( h_x \) is Lipschitz for \( x \) outside of a ball of the origin.

**Proof of Lemma 2.2.** Throughout the proof, we will use the following result from Lemma 5.1 in ref. [40]:
\[
|\tilde{\theta}_{0,x} - \tilde{\theta}_{0,z}| \leq 4 \max \left( \frac{1}{\|x\|}, \frac{1}{\|z\|} \right) \|x - z\|.
\] (2.35)
For any $x, z \neq 0$, we have that

$$
\|h_x - h_z\| \leq \frac{1}{2d} \left| \psi_{d,x} \zeta_{0,x} - \psi_{d,z} \zeta_{0,z} \right| \|x\| + \frac{1}{2d} \|x - z\| + \frac{1}{2d} \left| \frac{2 \sin \tilde{\theta}_{d,x}}{\pi} \hat{x} - \frac{2 \sin \tilde{\theta}_{d,z}}{\pi} \hat{z} \right|
$$

(I)

$$
+ \frac{\|x\|}{2d} \left| \psi_{d,x} \sum_{i=0}^{d-1} \frac{\sin \tilde{\theta}_{i,x}}{\pi} \zeta_{i+1,x} \hat{x} - \psi_{d,z} \sum_{i=0}^{d-1} \frac{\sin \tilde{\theta}_{i,z}}{\pi} \zeta_{i+1,z} \hat{z} \right|
$$

(II)

$$
+ \frac{\|x\|}{2d} \left| \psi_{d,x} \sum_{i=0}^{d-1} \frac{\sin \tilde{\theta}_{i,x}}{\pi} \zeta_{i+1,x} \hat{x} - \psi_{d,z} \sum_{i=0}^{d-1} \frac{\sin \tilde{\theta}_{i,z}}{\pi} \zeta_{i+1,z} \hat{z} \right|
$$

(III)

We will focus on bounding each of the individual quantities.

(I): The triangle inequality gives $|\psi_{d,x} \zeta_{d,x} - \psi_{d,z} \zeta_{d,z}| \leq |\psi_{d,x}| |\zeta_{d,x} - \zeta_{d,z}| + |\zeta_{d,z}| |\psi_{d,x} - \psi_{d,z}|$.

By (2.24) and (2.29), we have $\max\{|\psi_{d,x}|, |\zeta_{d,x}|\} \leq 1$ for all $x \neq 0$. In addition,

$$
|\psi_{d,x} - \psi_{d,z}| = \frac{2}{\pi} |\tilde{\theta}_{d,x} - \tilde{\theta}_{d,z}|.
$$

(2.36)

Since $g'(\theta) \in [0, 1]$ for all $\theta \in [0, \pi]$ and $\tilde{\theta}_{i,x} = g(\tilde{\theta}_{i-1,x})$, we have that $|\tilde{\theta}_{i,x} - \tilde{\theta}_{i,z}| \leq |\tilde{\theta}_{i-1,x} - \tilde{\theta}_{i-1,z}|$. Repeatedly applying this inequality for each $i \in [d]$, we attain

$$
|\tilde{\theta}_{d,x} - \tilde{\theta}_{d,z}| \leq |\tilde{\theta}_{0,x} - \tilde{\theta}_{0,z}| \leq 4 \max\left( \frac{1}{\|x\|}, \frac{1}{\|z\|} \right) \|x - z\|
$$

(2.37)

where we used (2.35) in the last inequality. Hence combining (2.36) and (2.37), we get

$$
|\psi_{d,x} - \psi_{d,z}| = \frac{2}{\pi} |\tilde{\theta}_{d,x} - \tilde{\theta}_{d,z}| \leq \frac{8}{\pi} \max\left( \frac{1}{\|x\|}, \frac{1}{\|z\|} \right) \|x - z\|.
$$

Using the definition of $\zeta_{0,x}$, another application of (2.35) gives

$$
|\zeta_{0,x} - \zeta_{0,z}| \leq \frac{d}{\pi} |\tilde{\theta}_{0,x} - \tilde{\theta}_{0,z}| \leq \frac{4d}{\pi} \max\left( \frac{1}{\|x\|}, \frac{1}{\|z\|} \right) \|x - z\|.
$$

Combining our results, if $K_1 \coloneqq \frac{8+4d}{\pi}$ then

$$
(I) \leq \frac{\|x\|}{2d} K_1 \max\left( \frac{1}{\|x\|}, \frac{1}{\|z\|} \right) \|x - z\|
$$

(2.38)

(II): Observe that we have

$$
\frac{2 \sin \tilde{\theta}_{d,x}}{\pi} \hat{x} = \frac{2 \sin \tilde{\theta}_{d,x}}{\pi} \hat{x} + O_1\left( \frac{2}{\pi} \|\hat{x} - \hat{z}\| \right)
$$

$$
= \left( \frac{2 \sin \tilde{\theta}_{d,z}}{\pi} + O_1\left( \frac{2}{\pi} |\tilde{\theta}_{d,x} - \tilde{\theta}_{d,z}| \right) \right) \hat{z} + O_1\left( \frac{2}{\pi} \|\hat{x} - \hat{z}\| \right)
$$

$$
= \frac{2 \sin \tilde{\theta}_{d,z}}{\pi} \hat{z} + O_1\left( \frac{2}{\pi} \cdot 4 + \frac{4}{\pi} \right) \max\left( \frac{1}{\|x\|}, \frac{1}{\|z\|} \right) \|x - z\|
$$
where the second line follows from $| \sin \theta_1 - \sin \theta_2 | \leq |\theta_1 - \theta_2|$ and the third from (2.37) and $\| \hat{x} - \hat{z} \| \leq 2 \max(\frac{1}{\|x\|}, \frac{1}{\|z\|}) \|x - z\|$ Thus if $K_2 := \frac{12}{\pi}$, 

\[
(II) \leq \frac{\|x_s\|}{2^d} K_2 \max \left( \frac{1}{\|x\|}, \frac{1}{\|z\|} \right) \|x - z\| \tag{2.39}
\]

(III): The final term follows from

\[
\psi_{d,x} \sum_{i=0}^{d-1} \frac{1}{\pi} \sin \frac{\theta_i,x}{\pi} \xi_{i+1,x} \hat{x} = \psi_{d,z} \sum_{i=0}^{d-1} \frac{1}{\pi} \sin \frac{\theta_i,x}{\pi} \xi_{i+1,x} \hat{x} + O_1 \left( \frac{8d}{\pi^2} \max \left( \frac{1}{\|x\|}, \frac{1}{\|z\|} \right) \|x - z\| \right)
\]

\[
= \psi_{d,z} \sum_{i=0}^{d-1} \frac{1}{\pi} \sin \frac{\theta_i,z}{\pi} \xi_{i+1,x} \hat{x}
\]

\[
+ O_1 \left( \frac{4d}{\pi} + \frac{8d}{\pi^2} \right) \max \left( \frac{1}{\|x\|}, \frac{1}{\|z\|} \right) \|x - z\|
\]

\[
= \psi_{d,z} \sum_{i=0}^{d-1} \frac{1}{\pi} \sin \frac{\theta_i,z}{\pi} \xi_{i+1,x} \hat{x} + \frac{1}{\pi} \sum_{i=0}^{d-1} O_1 \left( \frac{d-i-1}{\pi} |\theta_i,x - \theta_i,z| \right)
\]

\[
+ O_1 \left( \frac{4d}{\pi} + \frac{8d}{\pi^2} \right) \max \left( \frac{1}{\|x\|}, \frac{1}{\|z\|} \right) \|x - z\|
\]

\[
= \psi_{d,z} \sum_{i=0}^{d-1} \frac{1}{\pi} \sin \frac{\theta_i,z}{\pi} \xi_{i+1,x} \hat{x}
\]

\[
+ O_1 \left( \frac{4d}{\pi} + \frac{8d}{\pi^2} \right) \max \left( \frac{1}{\|x\|}, \frac{1}{\|z\|} \right) \|x - z\|
\]

\[
= \psi_{d,z} \sum_{i=0}^{d-1} \frac{1}{\pi} \sin \frac{\theta_i,z}{\pi} \xi_{i+1,x} \hat{x}
\]

\[
+ O_1 \left( \frac{2d^2}{\pi^2} + \frac{4d}{\pi} + \frac{8d}{\pi^2} \right) \max \left( \frac{1}{\|x\|}, \frac{1}{\|z\|} \right) \|x - z\|
\]

\[
= \psi_{d,z} \sum_{i=0}^{d-1} \frac{1}{\pi} \sin \frac{\theta_i,z}{\pi} \xi_{i+1,x} \hat{x}
\]

\[
+ O_1 \left( \frac{2d}{\pi} + \frac{2d^2}{\pi^2} + \frac{4d}{\pi} + \frac{8d}{\pi^2} \right) \max \left( \frac{1}{\|x\|}, \frac{1}{\|z\|} \right) \|x - z\|
\]

where the first line follows from Equations (2.36) and (2.37) and using $| \sin \frac{\theta_i,x}{\pi} | \xi_{i+1,x} \| \| x \| \| z \| \leq 1$ for any $i = 0, \ldots, d - 1$ and $x$; the second line from (2.37); the third from $| \xi_{i+1,x} - \xi_{i+1,z} | \leq \frac{d-i-1}{\pi} |\theta_i,x - \theta_i,z| \xi_{i+1,x} \hat{x}$; the fourth from $|\theta_i,x - \theta_i,z| \leq |\theta_0,x - \theta_0,z|$, (2.35), and $\sum_{i=0}^{d-1} (d-i-1) = \frac{1}{2}(d-1)d$;
and the fifth from $| \sin \tilde{\gamma}_{i, z} \xi_{i+1, z}| \leq 1$ for all $i = 0, \ldots, d - 1$ and $\| \tilde{x} - \tilde{z} \| \leq 2 \max(\frac{1}{\| x \|}, \frac{1}{\| z \|}) \| x - z \|$. Combining our results, we have that if $K_3 := \frac{8d + 2d^2}{\pi^2} + \frac{6d}{\pi}$ then

$$(III) \leq \frac{\| x \|}{2^d} K_3 \max \left( \frac{1}{\| x \|}, \frac{1}{\| z \|} \right) \| x - z \|. \quad (2.40)$$

Thus for all $x, z \neq 0$, using Equations (2.38), (2.39), and (2.40), we conclude that

$$\| h_x - h_z \| \leq (I) + \frac{1}{2^d} \| x - z \| + (II) + (III)$$

$$\leq \frac{1}{2^d} \left( \| x \| (K_1 + K_2 + K_3) \max \left( \frac{1}{\| x \|}, \frac{1}{\| z \|} \right) + 1 \right) \| x - z \|$$

$$= \left( \frac{(2d^2 + (10\pi + 8)d + 20\pi)}{\pi^2} \frac{\| x \|}{2^d} \max \left( \frac{1}{\| x \|}, \frac{1}{\| z \|} \right) + \frac{1}{2^d} \right) \| x - z \|.$$

Then if $x, y \notin B(0, r)$, we can further conclude that

$$\| h_x - h_y \| \leq \left( \frac{(2d^2 + (10\pi + 8)d + 20\pi)}{r\pi^2} \frac{\| x \|}{2^d} + \frac{1}{2^d} \right) \| x - y \|.$$

□

We can now show that $h_x$ approximates any $v_x \in \partial f(x)$, which is formalized in Lemma 2.3. Prior to this proof, we define

$$w_x := \Lambda^T_x (\Lambda_x x - \Phi_{x_d, x_{s, d}} \Lambda_{x_{s, d}}). \quad (2.41)$$

The key idea is that the RRCP and WDC together imply $v_x \approx w_x$ and the WDC further implies $w_x \approx h_x$ which is shown in Lemma 2.18.

**Proof of Lemma 2.3.** Suppose $f$ is differentiable at $x$ so that $v_x = \bar{v}_x - q_x$ where $\bar{v}_x = \Lambda^T_x A_{x_d}^T (A_{x_d} \Lambda_x x - A_{x_{s, d}} \Lambda_{x_{s, d}})$ and $q_x = \Lambda^T_x A_{x_d}^T \eta$. Observe that

$$\| \bar{v}_x - w_x \| \leq \left\| \Lambda^T_x (A_{x_d}^T A_{x_d} - I_{n_d}) \Lambda_x \right\| \| x \| + \left\| \Lambda^T_x (A_{x_d}^T A_{x_{s, d}} - \Phi_{x_d, x_{s, d}}) \Lambda_{x_{s, d}} \right\| \| x_{s, d} \|.$$

By the local linearity of $G$, for sufficiently small $z \in \mathbb{R}^k$, we have $G(x + z) - G(x) = \Lambda x z$. Hence by the RRCP, we have for sufficiently small $z, \tilde{z} \in \mathbb{R}^k$,

$$| \langle (A_{x_d}^T A_{x_d} - I_{n_d}) \Lambda_x z, \Lambda_x z \rangle | \leq L \varepsilon \| \Lambda_x \|^2 \| z \|^2$$

and

$$| \langle (A_{x_d}^T A_{x_{s, d}} - \Phi_{x_d, x_{s, d}}) \Lambda_x z, \Lambda_x \tilde{z} \rangle | \leq L \varepsilon \| \Lambda_x \| \| \Lambda_{x_{s, d}} \| \| z \| \| \tilde{z} \|.$$
Since this holds for any $z, \tilde{z} \in \mathbb{R}^k$, we conclude that
\[ \left\| \Lambda_x^T (\Lambda_x^T A_{x,d} - I_n_d) \Lambda_x \right\| \leq L \| \Lambda_x \|^2 \text{ and } \left\| \Lambda_x^T (A_{x,d} A_{x,s,d} - \Phi_{x,d,x,s,d}) \Lambda_{x_*} \right\| \leq L \| \Lambda_x \| \| \Lambda_{x_*} \| . \]

This implies
\[
\left\| \tilde{v}_x - w_x \right\| \leq L \varepsilon \left( \| \Lambda_x \|^2 + \| \Lambda_x \| \| \Lambda_{x_*} \| \right) \max(\|x\|, \|x_*\|) \\
\leq 2L \varepsilon \left( \frac{1}{2} + \varepsilon \right)^d \max(\|x\|, \|x_*\|)
\]

where the last inequality follows by the WDC. Furthermore, by Lemma 2.18, we have that
\[ \|w_x - h_x\| \leq \frac{90d^3 \sqrt{\varepsilon}}{2d} \max(\|x\|, \|x_*\|). \]

Combining these two bounds, we have
\[
\left\| \tilde{v}_x - h_x \right\| \leq \left\| \tilde{v}_x - w_x \right\| + \|w_x - h_x\| \\
\leq \sqrt{\varepsilon} \left( 2L \frac{(1 + 2\varepsilon)^d}{2d} + 90 \frac{d^3}{2d} \right) \max(\|x\|, \|x_*\|) \\
\leq \sqrt{\varepsilon} K \frac{d^3}{2d} \max(\|x\|, \|x_*\|) \tag{2.42}
\]

for some universal constant $K$ where the third inequality follows since $2\varepsilon d \leq 1 \implies (1 + 2\varepsilon)^d \leq e^{2\varepsilon d} \leq 1 + 4\varepsilon d$ so choosing $\varepsilon < 1/(4d)$ implies $(1 + 2\varepsilon)^d \leq 2$. Lastly, to bound $\|q_x\|$, observe that
\[
\|q_x\| \leq \|A_{x,d} \Lambda_x \| \| \eta \| \leq \sqrt{1 + L \varepsilon \| \Lambda_x \| \| \eta \|} \leq \sqrt{\frac{13}{12} (1 + L \varepsilon)} \frac{1}{2d/2} \| \eta \| \leq \frac{2}{2d/2} \| \eta \|. \tag{2.43}
\]

where in the second inequality we used (2.15) and in the third inequality we used (2.14). The last inequality follows by choosing $\varepsilon$ such that $\sqrt{\frac{13}{12} (1 + L \varepsilon)} \leq 2$. Then we can combine (2.42) and (2.43) to obtain
\[
\left\| v_x - h_x \right\| \leq \left\| \tilde{v}_x - h_x \right\| + \|q_x\| \leq K \frac{d^3 \sqrt{\varepsilon}}{2d} \max(\|x\|, \|x_*\|) + \frac{2}{2d/2} \| \eta \|. 
\]

When $f$ is not differentiable at $x$, we can use (2.3) to write $v_x = \sum_{\ell=1}^{s} c_\ell v_\ell$ where $c_\ell \geq 0$ for $\ell \in [s]$. Moreover, note that for each $v_\ell$, there exists a direction $w_\ell$ such that $v_\ell = \lim_{\delta_\ell \to 0^+} \nabla f(x + \delta_\ell w_\ell)$ and $f$ is differentiable at $x + \delta_\ell w_\ell$ for sufficiently small $\delta_\ell > 0$. Appealing to the continuity of $h_x$ for $x \neq 0$, we obtain
\[
\left\| v_x - h_x \right\| \leq \sum_{\ell=1}^{s} c_\ell \| v_\ell - h_x \| = \sum_{\ell=1}^{s} c_\ell \left\| \lim_{\delta_\ell \to 0^+} \nabla f(x + \delta_\ell w_\ell) - h_x \right\|
\]
\[
\begin{align*}
&= \sum_{\ell=1}^{s} c_{\ell} \lim_{\delta_{\ell} \to 0^+} \| \nabla f(x + \delta_{\ell} w_{\ell}) - h_{x+\delta_{\ell} w_{\ell}} \|
\\
&= \sum_{\ell=1}^{s} c_{\ell} \lim_{\delta_{\ell} \to 0^+} \| v_{x+\delta_{\ell} w_{\ell}} - h_{x+\delta_{\ell} w_{\ell}} \|
\\
&\leq \sum_{\ell=1}^{s} c_{\ell} \left( K \frac{d^3 \sqrt{\epsilon}}{2^d} \max(\|x\|, \|x^*\|) + \frac{2}{2^d/2} \|\eta\| \right)
\\
&= K \frac{d^3 \sqrt{\epsilon}}{2^d} \max(\|x\|, \|x^*\|) + \frac{2}{2^d/2} \|\eta\|.
\end{align*}
\]

\[\square\]

We now establish a technical result that shows \( w_x \) is approximated by \( h_x \). Prior to this proof, we highlight the following result that summarizes some useful bounds from [38]:

**Lemma 2.17** (Results from Lemma 5 in ref. [38]). Fix \( 0 < \epsilon < d^{-4}(1/16\pi)^2 \) and let \( d \geq 2 \). Let \( W_i \) satisfy the WDC with constant \( \epsilon \) for \( i = 1, \ldots, d \). Then for any non-zero \( x, z \in \mathbb{R}^k \), the following hold:

\[
\begin{align*}
\left\| \Lambda^T_x \Lambda_z z - \tilde{h}_{x,z} \right\| &\leq 24 \frac{d^3 \sqrt{\epsilon}}{2^d} \|z\|, \\
\langle \Lambda_x x, \Lambda_z z \rangle &\geq \frac{1}{4\pi} \frac{1}{2^d} \|x\| \|z\|, \\
\left| \|z_d\| - \|z\| \right| &\leq 8d \epsilon \|z\| \|x\|, \\
|\theta_d - \overline{\theta}_d| &\leq 4d \sqrt{\epsilon}
\end{align*}
\]

where \( \theta_d := \angle(x_d, z_d) \), \( \overline{\theta}_d := g^{\circ d}(\angle(x, z)) \), and the vector \( \tilde{h}_{x,z} \) is defined as

\[
\tilde{h}_{x,z} := \frac{1}{2^d} \left[ \left( \prod_{i=0}^{d-1} \frac{\pi - \overline{\theta}_i}{\pi} \right) z + \sum_{i=0}^{d-1} \sin \overline{\theta}_i \left( \prod_{j=i+1}^{d-1} \frac{\pi - \overline{\theta}_j}{\pi} \right) \|z\| x \right]
\]

with \( \overline{\theta}_0 := \angle(x, z) \) and \( \overline{\theta}_i := g(\overline{\theta}_{i-1}) \) for \( i \in [d] \).

We now establish that \( w_x \) is approximated by \( h_x \).

**Lemma 2.18.** Fix \( 0 < \epsilon < d^{-4}(1/16\pi)^2 \). Let \( W_i \) satisfy the WDC with constant \( \epsilon \) for \( i = 1, \ldots, d \). For any non-zero \( x \in \mathbb{R}^k \), we have

\[
\|w_x - h_x\| \leq \frac{90d^3}{2^d} \sqrt{\epsilon} \max(\|x\|, \|x^*\|)
\]
where \( w_x \) is defined by (2.41).

**Proof.** Fix \( x \in \mathbb{R}^k \setminus \{0\} \) and set \( \vartheta_d := \angle(x_d, x_{*d}) \). Note that by the definition of \( \Phi_{x,w} \) and \( M_{z \rightarrow \hat{w}} \), \( w_x \) can be written as

\[
w_x = \Lambda_x^T \Lambda_x x - \frac{\pi - 2\vartheta_d}{\pi} \Lambda_x^T \Lambda_x x_{*} - \frac{2 \sin \vartheta_d}{\pi} \frac{\|\Lambda_x x_{*}\|}{\|\Lambda_x x\|} \Lambda_x^T \Lambda_x x
\]

where \( \vartheta_d := \angle(x_d, x_{*d}) \). Observe that

\[
\|w_x - h_x\| \leq \left\| \Lambda_x^T \Lambda_x x - \frac{1}{2d} x \right\| + \left\| \frac{\pi - 2\vartheta_d}{\pi} \Lambda_x^T \Lambda_x x_{*} - \frac{\pi - 2\overline{\vartheta}_d}{\pi} \hat{h}_{x,x_{*}} \right\|
\]

\[
+ \left\| \frac{2 \sin \vartheta_d}{\pi} \frac{\|\Lambda_x x_{*}\|}{\|\Lambda_x x\|} \Lambda_x^T \Lambda_x x - \frac{2 \sin \overline{\vartheta}_d}{\pi} \frac{\|x_{*}\|}{\|x\|} \frac{1}{2d} x \right\|
\]

where \( \overline{\vartheta}_d := g^{zd}(\angle(x, x_{*})) \) and \( \hat{h}_{x,x_{*}} \) is defined in (2.48).

We focus on bounding each individual quantity separately. For the first term, we have that by (2.44) in Lemma 2.17,

\[
\Lambda_x^T \Lambda_x x = \frac{1}{2d} x + O_1 \left( \frac{24d^3}{2d} \right) \sqrt{\epsilon} \max(\|x\|, \|x_{*}\|).
\]  

(2.49)

For the second term, observe that

\[
\frac{\pi - 2\overline{\vartheta}_d}{\pi} \Lambda_x^T \Lambda_x x_{*} = \frac{\pi - 2\overline{\vartheta}_d}{\pi} \hat{h}_{x,x_{*}} + O_1 \left( \frac{24d^3 \sqrt{\epsilon}}{2d} \|x\| \right)
\]

\[
= \left( \frac{\pi - 2\overline{\vartheta}_d}{\pi} + O_1 \left( \frac{24d^3 \sqrt{\epsilon}}{2d} \|x\| \right) \right) \hat{h}_{x,x_{*}} + O_1 \left( \frac{24d^3 \sqrt{\epsilon}}{2d} \|x\| \right)
\]

\[
= \frac{\pi - 2\overline{\vartheta}_d}{\pi} \hat{h}_{x,x_{*}} + O_1 \left( \frac{24d^3 \sqrt{\epsilon}}{2d} \|x\| \right)
\]  

(2.50)

where in the first equality we used (2.44) and in the second we used (2.47) and the fact that \( \|\hat{h}_{x,x_{*}}\| \leq 2^{-d} (1 + d/\pi) \|x_{*}\| \). For the final term, observe that

\[
\frac{2 \sin \vartheta_d}{\pi} \frac{\|\Lambda_x x_{*}\|}{\|\Lambda_x x\|} \Lambda_x^T \Lambda_x x = \frac{2 \sin \vartheta_d}{\pi} \frac{\|\Lambda_x x_{*}\|}{\|\Lambda_x x\|} \left( \frac{1}{2d} x + O_1 \left( \frac{24d^3 \sqrt{\epsilon}}{2d} \|x\| \right) \right)
\]

\[
= \frac{2 \sin \vartheta_d}{\pi} \frac{\|\Lambda_x x_{*}\|}{\|\Lambda_x x\|} \left( \frac{1}{2d} x + O_1 \left( \frac{24d^3 \sqrt{\epsilon}}{2d} \|x\| \right) \right)
\]

\[
= \frac{2 \sin \vartheta_d}{\pi} \left( \frac{|x_{*}|}{\|x\|} + O_1 \left( \frac{8d \epsilon \|x_{*}\|}{\|x\|} \right) \right) \frac{1}{2d} x
\]  

(2.51)
\[ + O_1 \left( \frac{4 \cdot 24d^3 \sqrt{\varepsilon}}{\pi^{2d}} \| x_* \| \right) \]
\[ = \frac{2 \sin \theta_d \| x_* \|}{\pi} \frac{1}{\| x \|} 2^d x + O_1 \left( \frac{16d \varepsilon}{\pi^{2d}} \| x \| + \frac{4 \cdot 24d^3 \sqrt{\varepsilon}}{\pi^{2d}} \| x_* \| \right) \]
\[ = \left( \frac{2 \sin \theta_d}{\pi} + O_1 \left( \frac{8d \sqrt{\varepsilon}}{\pi} \right) \right) \| x_* \| \frac{1}{\| x \|} \frac{1}{2^d} x \]
\[ + O_1 \left( \frac{16d \varepsilon}{\pi^{2d}} \| x \| + \frac{4 \cdot 24d^3 \sqrt{\varepsilon}}{\pi^{2d}} \| x_* \| \right) \]
\[ = \frac{2 \sin \theta_d}{\pi} \| x_*= \frac{1}{\| x \|} \frac{1}{2^d} x \quad (2.52) \]
\[ + O_1 \left( \frac{1}{2^d} \left( \frac{24d + 4 \cdot 24d^3}{\pi} \right) \right) \sqrt{\varepsilon} \max(\| x \|, \| x_* \|). \quad (2.54) \]

where the first line follows from (2.44); the second line from \( \sin \theta \leq 1 \); the third line from (2.46); and the fifth line from (2.47). Combining Equations (2.49), (2.50), and (2.54) achieves the desired result. \( \Box \)

2.5.3 | Proofs for Section 2.4.3

We first establish Lemma 2.6 which shows that the zeros of \( h_x \) occur near \( x_* \) and a particular negative multiple \(-\rho_d x_*\). Here the lemma is stated more precisely.

**Proposition 2.19.** Suppose \( \beta > 0 \) obeys \( 24\pi d^6 \sqrt{\beta} \leq 1 \) and define \( S_\beta \) as in (2.2). If \( x \in S_\beta \), then either

\[ |\bar{\theta}_{0,x} - \pi| \leq 82\pi d^4 \beta \quad \text{and} \quad \| x \| - \| x_* \| \leq 838\pi d^3 \beta \| x_* \| \]

or

\[ |\bar{\theta}_{0,x} - \pi| \leq 24\pi^2 d^4 \sqrt{\beta} \quad \text{and} \quad \| x \| - \rho_d \| x_* \| \leq 3517d^8 \sqrt{\beta} \| x_* \|. \]

*In particular, we have*

\[ S_\beta \subset B(x_*, 70000\pi^2 d^9 \beta \| x_* \|) \cup B(-\rho_d x_* , 77422\pi^2 d^{12} \sqrt{\beta} \| x_* \|). \]

*Additionally, \( \rho_d \to 1 \) as \( d \to \infty \).*

**Proof of Proposition 2.19.** Without loss of generality, let \( x_* = e_1 \) and \( \| x_* \| = 1 \) where \( e_1 \) is the first standard basis vector in \( \mathbb{R}^k \). We also let \( x = r \cos \bar{\theta}_0 e_1 + r \sin \bar{\theta}_0 e_2 \) where \( \bar{\theta}_0 = \angle(x_*, x_*). \) For
simplicity, we use the shorthand notation \( \bar{\theta}_i = \bar{\theta}_{i,x} \) for \( i \in [d] \). Set

\[
\xi = \left( \frac{\pi - 2\bar{\theta}_d}{\pi} \right) \left( \prod_{i=0}^{d-1} \frac{\pi - \bar{\theta}_i}{\pi} \right)
\]

and

\[
\alpha = \frac{2 \sin \bar{\theta}_d}{\pi} + \left( \frac{\pi - 2\bar{\theta}_d}{\pi} \right) \sum_{i=0}^{d-1} \frac{\sin \bar{\theta}_i}{\pi} \left( \prod_{j=i+1}^{d-1} \frac{\pi - \bar{\theta}_j}{\pi} \right).
\]

Note that we can write

\[
h_x = \frac{1}{2^d}(-\xi \hat{x}_* + (r - \alpha)\hat{x})
\]

Then if \( x \in S_\beta \), we have that

\[
| - \xi + \cos \bar{\theta}_0 (r - \alpha) | \leq \beta M \tag{2.55}
\]

and

\[
| \sin \bar{\theta}_0 (r - \alpha) | \leq \beta M \tag{2.56}
\]

where \( M := \max(r, 1) \).

To prove the Proposition, we first show that it is sufficient to only consider the small and large angle case. Then, we show that in the small and large angle case, \( x \approx x_* \) and \( x \approx -\rho_d x_* \), respectively. We begin by proving that \( \max(||x||, ||x_*||) \leq 6d \) for any \( x \in S_\beta \).

**Bound on maximal norm in \( S_\beta \):** It suffices to show that \( r \leq 6d \). Suppose \( r > 1 \) since if \( r \leq 1 \), the result is immediate. Then either \( |\sin \bar{\theta}_0| \geq 1/\sqrt{2} \) or \( |\cos \bar{\theta}_0| \geq 1/\sqrt{2} \). If \( |\sin \bar{\theta}_0| \geq 1/\sqrt{2} \) then (2.56) gives

\[
|r - \alpha| \leq \sqrt{2}\beta r \implies (1 - \sqrt{2}\beta)r \leq |\alpha|.
\]

But

\[
|\alpha| \leq \frac{2}{\pi} |\sin \bar{\theta}_d| + \left| \left( \frac{\pi - 2\bar{\theta}_d}{\pi} \right) \sum_{i=0}^{d-1} \frac{\sin \bar{\theta}_i}{\pi} \left( \prod_{j=i+1}^{d-1} \frac{\pi - \bar{\theta}_j}{\pi} \right) \right| \leq 1 + \frac{d}{\pi}
\]

where the second inequality used Equations (2.26) and (2.29). Thus

\[
r \leq \frac{1 + \frac{d}{\pi}}{1 - \sqrt{2}\beta} \leq 2 \left( 1 + \frac{d}{\pi} \right) \leq 2 + d \leq 2d
\]

provided \( \beta < 1/4 \) and \( d \geq 2 \). If \( |\cos \bar{\theta}_0| \geq 1/\sqrt{2} \), then (2.55) gives

\[
|r - \alpha| \leq \sqrt{2}(\beta r + |\xi|) \implies (1 - \sqrt{2}\beta)r \leq \sqrt{2}|\xi| + \alpha.
\]
But by (2.24),

\[ |\xi| = \left| \left( \frac{\pi - 2\bar{\theta}_d}{\pi} \right) \left( \prod_{i=0}^{d-1} \frac{\pi - \bar{\theta}_i}{\pi} \right) \right| \leq 1 \text{ since } \bar{\theta}_i \in [0, \pi/2] \forall i \geq 1. \]

Hence if \( \beta < 1/4 \),

\[ r \leq \frac{\sqrt{2} + 2d}{1 - \sqrt{2\beta}} \leq 2\sqrt{2} + 4d \leq \sqrt{2d + 4d} \leq 6d. \]

Thus in any case, \( r \leq 6d \implies M \leq 6d \).

We now show that it is sufficient to only consider the small angle case \( \bar{\theta}_0 \approx 0 \) and the large angle case \( \bar{\theta}_0 \approx \pi \).

**Sufficiency:** We have three possible cases:

- \( \sin \bar{\theta}_0 \leq 48\pi d^4 \beta \): Then we have that \( \bar{\theta}_0 = O_1(82\pi d^4 \beta) \) or \( \bar{\theta}_0 = \pi + O_1(82\pi d^4 \beta) \).
- \( \sin \bar{\theta}_0 > 48\pi d^4 \beta \) and \( |r - \alpha| \geq \sqrt{\beta M} \): Observe that due to Equation (2.56), we have that \( |r - \alpha| \leq \frac{\beta M}{\sin \bar{\theta}_0} \). Thus using this inequality in Equation (2.55), we have that

\[ |\xi| \leq \beta M + \frac{\beta M}{\sin \bar{\theta}_0} \leq \frac{2\beta M}{\sin \bar{\theta}_0} \leq \frac{2\beta M}{48\pi d^4 \beta} \leq \frac{12d}{48\pi d^4} = \frac{1}{4\pi} d^{-3} \tag{2.57} \]

where we used the assumption \( \sin \bar{\theta}_0 > 48\pi d^4 \beta \) in the second to last inequality and \( M \leq 6d \) in the last inequality. In addition, (2.30) implies

\[ |\pi - 2\bar{\theta}_d| \geq \left| \pi - 2\cos^{-1} \left( \frac{1}{\pi} \right) \right| \geq \frac{1}{2}. \tag{2.58} \]

Combining this inequality with (2.25) and (2.57), we obtain

\[ \frac{1}{2\pi} \left( \frac{\pi - \bar{\theta}_0}{\pi} \right) d^{-3} \leq |\xi| \leq \frac{1}{4\pi} d^{-3}. \]

From this, we can conclude that \( \bar{\theta}_0 \geq \frac{\pi}{2} \). Moreover, since \( |r - \alpha| \geq \sqrt{\beta M} \), then (2.56) implies that \( |\sin \bar{\theta}_0| \leq \sqrt{\beta} \) so we must have that \( \bar{\theta}_0 = \pi + O_1(2\sqrt{\beta}) \) since \( \bar{\theta}_0 \geq \frac{\pi}{2} \) and \( \beta < 1 \).

- \( |r - \alpha| \leq \sqrt{\beta M} \): Then (2.55) implies

\[ |\xi| \leq 2\sqrt{\beta M}. \]

But note that by (2.25),

\[ \xi = \left( \frac{\pi - 2\bar{\theta}_d}{\pi} \right) \left( \prod_{i=0}^{d-1} \frac{\pi - \bar{\theta}_i}{\pi} \right) \geq \frac{(\pi - 2\bar{\theta}_d)(\pi - \bar{\theta}_0)}{d^3\pi^2}. \]
In addition, since $|\pi - 2\tilde{\theta}_d| \geq \frac{1}{2}$ by (2.58), we have

$$|\xi| \geq \frac{|(\pi - 2\tilde{\theta}_d)(\pi - \tilde{\theta}_0)|}{d^3\pi^2} \geq \frac{|\pi - \tilde{\theta}_0|}{2d^3\pi^2}$$

which implies

$$|\pi - \tilde{\theta}_0| \leq 4d^3\pi^2 \sqrt{\beta}M \leq 24d^4\pi^2 \sqrt{\beta}.$$ 

Thus $\tilde{\theta}_0 = \pi + O_1(24d^4\pi^2 \sqrt{\beta})$.

Since only one of these situations can hold, it suffices to consider either the small angle case $\tilde{\theta}_0 = O_1(82\pi d^4 \beta)$ or the large angle case $\tilde{\theta}_0 = \pi + O_1(24d^4\pi^2 \sqrt{\beta})$. Now, we show that in the small angle case, $x \approx x_*$, while in the large angle case, $x \approx -\rho_d x_*$.

**Small angle case**: Assume $\tilde{\theta}_0 = O_1(\delta)$ where we set $\delta := 82\pi d^4 \beta$. Note that since $\tilde{\theta}_i \leq \tilde{\theta}_0 \leq \delta$ for each $i$, we have that

$$\prod_{i=0}^{d-1} \frac{\pi - \tilde{\theta}_i}{\pi} \geq \left(1 - \frac{\delta}{\pi}\right)^d = 1 + O_1\left(\frac{2d\delta}{\pi}\right)$$

provided $d\delta/\pi \leq 1/2$. Hence

$$\xi = \left(\frac{\pi - 2\tilde{\theta}_d}{\pi}\right)\left(\prod_{i=0}^{d-1} \frac{\pi - \tilde{\theta}_i}{\pi}\right) \geq \left(1 + O_1\left(\frac{2\delta}{\pi}\right)\right)\left(1 + O_1\left(\frac{2d\delta}{\pi}\right)\right)$$

where we used (2.27) in the second inequality. In addition, $|\sin \tilde{\theta}_d| \leq |\tilde{\theta}_d| \leq \delta$ and (2.26) imply that

$$\left|\sum_{i=0}^{d-1} \frac{\sin \tilde{\theta}_i}{\pi}\left(\prod_{j=i+1}^{d-1} \frac{\pi - \tilde{\theta}_j}{\pi}\right)\right| \leq \frac{d}{\pi} |\sin \tilde{\theta}_d| \leq d\delta.$$ 

Hence

$$\alpha = \frac{2\sin \tilde{\theta}_d}{\pi} + \left(\frac{\pi - 2\tilde{\theta}_d}{\pi}\right)\sum_{i=0}^{d-1} \frac{\sin \tilde{\theta}_i}{\pi}\left(\prod_{j=i+1}^{d-1} \frac{\pi - \tilde{\theta}_j}{\pi}\right)$$

$$= O_1\left(\frac{2\delta}{3\pi}\right) + \left(1 + O_1\left(\frac{2d\delta}{\pi}\right)\right)O_1(d\delta)$$

$$= O_1\left(\frac{(4 + 3d\pi + 6d^2)\delta}{3\pi}\right)$$
where we used $\delta < 1$ in the last equality. Thus since $| - \xi + \cos \tilde{\theta}_0 (r - \alpha)| \leq \beta M$ and $M \leq 6d$, we attain
\[
-(1 + O_1 \left( \frac{2\delta}{\pi} \right)) \left( 1 + O_1 \left( \frac{2d\delta}{\pi} \right) \right) + (1 + O_1(\delta)) \left( r + O_1 \left( \frac{4 + 3d\pi + 6d^2}{3\pi} \delta \right) \right) = O_1(6d\beta).
\]
Rearranging, this gives
\[
r - 1 = O_1 \left( \frac{2d\delta}{\pi} + \frac{2\delta}{\pi} + \frac{16d\delta^2}{\pi} + (\delta + 1) \left( \frac{4 + 3d\pi + 6d^2}{3\pi} \delta \right) \right) + O_1(12d\beta) + O_1(6d\beta)
\]
\[
= O_1 \left( \frac{(12d + 12 + 48d)\delta}{3\pi} + (2\varepsilon + 1)(4 + 3\pi d + 12d)\delta + 18d\beta \right)
\]
\[
= O_1(10d\delta + 18d\beta)
\]
\[
= O_1(838\pi d^5\beta)
\]
where we used $\delta < 1/2$ and $d \geq 2$ in the second to last equality and the definition of $\delta$ in the final equality.

**Large angle case:** Assume $\tilde{\theta}_0 = \pi + O_1(\delta)$ where $\delta := \frac{24d^4\pi^2}{\sqrt{\beta}}$. We first prove that $\alpha$ is close to $\rho_d$. Recall that $\tilde{\theta}_d = \tilde{\theta}_d + O_1(d\delta)$. Then by the mean value theorem:
\[
| \sin \tilde{\theta}_d - \sin \tilde{\theta}_d | \leq | \tilde{\theta}_d - \tilde{\theta}_d | \leq d\delta
\]
so $\sin \tilde{\theta}_d = \sin \tilde{\theta}_d + O_1(d\delta)$. Let $\Gamma_d := \sum_{i=0}^{d-1} \frac{\sin \tilde{\theta}_i}{\pi} \left( \prod_{j=i+1}^{d-1} \frac{\pi - \tilde{\theta}_j}{\pi} \right)$ and note that $\rho_d = \frac{2 \sin \tilde{\theta}_d}{\pi} + \left( \frac{\pi - 2\tilde{\theta}_d}{\pi} \right) \Gamma_d$. In [38], it was shown that if $d^2 \delta / \pi \leq 1$, then $|\Gamma_d| \leq d$ and
\[
\sum_{i=0}^{d-1} \frac{\sin \tilde{\theta}_i}{\pi} \left( \prod_{j=i+1}^{d-1} \frac{\pi - \tilde{\theta}_j}{\pi} \right) = \Gamma_d + O_1(3d^3\delta).
\]
By the condition, $d^2 \delta / \pi \leq 1$, we require $\sqrt{\beta} \leq \frac{1}{24\pi d^6}$. Thus for sufficiently small $\beta$, we have
\[
\alpha = \frac{2 \sin \tilde{\theta}_d}{\pi} + \left( \frac{\pi - 2\tilde{\theta}_d}{\pi} \right) \sum_{i=0}^{d-1} \frac{\sin \tilde{\theta}_i}{\pi} \left( \prod_{j=i+1}^{d-1} \frac{\pi - \tilde{\theta}_j}{\pi} \right)
\]
\[
= \frac{2 \sin \tilde{\theta}_d}{\pi} + O_1 \left( \frac{2d\delta}{\pi} \right) + \left( \frac{\pi - 2\tilde{\theta}_d}{\pi} + O_1 \left( \frac{2d\delta}{\pi} \right) \right) (\Gamma_d + O_1(3d^3\delta))
\]
\[
= \rho_d + O_1 \left( \frac{2d\delta}{\pi} \right) + \Gamma_d O_1 \left( \frac{2d\delta}{\pi} \right) + \left( \frac{\pi - 2\tilde{\theta}_d}{\pi} + O_1 \left( \frac{2d\delta}{\pi} \right) \right) O_1(3d^3\delta) + O_1 \left( \frac{6d^4\delta^2}{\pi} \right)
\]
\[
= \rho_d + O_1 \left( \frac{2d\delta}{\pi} \right) + O_1 \left( \frac{2d^2\delta}{\pi} \right) + O_1(3d^3\delta) + O_1 \left( \frac{6d^4\delta^2}{\pi} \right)
\]
\[
= \rho_d + O_1(7d^4\delta).
We now prove $r$ is close to $\rho_d$. Since $x \in S_\beta$,

$$| - \beta + \cos \bar{\delta}_0(r - \alpha)| \leq \beta M.$$ 

Also note that $|\beta| \leq \delta/\pi$ by (2.28). Since $\cos \bar{\delta}_0 = 1 + O_1(\delta^2/2)$, we have that

$$O_1(\delta/\pi) + (1 + O_1(\delta^3/2))(r - \rho_d + O_1(7d^4\delta)) = O_1(\beta M).$$

Using $r \leq 6d, \rho_d \leq 2d$, and $\delta = 24d^4\pi^2\sqrt{\beta} \leq 1$, we get

$$r - \rho_d + O_1\left(\frac{\delta^2}{2}\right)(r - \rho_d) + O_1(7d^4\delta) + O_1\left(\frac{7d^4\delta^3}{2}\right) = O_1(\beta M) + O_1\left(\frac{\delta}{\pi}\right)$$

$$\Rightarrow r - \rho_d = O_1\left(4d\delta^2 + 7d^4\delta + \frac{7d^4\delta^3}{2} + 6d\beta + \frac{\delta}{\pi}\right)$$

$$= O_1\left(6d\beta + \delta\left(4d + 7d^4 + \frac{7d^4\delta^3}{2} + 1\right)\right)$$

$$= O_1\left(\left(6d + 24d^4\pi^2\left(4d + \frac{21d^4}{2} + \frac{1}{\pi}\right)\right)\sqrt{\beta}\right)$$

$$= O_1(3517d^8\sqrt{\beta}).$$

Finally, to complete the proof we use the inequality

$$\|x - x_*\| \leq \||x| - \|x_*|| + (\|x_*\| + \|x\| - \|x_*\|)\bar{\delta}_0.$$ 

This inequality states that if a two dimensional point is known to be within $\Delta r$ of magnitude $r$ and an angle $\Delta \bar{\theta}$ away from 0, then it is at most a Euclidean distance of $\Delta r + (r + \Delta r)\Delta \bar{\theta}$ away from the point $(r,0)$ in polar coordinates. Thus for $\bar{\delta}_0 = O_1(82\pi d^4\beta)$, we have $r = 1 + O_1(838\pi d^5\beta)$ so

$$\|x - x_*\| \leq 838\pi d^5\beta + (1 + 838\pi d^5\beta)82\pi d^4\beta \leq 70000\pi^2 d^9\beta.$$ 

Then if $\bar{\delta}_0 = \pi + O_1(24d^4\pi^2\sqrt{\beta}), r = \rho_d + O_1(3517d^8\sqrt{\beta})$ so that

$$\|x + \rho_d x_*\| \leq 3517d^8\sqrt{\beta} + (\rho_d + 3517d^8\sqrt{\beta})24d^4\pi^2\sqrt{\beta} \leq 77422\pi^2 d^{12}\sqrt{\beta}.$$ 

Hence we attain

$$S_\beta \subset B(x_*, 70000\pi^2 d^9\beta) \cup B(-\rho_d x_*, 77422\pi^2 d^{12}\sqrt{\beta}).$$
The result that $\rho_d \to 1$ as $d \to \infty$ follows from the following facts: by (2.31), we have that $\tilde{\theta}_d \to 0$ as $d \to \infty$ which implies $\frac{2\sin \tilde{\theta}_d}{\pi} \to 0$ as $d \to \infty$. Moreover, in ref. [38], it was shown that

$$
\sum_{i=0}^{d-1} \frac{\sin \tilde{\theta}_i}{\pi} \left( \prod_{j=i+1}^{d-1} \frac{\pi - \tilde{\theta}_j}{\pi} \right) \to 1 \text{ as } d \to \infty.
$$

Hence

$$
\left( \frac{\pi - 2\tilde{\theta}_d}{\pi} \right) \sum_{i=0}^{d-1} \frac{\sin \tilde{\theta}_i}{\pi} \left( \prod_{j=i+1}^{d-1} \frac{\pi - \tilde{\theta}_j}{\pi} \right) \to 1 \text{ as } d \to \infty
$$

so $\rho_d \to 1$ as $d \to \infty$.

We now aim to show that the objective function value for points near the minimizer are lower than near the negative multiple which is formally stated in Lemma 2.7. We first define

$$
f_0(x) := \frac{1}{2} \left\| |AG(x)| - |AG(x_*)| \right\|^2
$$

which is the objective function without noise and $f_\eta(x) = f_0(x) - \langle |AG(x)| - |AG(x_*)|, \eta \rangle$. Then note that $f(x) = f_\eta(x) + \frac{1}{2} \|\eta\|^2$. We will first show that the objective function without noise can be closely approximated by a particular function $F$ which is defined by

$$
F(x) := \frac{1}{2d+1} \left( \|x\|^2 + \|x_*\|^2 \right) - \frac{1}{2d} \left( \frac{\pi - 2\tilde{\theta}_d}{\pi} \right) \left( \prod_{j=0}^{d-1} \frac{\pi - \tilde{\theta}_j}{\pi} \right) \langle x, x_* \rangle
$$

where $\tilde{\theta}_0 := \angle(x, y)$ and $\tilde{\theta}_i := g(\tilde{\theta}_{i-1})$ for $i \in [d]$. This result is formalized in the following lemma:

**Lemma 2.20.** Fix $0 < \varepsilon < 1/(16 \pi d^2)^2$. Suppose that $A \in \mathbb{R}^{m \times n_d}$ satisfies the RRCP with respect to $G$ with constant $\varepsilon$ and $G$ is such that each $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ satisfies the WDC with constant $\varepsilon$ for $i \in [d]$. Then we have that for all non-zero $x, x_* \in \mathbb{R}^k$:

$$
|f_0(x) - F(x)| \leq \frac{(L + 12)d^3 \sqrt{\varepsilon}}{2d} \|x\|^2 + \frac{(L + 12)d^3 \sqrt{\varepsilon}}{2d} \|x_*\|^2 + \frac{2L \varepsilon}{2d} \|x\| \|x_*\| + \frac{1}{2d} \left[ 24d^3 + 8d \left( 1 + \frac{d}{\pi} \right) + \frac{48d + 48d^3}{\pi} \right] \sqrt{\varepsilon} \|x\| \|x_*\|.
$$
Proof of Lemma 2.20. Fix $x, x_s \in \mathbb{R}^k \setminus \{0\}$. For notational simplicity, define

$$
\xi_{x, x_s} := \frac{\pi - 2\tilde{\theta}_d}{\pi} \left( \prod_{i=0}^{d-1} \frac{\pi - \tilde{\theta}_i}{\pi} \right) \langle x, x_s \rangle 
+ \left( \frac{\pi - 2\tilde{\theta}_d}{\pi} \sum_{i=0}^{d-1} \frac{\sin \tilde{\theta}_i}{\pi} \left( \prod_{j=i+1}^{d-1} \frac{\pi - \tilde{\theta}_i}{\pi} \right) + \frac{2 \sin \tilde{\theta}_d}{\pi} \right) \|x_s\| \|x\|.
$$

Then observe that $F$ can be written more compactly as $F(x) = \frac{1}{2d+1} (\|x\|^2 + \|x_s\|^2) - \frac{1}{2d+1} \xi_{x, x_s}$. Then the following bound shows we need to approximate three particular terms:

$$
|f_0(x) - F(x)| \leq \frac{1}{2} \left| \|A_{x_d} x_d\|^2 - \frac{1}{2d} \|x\|^2 \right| + \frac{1}{2} \left| \|A_{x_s,d} x_{s,d}\|^2 - \frac{1}{2d} \|x_s\|^2 \right| 
+ \left| \langle A_{x_d} x_d, A_{x_s,d} x_{s,d} \rangle - \frac{1}{2d} \xi_{x, x_s} \right|.
$$

Bounds on the first two terms follow directly by the RRCP and WDC in the following way. Note that

$$
\left| \|A_{x_d} x_d\|^2 - \frac{1}{2d} \|x\|^2 \right| \leq \left| \|A_{x_d} x_d\|^2 - \|x_d\|^2 \right| + \left| \|x_d\|^2 - \frac{1}{2d} \|x\|^2 \right|.
$$

Since $A$ satisfies the RRCP with respect to $G$, we have that

$$
\left| \|A_{x_d} x_d\|^2 - \|x_d\|^2 \right| \leq L\varepsilon \|A_x\|^2 \|x\|^2 \leq L\varepsilon \left( \frac{1}{2} + \varepsilon \right)^d \|x\|^2
$$

where the last inequality follows by the WDC. Then by (2.44), we have

$$
\left| \|x_d\|^2 - \frac{1}{2d} \|x\|^2 \right| \leq 24 \frac{d^3 \sqrt{\varepsilon}}{2d} \|x\|^2.
$$

Using these two bounds, we have

$$
\left| \|A_{x_d} x_d\|^2 - \frac{1}{2d} \|x\|^2 \right| \leq L\varepsilon \left( \frac{1}{2} + \varepsilon \right)^d \|x\|^2 + 24 \frac{d^3 \sqrt{\varepsilon}}{2d} \|x\|^2 
\leq \frac{(2L + 24)d^3 \sqrt{\varepsilon}}{2d} \|x\|^2
$$

(2.60)

since $(1 + 2\varepsilon)^d \leq e^{2\varepsilon d} \leq 1 + 4\varepsilon d \leq 2$ for $\varepsilon \leq 1/(4d)$. By the same logic, we have that

$$
\left| \|A_{x_s,d} x_{s,d}\|^2 - \frac{1}{2d} \|x_s\|^2 \right| \leq \frac{(2L + 24)d^3 \sqrt{\varepsilon}}{2d} \|x_s\|^2.
$$

(2.61)
For the last term, note that
\[
\left| \langle A_{x_d} x_d, A_{x_{s,d}} x_{s,d} \rangle - \frac{1}{2^d} \xi_{x,x_s} \right| \leq \left| \langle A_{x_d} x_d, A_{x_{s,d}} x_{s,d} \rangle - \langle \Phi_{x_d,x_{s,d}} x_d, x_{s,d} \rangle \right| \\
+ \left| \langle \Phi_{x_d,x_{s,d}} x_d, x_{s,d} \rangle - \frac{1}{2^d} \xi_{x,x_s} \right|.
\]

For the first term, the RRCP and WDC imply
\[
\left| \langle A_{x_d} x_d, A_{x_{s,d}} x_{s,d} \rangle - \langle \Phi_{x_d,x_{s,d}} x_d, x_{s,d} \rangle \right| \leq L \varepsilon \left( \frac{1}{2} + \varepsilon \right)^d \|x\| \|x_s\| \leq \frac{2L \varepsilon}{2^d} \|x\| \|x_s\| \quad (2.62)
\]
for \( \varepsilon \leq 1/(4d) \). For the second term, by the definition of \( \Phi_{z,w} \) and \( M_{z \leftrightarrow w} \), we have
\[
\left| \langle \Phi_{x_d,x_{s,d}} x_d, x_{s,d} \rangle - \frac{1}{2^d} \xi_{x,x_s} \right| \leq \|x\| \underbrace{\left| \frac{\pi - 2 \theta_d}{\pi} \Lambda_x^T \Lambda x_s - \frac{\pi - 2 \theta_d}{\pi} \hat{h}_{x,x_s} \right|}_{(I)} \\
+ \|x\| \underbrace{\left| \frac{2 \sin \theta_d \|x_{s,d}\|}{\pi} \Lambda_x^T \Lambda x - \frac{2 \sin \theta_d}{\pi} \|x\| \frac{1}{2^d} \|x\| \right|}_{(II)}
\]
where \( \hat{h}_{x,x_s} \) is defined in (2.48). It was shown in the proof of Lemma 2.18 that
\[
(I) \leq \frac{1}{2^d} \left( 24d^3 + \frac{8d}{\pi} \left( 1 + \frac{d}{\pi} \right) \right) \sqrt{\varepsilon \|x_s\|}
\]
and
\[
(II) \leq \frac{1}{2^d} \left( \frac{24d(1 + 2\varepsilon)^d + 48d^3}{\pi} \right) \sqrt{\varepsilon \|x_s\|}.
\]

Combining the results for (2.62), (I), and (II) we have
\[
\left| \langle A_{x_d} x_d, A_{x_{s,d}} x_{s,d} \rangle - \frac{1}{2^d} \xi_{x,x_s} \right| \leq \frac{2L \varepsilon}{2^d} \|x\| \|x_s\| \frac{1}{2^d} \left( 24d^3 + \frac{8d}{\pi} \left( 1 + \frac{d}{\pi} \right) \right) \sqrt{\varepsilon \|x_s\|} \\
+ \frac{1}{2^d} \left( \frac{24d(1 + 2\varepsilon)^d + 48d^3}{\pi} \right) \sqrt{\varepsilon \|x_s\|} \quad (2.63)
\]
Combining Equations (2.60), (2.61), and (2.63) achieves the desired result. \( \square \)

Now that we have established that the objective function without noise can be approximated by \( F \), we now show that \( F \) satisfies particular quadratic upper and lower bounds to establish the desired properties of the true objective function \( f \):

Lemma 2.21. Fix $0 < r < \frac{1}{4d^2\pi}$ and let $\kappa := \min_{d \geq 2} \rho_d > 0$. Then for any $\phi_d \in [\rho_d, 1]$, we have that

$$F(x) \leq \frac{\|x_*\|^2}{2d+1} \left( \phi_d^2 - 2\phi_d + \frac{45d}{\kappa^3} r \right) + \frac{\|x_*\|^2}{2d+1} \forall x \in B(\phi_d x_*, \|x_*\|). \tag{2.64}$$

$$F(z) \geq \frac{\|x_*\|^2}{2d+1} \left( \phi_d^2 - 2\rho_d \phi_d - 139d^4 r \right) + \frac{\|x_*\|^2}{2d+1} \forall z \in B(-\phi_d x_*, \|x_*\|). \tag{2.65}$$

Proof of Lemma 2.21. Define $\psi_d := \frac{\pi - \bar{\delta}_d}{\pi}, \xi_{i+1} := \prod_{j=i+1}^{d-1} \frac{\pi - \bar{\delta}_j}{\pi}$, and $\alpha_i := \frac{\sin \bar{\delta}_i}{\pi}$. Then note that we can write $F$ as

$$F(x) := \frac{1}{2d+1}(\|x\|^2 + \|x_*\|^2) - \frac{1}{2d} \left( \psi_d \xi_0(x, x_*) + \left( 2\alpha_d + \psi_d \sum_{i=0}^{d-1} \alpha_i \xi_{i+1} \right) \|x\| \|x_*\| \right).$$

Fix $x \in B(\phi_d x_*, r \|x_*\|)$. Then observe that we have $\theta_0 \leq \frac{\pi r}{2\phi_d}$ and $(\phi_d - r) \|x_*\| \leq \|x\| \leq (\phi_d + r) \|x_*\|$. Furthermore, $\cos \theta_0 \geq 1 - \frac{\theta_0^2}{2}$. Thus, we have the following bounds:

$$\psi_d \geq 1 - \frac{r}{\phi_d}, \xi_0 \geq \prod_{i=0}^{d-1} \left( 1 - \frac{r}{2\phi_d} \right), \text{ and } \cos \theta_0 \geq 1 - \frac{\pi^2 r^2}{8\phi_d^2}.$$

Hence we see that

$$F(x) - \frac{\|x_*\|^2}{2d+1} \leq \frac{\|x\|^2}{2d+1} - \frac{1}{2d} \psi_d \xi_0 \cos \theta_0 \|x\| \|x_*\|$$

$$\leq \frac{1}{2d+1}(\phi_d^2 + r^2) \|x_*\|^2 - \frac{1}{2d} \psi_d \xi_0 \cos \theta_0 (\phi_d - r) \|x_*\|^2$$

$$\leq \frac{\|x_*\|^2}{2d+1} \left( \phi_d^2 + 2r\phi_d + r^2 - 2 \left( 1 - \frac{r}{\phi_d} \right) \left( 1 - \frac{dr}{\phi_d} \right) \left( 1 - \frac{\pi^2 r^2}{8\phi_d^2} \right) \right).$$

where in the first inequality we used $2\alpha_d + \psi_d \sum_{i=0}^{d-1} \alpha_i \xi_{i+1} \geq 0$. Noting that $\phi_d \in [\rho_d, 1]$ and $r < 1$, with some algebra we attain

$$\phi_d^2 + 2r\phi_d + r^2 - 2 \left( 1 - \frac{r}{\phi_d} \right) \left( 1 - \frac{dr}{\phi_d} \right) \left( 1 - \frac{\pi^2 r^2}{8\phi_d^2} \right)$$

$$\leq \phi_d^2 - 2\phi_d + \frac{(8d + (2d + 1)\pi^2 + 7)r}{\phi_d^3}$$

$$\leq \phi_d^2 - 2\phi_d + \frac{45d}{\kappa^3} r$$
so we may conclude that for \( x \in B(\phi_d x_*, r \| x_* \|) \),

\[
F(x) - \frac{\| x_* \|^2}{2d+1} \leq \frac{\| x_* \|^2}{2d+1} \left( \phi_d^2 - 2\phi_d + \frac{45d}{\chi^3} r \right).
\]

Fix \( x \in B(-\phi_d x_*, r \| x_* \|) \). Then note that we have \( \pi - \vartheta_0 \leq \frac{\pi^2}{2} r \) and \( (\phi_d - r) \| x_* \| \leq \| x \| \leq (\phi_d + r) \| x_* \| \). Furthermore, for sufficiently small \( r > 0 \), we have that \( \langle x, x_* \rangle \leq 0 \) so that \( -\psi_d \vartheta_0 (x, x_*) \geq 0 \) (note that \( \vartheta_0 \leq \pi/2 \)). Thus

\[
F(x) - \frac{1}{2d+1} \| x_* \|^2 = \frac{1}{2d+1} \| x \|^2 - \frac{1}{2d} \left( \psi_d \vartheta_0 (x, x_*) + (\psi_d \sum_{i=0}^{d-1} \alpha_i \vartheta_{i+1} + 2\alpha_d) \| x \| \| x_* \| \right)
\]

\[
\geq \frac{1}{2d+1} (\phi_d - r)^2 \| x_* \|^2 - \frac{1}{2d} (\psi_d \sum_{i=0}^{d-1} \alpha_i \vartheta_{i+1} + 2\alpha_d)(\phi_d + r) \| x_* \|.
\]

Note that we have \( \vartheta_0 = \pi + O(1/(r \pi^2/2)) \). As shown in Proposition 2.19, if \( d^2 (r \pi^2/2)/\pi \leq 1 \), then we have that

\[
\psi_d \sum_{i=0}^{d-1} \alpha_i \vartheta_{i+1} + 2\alpha_d = \rho_d + O(1/(7d^4r \pi^2/2)).
\]

Hence we have

\[
F(x) - \frac{\| x_* \|^2}{2d+1} \geq \frac{1}{2d+1} (\phi_d - r)^2 \| x_* \|^2 - \frac{1}{2d} (\rho_d + 7\pi^2 d^4 r/2) (\phi_d + r) \| x_* \|^2
\]

\[
= \frac{\| x_* \|^2}{2d+1} (\phi_d^2 - 2r \phi_d + r^2)
\]

\[
- 2(\rho_d \phi_d + r \rho_d + 7\pi^2 d^4 r \phi_d/2 + 7\pi^2 d^4 r^2/2) \| x_* \|^2
\]

\[
\geq \frac{\| x_* \|^2}{2d+1} (\phi_d^2 - 2r - 2\rho_d \phi_d - 2r - 7\pi^2 d^4 r - 7\pi^2 d^4 r^2)
\]

\[
\geq \frac{\| x_* \|^2}{2d+1} (\phi_d^2 - 2\rho_d \phi_d - 139d^4 r)
\]

where we used the fact that \( \phi_d \in [\rho_d, 1] \) and \( 0 < r < 1 \). This completes the proof.

With this result, we are equipped to prove Lemma 2.7.

**Proof of Lemma 2.7.** By the same argument for (2.43), we have that \( \| \langle A_{x_d} A_* x, \eta \rangle \| \leq \frac{2}{2d/2} \| x \| \| \eta \| \) for any \( x \in \mathbb{R}^k \). Thus for \( x \in B(\phi_d x_*, \varphi \| x_* \|) \),

\[
|\langle AG(x) \rangle - |AG(x_*)\rangle, \eta) | \leq |\langle A_{x_d} A_* x, \eta \rangle | + |\langle A_{x_*} A_* x_*, \eta \rangle | \leq (\| x \| + \| x_* \|) \| \eta \| \frac{2}{2d/2} \| \eta \|
\]

\[
\leq (\varphi \| x_* \| + 2\| x_* \|) \| \eta \| \frac{2}{2d/2} \| \eta \|
\]
where we used the fact that \( \|x\| \leq (\phi_d + \varphi)\|x_*\| \leq (1 + \varphi)\|x_*\| \) in the last inequality.

Let \( \kappa := \min_{d \geq 2} \rho_d \). If \( x \in B(\phi_d x_*, \varphi \|x_*\|) \) and \( K_d := 24d^3 + 8d(1 + \frac{d}{\pi}) + \frac{48d^4 + 48d^3}{\pi} \), then Lemma 2.20 and Lemma 2.21 give

\[
\begin{align*}
f_\eta(x) & \leq F(x) + |f_0(x) - F(x)| + |(|AG(x)| - |AG(x_*)|, \eta)| \\
& \leq \frac{\|x_*\|^2}{2d+1} \left( \phi_d^2 - 2\phi_d + \frac{45d}{\kappa^3} \varphi \right) + \frac{\|x_*\|^2}{2d+1} + \frac{(L + 12)d^3\sqrt{\varepsilon}}{2d} \|x\|^2 \\
& \quad + \frac{(L + 12)d^3\sqrt{\varepsilon}}{2d} \|x_*\|^2 + \frac{2L\varepsilon}{2d} \|x\| \|x_*\| + \frac{1}{2d} K_d \sqrt{\varepsilon} \|x\| \|x_*\| \\
& \quad + (\varphi \|x_*\| + 2\|x_*\|) \frac{2}{2d/2} \|\eta\|
\end{align*}
\]

where \( K_d := 6(L + 12)d^3 + 8L + 4K_d \), in the second inequality we used \( \|x\| \leq (\phi_d + \varphi)\|x_*\| \), and in the last inequality we used \( \varepsilon < \sqrt{\varepsilon}, \rho_d \leq 1 \) and \( \varphi < 1 \).

Similarly, if \( z \in B(-\phi_d x_*, \varphi \|x_*\|) \), then

\[
\begin{align*}
f_\eta(z) & \geq F(z) - |f_0(z) - F(z)| - |(|AG(z)| - |AG(x_*)|, \eta)| \\
& \geq \frac{\|x_*\|^2}{2d+1} \left( \phi_d^2 - 2\phi_d + \frac{45d}{\kappa^3} \varphi \right) + \frac{\|x_*\|^2}{2d+1} - \frac{(L + 12)d^3\sqrt{\varepsilon}}{2d} \|x\|^2 \\
& \quad - \frac{(L + 12)d^3\sqrt{\varepsilon}}{2d} \|x_*\|^2 - \frac{2L\varepsilon}{2d} \|x\| \|x_*\| - \frac{1}{2d} K_d \sqrt{\varepsilon} \|x\| \|x_*\| \\
& \quad - (\varphi \|x_*\| + 2\|x_*\|) \frac{2}{2d/2} \|\eta\|
\end{align*}
\]

In sum, we have for \( x \in B(\phi_d x_*, \varphi \|x_*\|) \),

\[
f_\eta(x) \leq \frac{\|x_*\|^2}{2d+1} \left( 1 + \phi_d^2 - 2\phi_d + \frac{45d}{\kappa^3} \sqrt{\varepsilon} + K_d \sqrt{\varepsilon} \right) + (\varphi \|x_*\| + 2\|x_*\|) \frac{2}{2d/2} \|\eta\| \tag{2.66}
\]

while for \( z \in B(-\phi_d x_*, \varphi \|x_*\|) \),

\[
f_\eta(z) \geq \frac{\|x_*\|^2}{2d+1} \left( 1 + \phi_d^2 - 2\phi_d - 139d^4\sqrt{\varepsilon} - K_d \sqrt{\varepsilon} \right) - (\varphi \|x_*\| + 2\|x_*\|) \frac{2}{2d/2} \|\eta\|. \tag{2.67}
\]
Note that we require the lower bound in (2.67) to be larger than the upper bound in (2.66). Setting $\varphi = \varepsilon$ and using both $\varepsilon < \sqrt{\varepsilon}$ and $\|\eta\| \leq c_2 \frac{\|x\|}{2^{d/2}d^{4\delta}}$, we see that we require

$$\varepsilon \leq \left( \frac{2\phi_d(1 - \rho_d) - 8c_2/d^{4\delta}}{45d^3 + 2K_d + 139d^4} \right)^2.$$  

(2.68)

By Lemma 2.22, we have $1 - \rho_d \geq 1/(C(d + 2)^2)$ for some numerical constant $C$ and $\phi_d \geq \kappa$. Hence it suffices to have $\frac{c_2}{d^{4\delta}} \leq \frac{\kappa}{8C(d + 2)^2}$ and $\varphi = \varepsilon \leq r_1/d^{12}$ for some numerical constants $r_1$ and $c_2$.

□

**Lemma 2.22.** We have that $\rho_d$ satisfies $\min_{d \geq 2} \rho_d > 0$ and for some numerical constant $C$,

$$\frac{1}{C(d + 2)^2} \leq 1 - \rho_d \forall d \geq 2.$$

**Proof of Lemma 2.22.** Let

$$\Gamma_d := \sum_{i=0}^{d-1} \sin \frac{\delta_i}{\pi} \left( \prod_{j=i+1}^{d-1} \frac{\pi - \delta_j}{\pi} \right).$$

In Lemma A.4 of ref. [40], it has been established that $\Gamma_d \in [0,1]$ and $\min_{d \geq 2} \Gamma_d > 0$. By (2.31) and (2.32), we have that $\hat{\delta}_d \leq 3\pi/(d + 3)$ and $\hat{\delta}_d \geq \pi/(d + 1)$ for all $d \geq 2$. Since $\sin(2x) - 3x/4 \geq 0$ for all $x \in [0, \pi/3]$, observe that

$$\frac{2\sin \hat{\delta}_d}{\pi} \geq \frac{2\sin \left( \frac{\pi}{d+1} \right)}{\pi} \geq \frac{2}{\pi} \cdot \frac{3}{4} \left( \frac{\pi}{2(d+1)} \right) = \frac{3}{4(d+1)} \forall d \geq 2.$$

Thus for any $d \geq 2$,

$$\rho_d = \frac{2\sin \hat{\delta}_d}{\pi} + \left( \frac{\pi - 2\hat{\delta}_d}{\pi} \right) \Gamma_d \geq \frac{3}{4(d+1)} + \frac{\pi - \frac{6\pi}{d+3}}{\pi} \Gamma_d$$

$$\geq \frac{3}{4(d+1)} \Gamma_d + \frac{\pi - \frac{6\pi}{d+3}}{\pi} \Gamma_d$$

$$= \left( \frac{3}{4(d+1)} + \frac{d - 3}{d + 3} \right) \Gamma_d \geq \frac{1}{20} \Gamma_d$$

where the second inequality is due to $\Gamma_d \in [0,1]$. We conclude that $\min_{d \geq 2} \rho_d \geq 1/20 \min_{d \geq 2} \Gamma_d > 0$.

We now establish the lower bound on $1 - \rho_d$ for all $d \geq 2$. It was shown in Lemma A.4 of ref. [40] that $1 - \Gamma_d \geq \frac{1}{a_7(d+2)^2}$ for some numerical constant $a_7$. Observe that

$$\rho_d = \left( 1 - \frac{2\hat{\delta}_d}{\pi} \right) \Gamma_d + \frac{2}{\pi} \sin \hat{\delta}_d = \Gamma_d + \frac{2}{\pi} \left( \sin \hat{\delta}_d - \hat{\delta}_d \Gamma_d \right) \leq \Gamma_d + \frac{2}{\pi} \hat{\delta}_d (1 - \Gamma_d).$$
Furthermore, note that for all \( d \geq 2, \hat{\theta}_d \leq \hat{\theta}_2 = g(g(\pi)) = g(\pi/2) = \cos^{-1}(1/\pi) \). Hence for all \( d \geq 2, \)

\[
1 - \rho_d \geq 1 - \Gamma_d - \frac{2}{\pi} \hat{\theta}_d (1 - \Gamma_d) = (1 - \Gamma_d) \left( 1 - \frac{2}{\pi} \hat{\theta}_d \right) \geq \frac{1}{a_7 (d + 2)^2} \left( 1 - \frac{2}{\pi} \cos^{-1} \left( \frac{1}{\pi} \right) \right) \geq 0.2 \frac{1}{a_7 (d + 2)^2}.
\]

\[\square\]

2.5.4 | Proofs for Section 2.4.4

Here we prove the convexity-like property of \( f \) around the minimizer \( x_* \).

*Proof of Lemma 2.9.* Suppose our objective function \( f \) is differentiable at \( x \). Recall that the gradient of \( f \) is given by \( v_x = \bar{v}_x - q_x \) where \( \bar{v}_x = \Lambda_T^T (A_{x_d} \Lambda x - A_{x_*} \Lambda x_*), q_x = \Lambda_T^T A_{x_d} \eta \). We will first show that \( \bar{v}_x \) satisfies

\[
\left\| \bar{v}_x - \Lambda_T^T (\Lambda x x - \Lambda x_* x_*) \right\| \leq \frac{1}{16} \frac{1}{2^d} \| x - x_* \|.
\]

Note that by the triangle inequality, we have that

\[
\left\| \bar{v}_x - \Lambda_T^T (\Lambda x x - \Lambda x_* x_*) \right\| \leq \left\| \Lambda_T^T A_{x_d} (A_{x_d} \Lambda x - A_{x_*} \Lambda x_* x_*) - \Lambda_T^T (\Lambda x x - \Lambda x_* x_*) \right\|_{T_1} + \left\| \Lambda_T^T A_{x_d} (A_{x_d} - A_{x_*}) \Lambda x_* x_* \right\|_{T_2}.
\]

We will establish control of each of these terms separately.

*Controlling \( T_1 \):* Since \( f \) is differentiable at \( x \), note that by the local linearity of \( G \) we have that for sufficiently small \( z \in \mathbb{R}^k, G(x + z) - G(x) = \Lambda x z \). Hence for all \( z \), the RRCP implies that

\[
|\langle A_{x_d} \Lambda x z, A_{x_d} (\Lambda x x - \Lambda x_* x_*) \rangle - \langle A_{x_d} z, \Lambda x x - \Lambda x_* x_* \rangle| \leq L \varepsilon \| A_{x_d} \| \| \Lambda x x - \Lambda x_* x_* \| \| z \|.
\]

Since this holds for all \( z \), we have that

\[
\| \Lambda_T^T A_{x_d} (A_{x_d} \Lambda x x - A_{x_*} \Lambda x_* x_*) - \Lambda_T^T (\Lambda x x - \Lambda x_* x_*) \| \leq L \varepsilon \| A_{x_d} \| \| \Lambda x x - \Lambda x_* x_* \|.
\]

(2.69)
In addition, we have that by Lemma 2.11, if \( \varepsilon < 1/(200^4d^6) \) and \( x \in B(x_*, d \sqrt{\varepsilon} \|x_*\|) \) then

\[
\| \Lambda_xx - \Lambda_x x_* \| \leq \frac{1.2}{2^{d/2}} \| x - x_* \|. \tag{2.70}
\]

Combining (2.69), (2.70), and (2.14) in Lemma 2.15 we see that

\[
\| \Lambda_x^T A_{x_d}^T (A_{x_d}\Lambda_xx - A_{x_d}\Lambda_x x_*) - \Lambda_x^T (\Lambda_xx - \Lambda_x x_*) \| \leq \frac{1.2 \sqrt{\frac{13}{12} L \varepsilon}}{2^d} \| x - x_* \|. \tag{2.71}
\]

Thus choosing \( \varepsilon \) so that \( \varepsilon < 1/(32 \cdot 1.2 \sqrt{13/12L}) \) in (2.71) shows that

\[
T_1 = O_1 \left( \frac{1}{32} \right) \frac{1}{2^d} \| x - x_* \|. \tag{2.72}
\]

**Controlling \( T_2 \):** We will first show that for sufficiently small \( \varepsilon \),

\[
\| (A_{x_d} - A_{x_s,d}) \Lambda_{x_*} x_* \|^2 \leq \frac{1.44 (4L + \frac{48d}{\pi}) \sqrt{\varepsilon}}{2^d} \| x - x_* \|^2.
\]

Letting \( \{a_i\}_{i=1}^m \) denote the rows of \( A \), observe that we can write

\[
\| (A_{x_d} - A_{x_s,d}) \Lambda_{x_*} x_* \|^2 = \| (A_{x_d} - A_{x_s,d}) x_* \|^2 = \sum_{i=1}^m \left( \text{sgn}(\langle a_i, x_d \rangle) - \text{sgn}(\langle a_i, x_* \rangle) \right)^2 \langle a_i, x_* \rangle^2
\]

\[
\leq \sum_{i=1}^m \left( \text{sgn}(\langle a_i, x_d \rangle) - \text{sgn}(\langle a_i, x_* \rangle) \right)^2 \langle a_i, x_* \rangle^2
\]

\[
= \sum_{i=1}^m (1(\langle a_i, x_d \rangle \neq 0) + 1(\langle a_i, x_* \rangle \neq 0) - 2 \text{sgn}(\langle a_i, x_d \rangle \langle a_i, x_* \rangle) \langle a_i, x_* \rangle^2)
\]

\[
= \| A_{x_d} (x_d - x_* \rangle \|^2 + \| A_{x_s,d} (x_d - x_* \rangle \|^2
\]

\[
- 2 \langle x_d - x_* \rangle, A_{x_d}^T A_{x_s,d} (x_d - x_* \rangle \rangle.
\]

We first establish concentration of \( A_{x_d} (x_d - x_* \rangle \). Since \( A \) satisfies the RRCP with respect to \( G \), we have that

\[
| \langle (A_{x_d}^T A_{x_d} - I_{nd}) (x_d - x_* \rangle, x_d - x_* \rangle \rangle | \leq L \| x_d - x_* \|^2
\]

\[
| \langle (A_{x_d}^T A_{x_d} - I_{nd}) (x_d - x_* \rangle, x_d - x_* \rangle \rangle | \leq L \| x_d - x_* \|^2
\]
which ultimately gives
\[ \|A_{x_d}(x_d - x_{*,d})\|^2 \leq (1 + L\varepsilon)\|x_d - x_{*,d}\|^2. \]  
(2.73)

Likewise the same upper bound holds for \(A_{x_{*,d}}(x_d - x_{*,d})\):
\[ \|A_{x_{*,d}}(x_d - x_{*,d})\|^2 \leq (1 + L\varepsilon)\|x_d - x_{*,d}\|^2. \]  
(2.74)

We now aim to upper bound the inner product \(\langle x_d - x_{*,d}, A_{x_d}^T A_{x_{*,d}}(x_d - x_{*,d})\rangle\). We first note that since \(A\) satisfies the RRCP, we have
\[ \|\langle x_d - x_{*,d}, A_{x_d}^T A_{x_{*,d}}(x_d - x_{*,d})\rangle\| \leq L\varepsilon\|x_d - x_{*,d}\|^2. \]

Hence we have that
\[ \langle x_d - x_{*,d}, A_{x_d}^T A_{x_{*,d}}(x_d - x_{*,d})\rangle = \langle x_d - x_{*,d}, \Phi_{x_d,x_{*,d}}(x_d - x_{*,d})\rangle + O_1(L\varepsilon)\|x_d - x_{*,d}\|^2. \]  
(2.75)

But recall that \(x \in B(x_{*,d} \sqrt{\varepsilon}\|x_*\|)\) which implies \(|\theta_{0,x}d| \leq 2d \sqrt{\varepsilon}\). Since \(|\theta_{d,x}| \leq |\theta_{0,x}|\) we have \(|\theta_{d,x}| \leq 2d \sqrt{\varepsilon}\). Also Equation (2.47) gives
\[ |\theta_{d,x} - \bar{\theta}_{d,x}| \leq 4d \sqrt{\varepsilon}. \]

Hence we have that \(|\theta_{d,x}| \leq 6d \sqrt{\varepsilon}\). Thus \(\Phi_{x_d,x_{*,d}}\) is approximately an isometry since
\[ \|\Phi_{x_d,x_{*,d}} - I\| \leq \frac{2|\theta_{d,x}|}{\pi}\|I\| + \frac{2|\sin\theta_{d,x}|}{\pi}\|M_{x_d\leftrightarrow x_{*,d}}\| \leq \frac{24d \sqrt{\varepsilon}}{\pi}. \]

This implies that
\[ \langle x_d - x_{*,d}, \Phi_{x_d,x_{*,d}}(x_d - x_{*,d})\rangle = \|x_d - x_{*,d}\|^2 + O_1\left(\frac{24d \sqrt{\varepsilon}}{\pi}\right)\|x_d - x_{*,d}\|^2. \]  
(2.76)

Combining (2.75) and (2.76) we attain
\[ \langle x_d - x_{*,d}, A_{x_d}^T A_{x_{*,d}}(x_d - x_{*,d})\rangle = \|x_d - x_{*,d}\|^2 + O_1\left(\frac{24d \sqrt{\varepsilon}}{\pi} + L\varepsilon\right)\|x_d - x_{*,d}\|^2. \]

Note that this implies that
\[ -2\langle x_d - x_{*,d}, A_{x_d}^T A_{x_{*,d}}(x_d - x_{*,d})\rangle \leq \left(-2 + \frac{48d \sqrt{\varepsilon}}{\pi} + 2L\varepsilon\right)\|x_d - x_{*,d}\|^2. \]  
(2.77)
Returning to establishing concentration of \( (A_{xd} - A_{xs,d}) \Lambda_{xs,x_s} \), we can use (2.73), (2.74) and (2.77) to obtain

\[
\| (A_{xd} - A_{xs,d}) \Lambda_{xs,x_s} \|^2 \leq \| A_{xd} (x_d - x_{*,d}) \|^2 + \| A_{xs,d} (x_d - x_{*,d}) \|^2
- 2 (x_d - x_{*,d}, A^T_{xd} A_{xs,d} (x_d - x_{*,d}))
\leq \left( 2 + 2L \varepsilon - 2 + \frac{48d \sqrt{\varepsilon}}{\pi} + 2L \varepsilon \right) \| x_d - x_{*,d} \|^2
= \left( 4L \varepsilon + \frac{48d \sqrt{\varepsilon}}{\pi} \right) \| x_d - x_{*,d} \|^2.
\]

Using this inequality, Equation (2.70), and the fact that \( \varepsilon < \sqrt{\varepsilon} \), we attain

\[
\| (A_{xd} - A_{xs,d}) \Lambda_{xs,x_s} \|^2 \leq \left( 4L \varepsilon + \frac{48d \sqrt{\varepsilon}}{\pi} \right) \| x_d - x_{*,d} \|^2 \leq \frac{1.44(4L + \frac{48d}{\pi}) \sqrt{\varepsilon}}{2^d} \| x - x_* \|^2. \tag{2.78}
\]

Then by Equations (2.14) and (2.15) from Lemma 2.15, we have that

\[
\| A_{xd} \Lambda_x \| \leq \sqrt{1 + L \varepsilon} \| \Lambda_x \| \leq \sqrt{\frac{13}{12}} (1 + L \varepsilon) \frac{1}{2^{d/2}}. \tag{2.79}
\]

Combining (2.79) and (2.78) and choosing \( \varepsilon \) so that \( \sqrt{\frac{13}{12}} (1 + L \varepsilon) \leq 2 \), we attain

\[
\| A_{xd} \Lambda_x \| (A_{xd} - A_{xs,d}) \Lambda_{xs,x_s} \| \leq 2 \sqrt{1.44 \left( 4L + \frac{48d}{\pi} \right)} \sqrt{2^d} \| x - x_* \|.
\]

Thus if

\[
\varepsilon^{1/4} < \frac{1}{64 \sqrt{1.44 \left( 4L + \frac{48d}{\pi} \right)}}
\]

we attain

\[
\| A_{xd} \Lambda_x \| (A_{xd} - A_{xs,d}) \Lambda_{xs,x_s} \| \leq \frac{1}{32} 2^d \| x - x_* \|
\]

that is, \( T_2 \) satisfies

\[
T_2 = O_1 \left( \frac{1}{32} \right) 2^d \| x - x_* \|. \tag{2.80}
\]
Combining our results for $T_1$ and $T_2$ in Equations (2.72) and (2.80) we ultimately get

$$\left\| \frac{1}{2} (x - x_*) \right\| \leq \frac{1}{16} \frac{1}{2^d} \|x - x_*\|. \hspace{1cm} (2.81)$$

To finish establishing concentration of $\bar{u}_x$, we appeal to Lemma A.9 of ref. [40] which showed that if $\varepsilon < 1/(200^4 d^6)$ and $x \in B(x_*, d \sqrt{\varepsilon}\|x_*\|)$ then

$$\left\| \Lambda^T_x (\Lambda_x x - \Lambda_x x_*) - \frac{1}{2^d} (x - x_*) \right\| \leq \frac{1}{16} \frac{1}{2^d} \|x - x_*\|. \hspace{1cm} (2.82)$$

Thus by combining Equations (2.81) and (2.82), we finally attain

$$\left\| \frac{1}{2} (x - x_*) \right\| \leq \left\| \frac{1}{2} (x - x_*) \right\| + \left\| \Lambda^T_x (\Lambda_x x - \Lambda_x x_*) - \frac{1}{2^d} (x - x_*) \right\| \leq \frac{1}{8} \frac{1}{2^d} \|x - x_*\|$$

as desired. Including the bound on $\|q_x\|$ from (2.43), we achieve the final desired result:

$$\left\| \frac{1}{2} (x - x_*) \right\| \leq \left\| \frac{1}{2} (x - x_*) \right\| + \left\| q_x \right\| \leq \frac{1}{8} \frac{1}{2^d} \|x - x_*\| + \frac{2}{2^d/2} \|\eta\|. \hspace{1cm} \square$$



3 | GAUSSIAN MATRICES SATISFY THE RRCP

We set out to prove that Gaussian $A$ satisfies the RRCP with respect to $G$ with high probability. The particular result is stated as follows:

**Proposition 3.1** (RRCP). Fix $0 < \varepsilon < 1$. Let $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ have i.i.d. $\mathcal{N}(0, 1/n_i)$ entries for $i = 1, \ldots, d$. Let $A \in \mathbb{R}^{m \times nd}$ have i.i.d. $\mathcal{N}(0, 1/m)$ entries independent from $\{W_i\}$. Then if $m > \tilde{C}_\varepsilon dk \log(n_1 n_2 \ldots n_d)$, then with probability at least $1 - \tilde{\gamma} m^{4k} \exp(-\frac{\varepsilon}{2} m)$, we have that for all $x, z, x_1, x_2, x_3, x_4 \in \mathbb{R}^k$,

$$|\langle (A^T_{G(x)} A_{G(z)} - \Phi_{G(x), G(z)})(G(x_1) - G(x_2)), G(x_3) - G(x_4) \rangle| \leq L \varepsilon \|G(x_1) - G(x_2)\| \|G(x_3) - G(x_4)\|$$

Here $\tilde{\gamma}$ and $L$ are positive universal constants, $\tilde{C}_\varepsilon$ depends polynomially on $\varepsilon$, and $\tilde{C}_\varepsilon$ depends polynomially on $\varepsilon^{-1}$. 


We will prove Proposition 3.1 via the following steps:

1. We first establish that for any fixed non-zero \( y, w \in \mathbb{R}^n \), the inner product \( \langle A^T_y A_w u, v \rangle \) concentrates around its expectation \( \langle \Phi_{y,w} u, v \rangle \) for all \( u \) and \( v \) in a fixed \( k \)-dimensional subspace of \( \mathbb{R}^n \).
2. Then we show that this concentration holds uniformly for all \( y, w, u, v \) that live in the union of a finite number of \( k \)-dimensional subspaces of \( \mathbb{R}^n \).
3. To complete the proof, we apply the result from Step 2 for all \( y, w, u, v \) in the range of the generative model which precisely lives in the union of \( k \)-dimensional subspaces.

### 3.1 Concentration over a fixed subspace

We first show that the matrix \( A^T_y A_w \) concentrates around \( \Phi_{y,w} \) for any fixed \( y \neq w \) while acting on a fixed \( k \)-dimensional subspace \( T \) of \( \mathbb{R}^n \). We will refer to this result as the Restricted Concentration Property (RCP).

**Proposition 3.2** (Variant of Lemma 5.1 in ref. [10]; RCP). Fix \( 0 < \varepsilon < 1 \) and \( k < m \). Let \( A \in \mathbb{R}^{m \times n} \) have i.i.d. \( \mathcal{N}(0, 1/m) \) entries and fix \( y, w \in \mathbb{R}^n \setminus \{0\} \). Let \( T \subset \mathbb{R}^n \) be a \( k \)-dimensional subspace. Then if \( m \geq Ck \), we have that with probability exceeding \( 1 - 2 \exp(-c_1 m) \),

\[
|\langle A^T_y A_w u, u \rangle - \langle \Phi_{y,w} u, u \rangle| \leq \varepsilon \|u\|^2 \quad \forall \ u \in T
\]

and

\[
|\langle A^T_y A_w u, v \rangle - \langle \Phi_{y,w} u, v \rangle| \leq 3 \varepsilon \|u\|\|v\| \quad \forall \ u, v \in T.
\]

Furthermore, let \( U = \bigcup_{i=1}^M U_i \) and \( V = \bigcup_{j=1}^N V_j \) where \( U_i \) and \( V_j \) are subspaces of \( \mathbb{R}^n \) of dimension at most \( k \) for all \( i \in [M] \) and \( j \in [N] \). Then if \( m \geq 2Ck \)

\[
|\langle A^T_y A_w u, v \rangle - \langle \Phi_{y,w} u, v \rangle| \leq 3 \varepsilon \|u\|\|v\| \quad \forall \ u \in U, \ v \in V,
\]

with probability exceeding \( 1 - 2MN \exp(-c_1 m) \). Here \( c_1 \) depends polynomially on \( \varepsilon \) and \( C = \Omega(\varepsilon^{-1} \log \varepsilon^{-1}) \).

For the proof, we require the following large deviation inequality for subexponential random variables:

**Lemma 3.3** Corollary 5.17 in ref. [73]. Let \( Y_1, \ldots, Y_m \) be independent, centered, subexponential random variables. Let \( K = \max_{i \in [m]} \|Y_i\|_{\psi_1} \). Then for all \( \varepsilon > 0 \),

\[
P\left( \frac{1}{m} \left| \sum_{i=1}^m Y_i \right| \geq \varepsilon \right) \leq 2 \exp \left[ -c \min \left( \frac{\varepsilon^2}{K^2}, \frac{\varepsilon}{K} \right) m \right]
\]
where \( c > 0 \) is an absolute constant. Here \( \| \cdot \|_{\psi_1} \) is the subexponential norm: 
\[
\| X \|_{\psi_1} := \sup_{p \geq 1} p^{-1}(\mathbb{E} |X|^p)^{1/p}.
\]

We also require the following simple technical result.

**Proposition 3.4.** Fix \( y, w \in \mathbb{R}^n \setminus \{0\} \) and \( 0 < \varepsilon < 1 \). Let \( T \) be a subspace of \( \mathbb{R}^n \). If
\[
\left| \langle A_y^T A_w u, u \rangle - \langle \Phi_{y,w} u, u \rangle \right| \leq \varepsilon \| u \|^2 \quad \forall \ u \in T
\]
then
\[
\left| \langle A_y^T A_w u, v \rangle - \langle \Phi_{y,w} u, v \rangle \right| \leq 3 \varepsilon \| u \| \| v \| \quad \forall \ u, v \in T.
\]

With these two results, we are now equipped to prove Proposition 3.2.

**Proof of Proposition 3.2.** Without loss of generality, it suffices to show concentration over \( T \cap S^{n-1} \). For notational simplicity, set \( \Sigma_{y,w} := A_y^T A_w - \Phi_{y,w} \).

**Step 1: Approximation.** We first show that if concentration over an \( \varepsilon \)-net of \( T \cap S^{n-1} \) holds, then a continuity argument establishes concentration over all points in \( T \cap S^{n-1} \). Choose an \( \frac{\varepsilon}{14} \)-net \( Q_T \subset T \cap S^{n-1} \) such that \( |Q_T| \leq (42/\varepsilon)^k \) and for any \( u \in T \cap S^{n-1} \),
\[
\min_{q \in Q_T} \| u - q \| \leq \frac{\varepsilon}{14}.
\]

We will prove that
\[
|\langle \Sigma_{y,w} q, q \rangle| \leq \frac{\varepsilon}{8} \quad \forall \ q \in Q_T \implies |\langle \Sigma_{y,w} u, u \rangle| \leq \varepsilon \ \forall \ u \in T.
\]

Now, define
\[
\alpha^* := \inf \{ \alpha > 0 : |\langle \Sigma_{y,w} u, u \rangle| \leq \alpha \| u \|^2 \ \forall \ u \in T \}.
\]

We want to show that \( \alpha^* \leq \varepsilon \). Fix \( u \in T \cap S^{n-1} \). Then there exists a \( q \in Q_T \) such that \( \| u - q \| \leq \varepsilon/14 \). In addition, observe that \( u - q \in T \) since \( q \in Q_T \subset T \) so by (3.7),
\[
|\langle \Sigma_{y,w} (u - q), u - q \rangle| \leq \alpha^* \| u - q \|^2 \leq \alpha^* \frac{\varepsilon^2}{196}.
\]

Now, note that by the definition of \( \alpha^* \),
\[
|\langle \Sigma_{y,w} u, u \rangle| \leq \alpha^* \| u \|^2 \ \forall \ u \in T.
\]

Thus Proposition 3.4 gives
\[
|\langle \Sigma_{y,w} u, v \rangle| \leq 3 \alpha^* \| u \| \| v \| \ \forall \ u, v \in T.
\]
Applying this result to \( u - q \) and \( q \) gives
\[
|\langle \Sigma_{y,w}(u - q), q \rangle| \leq 3\alpha^* \|u - q\| \leq \alpha^* \frac{3\varepsilon}{14}. \tag{3.9}
\]

Let \( E \) be the event that \( |\langle \Sigma_{y,w}q, q \rangle| \leq \frac{\varepsilon}{8} \) for any \( q \in Q_T \). Using \( \langle \Sigma_{y,w}u, u \rangle = \langle \Sigma_{y,w}(u - q), u - q \rangle + 2\langle \Sigma_{y,w}u, q \rangle - \langle \Sigma_{y,w}q, q \rangle \) and \( \langle \Sigma_{y,w}u, q \rangle = \langle \Sigma_{y,w}(u - q), q \rangle + \langle \Sigma_{y,w}q, q \rangle \), we have that on \( E \),
\[
|\langle \Sigma_{y,w}u, u \rangle| \leq |\langle \Sigma_{y,w}(u - q), u - q \rangle| + 2|\langle \Sigma_{y,w}u, q \rangle| + |\langle \Sigma_{y,w}q, q \rangle|
\]
\[
\leq \alpha^* \frac{\varepsilon^2}{196} + \alpha^* \frac{3\varepsilon}{7} + \frac{3\varepsilon}{8}
\]
\[
= \alpha^* \left( \frac{\varepsilon^2}{196} + \frac{3\varepsilon}{7} \right) + \frac{3\varepsilon}{8}
\]
where we used (3.8), (3.9), and the event \( E \) in the third inequality. Thus
\[
|\langle \Sigma_{y,w}u, u \rangle| \leq \alpha^* \left( \frac{\varepsilon^2}{196} + \frac{3\varepsilon}{7} \right) + \frac{3\varepsilon}{8} \quad \forall u \in T \cap S^{n-1}. \tag{3.10}
\]

However, recall that \( \alpha^* \) was defined to be the smallest number such that
\[
|\langle \Sigma_{y,w}u, u \rangle| \leq \alpha^* \quad \forall u \in T \cap S^{n-1}.
\]
Hence \( \alpha^* \) must be smaller than the right hand side of (3.10), that is,
\[
\alpha^* \leq \alpha^* \left( \frac{\varepsilon^2}{196} + \frac{3\varepsilon}{7} \right) + \frac{3\varepsilon}{8} \implies \alpha^* \leq \frac{3\varepsilon}{8} \left( \frac{1}{1 - \frac{\varepsilon^2}{196} - \frac{3\varepsilon}{7}} \right) \leq \varepsilon
\]
since \( 0 < \varepsilon < 1 \). This establishes (3.6).

**Step 2: Concentration.** We now establish concentration for a fixed point \( u \in S^{n-1} \). Then observe that
\[
\left| \langle \Sigma_{y,w}u, u \rangle \right| = \frac{1}{m} \left| \sum_{i=1}^{m} Y_i \right|
\]
where \( Y_i = U_i - \mathbb{E}[U_i], U_i = \text{sgn}(\langle \tilde{a}_i, y \rangle \langle \tilde{a}_i, w \rangle) / \langle \tilde{a}_i, u \rangle^2 \), and each \( \tilde{a}_i \sim \mathcal{N}(0, I_n) \). Hence \( Y_i \) are independent, centered, subexponential random variables. We now estimate their subexponential norm prior to invoking Lemma 3.3.

By Remark 5.18 in ref. [73], the subexponential norm satisfies
\[
\|Y_i\|_{\psi_1} = \|U_i - \mathbb{E}[U_i]\|_{\psi_1} \leq 2\|U_i\|_{\psi_1}. \tag{3.11}
\]
Let \( Z_i := \langle \tilde{a}_i, u \rangle \sim \mathcal{N}(0, 1) \). Then \( \|Z_i\|_{\psi_2} \leq K_1 \) for some absolute constant \( K_1 \) where \( \| \cdot \|_{\psi_2} \) is the sub-gaussian norm. Observe that \( \mathbb{E} |U_i|^p \leq \mathbb{E} |Z_i^2|^p \) which implies \( U_i \|_{\psi_1} \leq \|Z_i^2\|_{\psi_1} \). Thus we
have
\[ \|Y_i\|_{\psi_1} \leq 2\|U_i\|_{\psi_1} \leq 2\|Z_i^2\|_{\psi_1} \leq 4\|Z_i\|_{\psi_2}^2 \leq 4K_1^2 \]
where we used Equation (3.11) in the first inequality and Lemma 5.14 in ref. [73] in the second to last inequality. Thus \( K = \max_{i \in [m]} \|Y_i\|_{\psi_1} \leq 4K_1^2 \) for an absolute constant \( K_1 \). Defining \( K_2 := 4K_1^2 \), Lemma 3.3 guarantees that for any fixed \( y, w, u \in \mathbb{R}^n \setminus \{0\} \) and \( \varepsilon > 0 \),
\[
\mathbb{P} \left( \left| \langle \Sigma_y, w, u \rangle \right| \geq \varepsilon \right) \leq 2 \exp(-c_0(\varepsilon)m) \tag{3.12}
\]
where \( c_0(\varepsilon) = c \min(\varepsilon^2/K_2^2, \varepsilon/K_2) \) and \( c > 0 \) is an absolute constant.

**Step 3: Union bound.** We now show concentration over \( Q_T \) holds. Recall that \( |Q_T| \leq (42/\varepsilon)^k \) so we can apply a union bound to (3.12) to attain
\[
\mathbb{P} \left( \left| \langle \Sigma_y, w, q \rangle \right| \geq \varepsilon \forall q \in Q_T \right) \leq 2 \left( \frac{42}{\varepsilon} \right)^k \exp \left( -c_0 \left( \frac{\varepsilon}{8} \right) m \right). \tag{3.13}
\]
By Equations (3.6) and (3.13), we conclude that
\[
\mathbb{P} \left( \left| \langle \Sigma_y, u, u \rangle \right| \geq \varepsilon \|u\|^2 \forall u \in T \right) \leq 2 \left( \frac{42}{\varepsilon} \right)^k \exp \left( -c_0 \left( \frac{\varepsilon}{8} \right) m \right).
\]
The probability bound in the proposition can be shown by noting that
\[
1 - 2(42/\varepsilon)^k \exp(-c_0(\varepsilon/8)m) = 1 - 2 \exp \left( -c_0(\varepsilon/8)m + k \log \left( \frac{42}{\varepsilon} \right) \right).
\]
Thus if
\[
\frac{2}{c_0(\varepsilon/8)} \log \left( \frac{42}{\varepsilon} \right) k \leq Ck \leq m
\]
where \( C = \Omega(\varepsilon^{-1} \log \varepsilon^{-1}) \), we have that the result holds with probability exceeding
\[
1 - 2 \exp \left( -c_0(\varepsilon/8)m + k \log \left( \frac{42}{\varepsilon} \right) \right) \geq 1 - 2 \exp(-c_1m)
\]
where \( c_1 = c_0(\varepsilon/8)/2 \). Applying Proposition 3.4 to our result gives (3.2) with the same probability. The extension to the union of subspaces follows by applying (3.2) to all subspaces of the form span(\( U_i, V_j \)) and using a union bound. Note that these subspaces have dimension at most \( 2k \), accounting for the extra factor of 2 in the bound on \( m \).

### 3.2 Uniform concentration over a union of subspaces

We will now set out to prove a stronger version of Proposition 3.2 that holds uniformly for all \( y \) and \( w \) in (possibly) different \( k \)-dimensional subspaces:
**Proposition 3.5** (Uniform RCP). Fix $0 < \varepsilon < 1$ and $k < m$. Let $A \in \mathbb{R}^{m \times n}$ have i.i.d. $\mathcal{N}(0, 1/m)$ entries. Let $T$, $W$, and $Y$ be fixed $k$-dimensional subspaces of $\mathbb{R}^n$. Then if $m \geq C \varepsilon k$, then with probability at least $1 - \tilde{C} \varepsilon^{4k} \exp(-\tilde{c} \varepsilon m)$, we have

$$\left| \langle A^T A_w, u \rangle - \langle \Phi_y, u \rangle \right| \leq L \varepsilon \|u\| \|y\| \quad \forall u, v \in T, w \in W, y \in Y \quad (3.14)$$

where $\gamma$ is a positive universal constant, $\tilde{C}$ depends on $\varepsilon$ and $C \varepsilon$ depends polynomially on $\varepsilon^{-1}$. Furthermore, let $U = \bigcup_{i=1}^{N_1} U_i$, $V = \bigcup_{j=1}^{N_2} V_j$, $W = \bigcup_{k=1}^{N_3} W_k$, and $Y = \bigcup_{\ell=1}^{N_4} Y_{\ell}$ be the union of at most $m$ $k$-dimensional subspaces of $\mathbb{R}^n$. Then if $m \geq 2 \tilde{C} k$,

$$\left| \langle A^T A_w, u \rangle - \langle \Phi_y, u \rangle \right| \leq L \varepsilon \|u\| \|v\| \quad \forall u \in U, v \in V, w \in W, y \in Y \quad (3.15)$$

with probability exceeding $1 - N_1 N_2 N_3 N_4 \tilde{C} \varepsilon^{4k} \exp(-\tilde{c} \varepsilon m)$. Here $L$ is a positive universal constant.

Note that Proposition 3.2 established concentration of $\langle A^T A_w, u \rangle$ around $\langle \Phi_y, u \rangle$ for $u$ and $y$ in a fixed $k$-dimensional subspace for fixed $y, w \in \mathbb{R}^n \setminus \{0\}$. We are interested in showing that this concentration holds uniformly for all $y$ and $w$ in the range of our generative model. The proof of this result uses an interesting fact from 1-bit compressed sensing which establishes that if a sufficient number of random hyperplanes cut the unit sphere, the diameter of each tessellation piece is small with high probability [60]. We state the theorem here for convenience:

**Theorem 3.6** (Theorem 2.1 in ref. [60]). Let $n, m, s > 0$ and set $\delta = C_1 \left( \frac{s}{m} \log(2n/s) \right)^{1/5}$. Let $a_i \in \mathbb{R}^n$ have i.i.d. $\mathcal{N}(0, 1)$ entries for $i \in [m]$. Then with probability at least $1 - C_2 \exp(-c \delta m)$, the following holds uniformly for all $u, \bar{u} \in \mathbb{R}^n$ that satisfy $\|u\|_2 = \|\bar{u}\|_2 = 1$, $\|u\|_1 \leq \sqrt{s}$, and $\|\bar{u}\|_1 \leq \sqrt{s}$ for $s \leq n$:

$$\langle a_i, \bar{u} \rangle \langle a_i, u \rangle \geq 0, \forall i \in [m] \implies \|\bar{u} - u\|_2 \leq \delta. \quad (3.16)$$

Here $C_1, C_2, c$ are positive universal constants.

We will use this result to prove the following: given a sufficient number of random hyperplanes and a $k$-dimensional subspace $Y$, there exists a finite set of points $Y_0 \subset Y$ that live in the interior of the tessellation pieces generated by the random hyperplanes such that any point in $Y$ can be closely approximated by a point in $Y_0$ with high probability.

**Lemma 3.7.** Fix $0 < \varepsilon < 1$. Let $A \in \mathbb{R}^{m \times n}$ have i.i.d. $\mathcal{N}(0, 1/m)$ entries with rows $\{a_\ell\}_{\ell=1}^m$. Let $Y$ be a $k$-dimensional subspace of $\mathbb{R}^n$. Define $E_{Y,A}$ to be the event that there exists a set $Y_0 \subset Y$ with the following properties:

- each $y_0 \in Y_0$ satisfies $\langle a_\ell, y_0 \rangle \neq 0$ for all $\ell \in [m]$,
- $|Y_0| \leq 10m^{2k}$, and
- for all $y \in Y$ such that $\|y\| = 1$, there exists a $y_0 \in Y_0$ such that $\|y - y_0\| \leq \varepsilon$.

If $m \geq \hat{C} k$, then $\mathbb{P}(E_{Y,A}) \geq 1 - C_2 \exp(-c \varepsilon m)$. Here $C_2$ and $c$ are positive absolute constants and $\hat{C}$ depends polynomially on $\varepsilon^{-1}$.
Proof of Lemma 3.7. By the rotational invariance of the Gaussian distribution, we may take \( Y \) to be in the span of the first \( k \) standard basis vectors. We may further without loss of generality assume \( A \in \mathbb{R}^{m \times k} \). We will invoke the following lemma which establishes that the unit sphere of \( Y \) is partitioned into at most \( 10m^{2k} \) regions by the rows \( \{a^T\}_{\ell=1}^{m} \) of \( A \) with probability 1:

**Lemma 3.8.** Let \( V \) be a subspace of \( \mathbb{R}^n \). Let \( A \in \mathbb{R}^{m \times n} \) have i.i.d. \( \mathcal{N}(0,1/m) \) entries. With probability 1,

\[
|\{\text{diag}(\text{sgn}(Av))A : v \in V\}| \leq 10m^{2\dim V}.
\]

In each tessellation piece defined by the rows of \( A \), choose a single point \( y_0 \) from \( Y \) with unit norm such that \( a^T_{\ell}y_0 \neq 0 \) for all \( \ell \in [m] \) (if such a point exists in the tessellation piece). Let \( Y_0 \) denote this collection of points and set \( I_{Y,A} := |Y_0| \). By Lemma 3.8 with \( V = Y \), the cardinality of \( Y_0 \) is bounded with probability 1: \( I_{Y,A} \leq 10m^{2k} \). Then observe that we can set the parameters \( n \) and \( s \) in Theorem 3.6 equal to \( k \) since \( A \in \mathbb{R}^{m \times k} \) and \( Y \) is in the span of the first \( k \) standard basis vectors. Then if \( m \geq (C_1^5 \log(2)/\varepsilon^5)k =: \hat{C}k \), we have that the quantity \( \delta \) in the theorem is bounded by \( \varepsilon \):

\[
\delta := C_1 \left( \frac{k}{m} \log(2) \right)^{1/5} \leq \varepsilon
\]

so \( \mathbb{P}(E_{Y,A}) \geq 1 - C_2 \exp(-c\varepsilon m) \) for some positive universal constants \( c, C_1, \) and \( C_2 \) and \( \hat{C} \) depends polynomially on \( \varepsilon^{-1} \).

We now proceed with the proof of the Uniform RCP.

**Proof of Proposition 3.5.** Let \( E_{Y,A} \) be the event defined in Lemma 3.7. By Lemma 3.7, we have that if \( m \geq \hat{C}k \), there exists an event \( E_{Y,A} \) with \( \mathbb{P}(E_{Y,A}) \geq 1 - C_2 \exp(-c\varepsilon m) \) on which there exists a finite subset \( Y_0 \) of \( Y \) with cardinality \( I_{Y,A} \leq 10m^{2k} \) such that for any \( y \in Y \) with \( \|y]\| = 1 \), there exists a \( y_0 \in Y_0 \) such that \( \|y - y_0\| \leq \varepsilon \). The analogous finite set \( W_0 \) with cardinality \( I_{W,A} \leq 10m^{2k} \) also exists on the event \( E_{W,A} \) with probability at least \( 1 - C_2 \exp(-c\varepsilon m) \). Thus if \( m \geq \hat{C}k \), the event \( E_{Y,W} := E_{Y,A} \cap E_{W,A} \) satisfies

\[
\mathbb{P}(E_{Y,W}) \geq 1 - 2C_2 \exp(-c\varepsilon m).
\]

We now establish concentration over \( Y_0 \) and \( W_0 \). Let \( E_0 \) be the event that

\[
\left| \langle A^T_{y_0} A_{w_0} u, v \rangle - \langle \Phi y_0, w_0 u, v \rangle \right| \leq 3\varepsilon \|u\| \|v\| \quad \forall \ u, v \in T, \ y_0 \in Y_0, \ w_0 \in W_0.
\]

By Proposition 3.2, if \( m \geq Ck \), we have that the following holds for fixed \( y_0 \in Y_0 \) and \( w_0 \in W_0 \) with probability exceeding \( 1 - 2 \exp(-c_1 m) \):

\[
\left| \langle A^T_{y_0} A_{w_0} u, v \rangle - \langle \Phi y_0, w_0 u, v \rangle \right| \leq 3\varepsilon \|u\| \|v\| \quad \forall \ u, v \in T.
\]
Furthermore, on $E_{Y,W}$, a union bound over all $y_0 \in Y_0$ and $w_0 \in W_0$ shows that

$$\mathbb{P}(E_0) \geq 1 - 2I_{Y,A}I_{W,A} \exp\left(-\frac{c_1m}{2}\right) \geq 1 - \gamma m^k \exp\left(-\frac{c_1m}{2}\right)$$

where $\gamma$ is a positive absolute constant and $c_1$ depends on $\epsilon$.

For the remainder of this proof, we work on the event $E_0 \cap E_{Y,W}$. Fix non-zero $y \in Y$ and $w \in W$. Define the following set:

$$\Omega_{y,w} := \{\ell \in [m] : \langle a_\ell, y \rangle = 0 \text{ or } \langle a_\ell, w \rangle = 0 \}.$$ 

Note that since $Y$ and $W$ are $k$-dimensional and any subset of $k$ rows of $A$ are linearly independent with probability 1, at most $k$ entries of $Ay$ are zero and similarly for $Aw$. Hence $|\Omega_{y,w}| \leq 2k$.

Furthermore, observe that

$$A_y^T A_w = \sum_{\ell=1}^{m} \operatorname{sgn}(\langle a_\ell, y \rangle \langle a_\ell, w \rangle) a_\ell a_\ell^T$$

by the definition of $\Omega_{y,w}$. However, on the event $E_{Y,W}$, there exists a $y_0 \in Y_0$ and $w_0 \in W_0$ such that for all $\ell \in \Omega_{y,w}$,

$$\operatorname{sgn}(\langle a_\ell, y \rangle) = \operatorname{sgn}(\langle a_\ell, y_0 \rangle) \text{ and } \operatorname{sgn}(\langle a_\ell, w \rangle) = \operatorname{sgn}(\langle a_\ell, w_0 \rangle)$$

that is, $y$ and $y_0$ (likewise $w$ and $w_0$) lie on the same side of each hyperplane defined by $\{a_\ell\}_{\ell=1}^m$.

Hence we have

$$A_y^T A_w = \sum_{\ell \in \Omega_{y,w}} \operatorname{sgn}(\langle a_\ell, y \rangle \langle a_\ell, w \rangle) a_\ell a_\ell^T = \sum_{\ell \in \Omega_{y,w}} \operatorname{sgn}(\langle a_\ell, y_0 \rangle \langle a_\ell, w_0 \rangle) a_\ell a_\ell^T$$

$$= A_{y_0}^T A_{w_0} - \sum_{\ell \in \Omega_{y,w}} \operatorname{sgn}(\langle a_\ell, y_0 \rangle \langle a_\ell, w_0 \rangle) a_\ell a_\ell^T$$

$$=: A_{y_0}^T A_{w_0} - \tilde{A}_{y_0}^T \tilde{A}_{w_0}. \quad (3.17)$$

We now use the following lemma which says that $\tilde{A}_{y_0}^T \tilde{A}_{w_0}$ is small when acting on $T$.

**Lemma 3.9.** Fix $0 < \epsilon < 1$ and $k < m$. Suppose that $A \in \mathbb{R}^{m \times n}$ has i.i.d. $\mathcal{N}(0, 1/m)$ entries. Let $T \subset \mathbb{R}^n$ be a $k$-dimensional subspace and $W_0$ and $Y_0$ be subsets of $\mathbb{R}^n$. Let $E$ be the event the following inequality holds for all $\Omega \subset [m]$ satisfying $|\Omega| \leq 2k$:

$$|\langle \tilde{A}_{y_0}^T \tilde{A}_{w_0} u, v \rangle| \leq \epsilon \|u\| \|v\| \quad \forall \ u, v \in T, \ w_0 \in W_0, \ y_0 \in Y_0 \quad (3.18)$$
where

\[ \tilde{A}_T^T \tilde{A}_{w_0} := \sum_{\ell \in \Omega} \text{sgn}(\langle a_\ell, y_0 \rangle \langle a_\ell, w_0 \rangle) a_\ell a_\ell^T. \]

Then there exists a \( \delta_\varepsilon > 0 \) such that if \( m \geq 9\varepsilon^{-1}k \) and \( 2k \leq \delta_\varepsilon m \), \( \mathbb{P}(E) \geq 1 - 2m \exp(-\varepsilon m/36) \).

Let \( E \) be the event defined in Lemma 3.9. On the event \( E \cap E_0 \cap E_{Y,W} \), we have that for all \( y \in Y \cap S^{n-1} \) and \( w \in W \cap S^{n-1} \), there exists a \( y_0 \in Y_0 \) and \( w_0 \in W_0 \) such that for any \( u, v \in T \),

\[
\left| \langle A_T^T A_w u, v \rangle - \langle \Phi_{y,w} u, v \rangle \right| = \left| \langle A_{y_0}^T A_{w_0} u, v \rangle - \langle \Phi_{y_0,w_0} u, v \rangle \right| \\
\leq \left| \langle A_{y_0}^T A_{w_0} u, v \rangle - \langle \Phi_{y,w} u, v \rangle \right| + \left| \langle \Phi_{y_0,w_0} u, v \rangle - \langle \Phi_{y,w} u, v \rangle \right| \\
\leq 3\varepsilon||u||||v|| + \frac{88}{\pi}\varepsilon||u||||v|| + \varepsilon||u||||v|| \\
=: L\varepsilon||u||||v||
\]

where we define \( L := 3 + \frac{88}{\pi} + 1 < 33 \). In the first equality, we used the event \( E_{Y,W} \) and (3.17). In the last inequality, we used the continuity of \( \Phi_{y,w} \) from Lemma 3.10 along with the event \( E_0 \cap E \).

Letting \( C_\varepsilon := 9\varepsilon^{-1}\hat{C} \) where \( \hat{C} \) is given by Lemma 3.7, we have that if \( m \geq C_\varepsilon k \), the event \( E \cap E_0 \cap E_{Y,W} \) holds with probability exceeding

\[ \mathbb{P}(E \cap E_0 \cap E_{Y,W}) \geq 1 - 2m \exp(-\varepsilon m/36) - \gamma m^{4k} \exp\left(-\frac{c_1 m}{2}\right) - 2C_2 \exp(-\varepsilon m) \]

\[ \geq 1 - \gamma m^{4k} \exp(-\hat{c}_\varepsilon m) \]

where \( \gamma \) is a positive absolute constant and \( \hat{c}_\varepsilon \) depends polynomially on \( \varepsilon \). The extension to the union of subspaces follows by applying (3.14) to all combinations of subspaces \( T_{ij} = \text{span}(U_i, V_j), W_k, \) and \( Y_\ell \) where each \( T_{ij} \) have dimension at most \( 2k \) and using a union bound. \( \square \)

### 3.3 Application to range of generative model

We now apply Proposition 3.5 to prove Proposition 3.1:

**Proof of Proposition 3.1.** For pedagogical purposes, we first establish the lemma in the \( d = 2 \) case. In order to apply Proposition 3.5, we will show that \( \{ G(x) - G(z) : x, z \in \mathbb{R}^k \} \) is a subset of the union of at most \( 10^6(n_1^2 n_2)^{2k} \) subspaces of dimensionality at most \( 2k \).

For fixed \( W_1, W_2 \), let \( A_{+,1} = \{ W_{1,+} : x \neq 0 \} \) and \( B_{+,2} = \{ W_{2,+} : x \neq 0 \} \). By Lemma 15 in ref. [38], there exists a probability one event, \( E \), over \( (W_1, W_2) \) on which \( |A_{+,1}| \leq 10n_1^k \) and \( |B_{+,2}| \leq ...
$10^2 n_1^k n_2^k$. On $E$,

$$|\{W_{2,+x} W_{1,+x} : x \neq 0\}| \leq 10^3 (n_1^2 n_2)^k.$$  

Note that $\dim \text{range}(W_{2,+x} W_{1,+x}) \leq k$ for all $x \neq 0$. Hence

$$\{G(x) : x \in \mathbb{R}^k\} \subset \{W_{2,+x} W_{1,+x} w : x, w \in S^{k-1}\} \subset V$$

where $V$ the union of at most $10^3 (n_1^2 n_2)^k$ subspaces of dimensionality at most $k$. This implies that

$$\{G(x) - G(z) : x, z \in \mathbb{R}^k\} \subset V'$$

where $V'$ is the union of at most $10^6 (n_1^2 n_2)^{2k}$ subspaces of dimensionality at most $2k$.

By applying the second half of Proposition 3.5 to the sets $V'$, $V'$, $V$, and $V$, we get that for fixed $W_1, W_2$,

$$|\langle (A^T G(x) A G(z) - \Phi G(x), G(z)) (G(x_1) - G(x_2)), G(x_3) - G(x_4) \rangle|$$

$$\leq L \varepsilon \|G(x_1) - G(x_2)\| \|G(x_3) - G(x_4)\|$$  

(3.19)

with probability at least

$$1 - 10^{3(2) + 6(2)} (n_1^2 n_2)^{2k+4k} \check{\gamma} m^{4k} e^{-\check{c}_{\varepsilon} m} \geq 1 - \check{\gamma} m^{4k} e^{-\check{c}_{\varepsilon} m/2},$$

provided $m \geq \hat{K} C_{\varepsilon}^d \bar{c}_{\varepsilon}^{-1} k \log(n_1 n_2) =: \hat{C}_{\varepsilon} k \log(n_1 n_2)$, where $\hat{\gamma}$ and $\hat{K}$ are positive universal constants, $\check{c}_{\varepsilon}$ depends polynomially on $\varepsilon$, and $C_{\varepsilon}$ depends polynomially on $\varepsilon^{-1}$.

Integrating over the probability space of $(W_1, W_2)$, independence of $A$ and $(W_1, W_2)$ implies that (3.19) holds for random $(W_1, W_2)$ with the same probability bound. Continuing from (3.19), we have

$$|\langle (A^T G(x) A G(z) - \Phi G(x), G(z)) (G(x_1) - G(x_2)), G(x_3) - G(x_4) \rangle|$$

$$\leq L \varepsilon \|G(x_1) - G(x_2)\| \|G(x_3) - G(x_4)\|$$

for all $x, z, x_1, x_2, x_3, x_4 \in \mathbb{R}^k$ with probability at least $1 - \check{\gamma} m^{4k} e^{-\check{c}_{\varepsilon} m/2}$ for some positive absolute constant $\check{\gamma}$ and $\check{c}_{\varepsilon}$ depends polynomially on $\varepsilon$.

The case for $d \geq 2$ follows similarly. We have

$$|\{I_{1=d} W_{i,+x} : x \neq 0\}| \leq 10^{(d^2)} (n_1^d n_2^{d-1} \cdots n_{d-1}^2 n_d)$$

on the probability 1 event. This implies that $G(x) : x \in \mathbb{R}^k \subset \{I_{1=d} W_{i,+x} w : x, w \in S^{k-1}\}$ is a subset of the union of at most $10^{(2d^2)} (n_1^d n_2^{d-1} \cdots n_{d-1}^2 n_d)^k$ subspaces of dimensionality at most $k$. Moreover, $G(x) - G(z) : x, z \in \mathbb{R}^k$ is a subset of the union of at most

$$10^{(2d^2)} (n_1^d n_2^{d-1} \cdots n_{d-1}^2 n_d)^{2k}$$
subspaces of dimensionality at most $2k$. Hence the analogous bound (3.19) holds for all $x, z, x_1, x_2, x_3, x_4 \in \mathbb{R}^k$ with probability at least

$$1 - 10^{(2d^2+4d^2)}(n_1^d n_2^{d-1} \cdots n_d^2)^{2k+4k} \gamma^4 m^{4k} e^{-\tilde{c}_1 m/2},$$

provided $m \geq \tilde{C}_\varepsilon dk \log(n_1 n_2 \cdots n_d)$, where $\gamma$ is a positive absolute constant, $\tilde{c}_1$ depends polynomially on $\varepsilon$, and $\tilde{C}_\varepsilon$ depends polynomially on $\varepsilon^{-1}$. □

3.4 RRCP supplementary results

Proof of Proposition 3.4. Fix $0 < \varepsilon < 1$. Suppose (3.4) holds and fix $u, v \in T$. Without loss of generality, assume $u$ and $v$ have unit norm. We will use the shorthand notation $\Phi = \Phi_{y,w}$. Since $T$ is a subspace, $u - v \in T$ so by (3.4),

$$\left| \langle A_{y}^T A_{w}(u - v), u - v \rangle - \langle \Phi(u - v), u - v \rangle \right| \leq \varepsilon \|u - v\|^2$$

or equivalently

$$\langle \Phi(u - v), u - v \rangle - \varepsilon \|u - v\|^2 \leq \langle A_{y}^T A_{w}(u - v), u - v \rangle \leq \langle \Phi(u - v), u - v \rangle + \varepsilon \|u - v\|^2. \quad (3.20)$$

Note that

$$\|u - v\|^2 = 2 - 2 \langle u, v \rangle,$$

$$\langle \Phi(u - v), u - v \rangle = \langle \Phi u, u \rangle + \langle \Phi v, v \rangle - 2 \langle \Phi u, v \rangle,$$

and

$$\langle A_{y}^T A_{w}(u - v), u - v \rangle = \langle A_{y}^T A_{w}u, u \rangle + \langle A_{y}^T A_{w}v, v \rangle - 2 \langle A_{y}^T A_{w}u, v \rangle$$

where we used the fact that $\Phi$ and $A_{y}^T A_{w}$ are symmetric. Rearranging (3.20) yields

$$2\left( \langle \Phi u, v \rangle - \langle A_{y}^T A_{w}u, v \rangle \right) \leq \left( \langle \Phi u, u \rangle - \langle A_{y}^T A_{w}u, u \rangle \right) + \left( \langle \Phi v, v \rangle - \langle A_{y}^T A_{w}v, v \rangle \right)$$

$$+ (2 - 2 \langle u, v \rangle)\varepsilon.$$

By assumption, the first two terms are bounded from above by $\varepsilon$. Thus

$$2\left( \langle \Phi u, v \rangle - \langle A_{y}^T A_{w}u, v \rangle \right) \leq 2\varepsilon + (2 - 2 \langle u, v \rangle)\varepsilon = 2(2 - \langle u, v \rangle)\varepsilon \leq 6\varepsilon$$

so

$$\langle \Phi u, v \rangle - \langle A_{y}^T A_{w}u, v \rangle \leq 3\varepsilon.$$

The lower bound is identical and establishes the desired result. □
Proof of Lemma 3.8. It suffices to prove the same upper bound for \( \|\text{sgn}(Av) : v \in V\| \). Let \( \ell = \text{dim } V \). By rotational invariance of Gaussians, we may take \( V = \text{span}(e_1, \ldots, e_\ell) \) without loss of generality. Without loss of generality, we may let \( A \) have dimensions \( m \times \ell \) and take \( V = \mathbb{R}^\ell \).

We will appeal to a classical result from sphere covering \([79]\). If \( m \) hyperplanes in \( \mathbb{R}^\ell \) contain the origin and are such that the normal vectors to any subset of \( \ell \) of those hyperplanes are independent, then the complement of the union of these hyperplanes is partitioned into at most

\[
2 \sum_{i=0}^{\ell-1} \binom{m-1}{i}
\]

disjoint regions. Each region uniquely corresponds to a constant value of \( \text{sgn}(Av) \) that has all non-zero entries. With probability 1, any subset of \( \ell \) rows of \( A \) are linearly independent, and thus,

\[
\|\text{sgn}(Av) : v \in \mathbb{R}^\ell, (Av)_i \neq 0 \forall i\| \leq 2 \sum_{i=0}^{\ell-1} \binom{m-1}{i} \leq 2\ell \left(\frac{em}{\ell}\right)^\ell \leq 10m^\ell
\]

where the first inequality uses the fact that \( \binom{m}{\ell} \leq \left(\frac{e}{m}\right)^m \) and the second inequality uses that \( 2\ell(e/\ell)^\ell \leq 10 \) for all \( \ell \geq 1 \).

For arbitrary \( v \), at most \( \ell \) entries of \( Av \) can be zero by linear independence of the rows of \( A \). At each \( v \), there exists a direction \( \delta \) such that \( (A(v + \delta v))_i \neq 0 \) for all \( i \) and for all \( \delta \) sufficiently small. Hence, \( \text{sgn}(Av) \) differs from one of \( \{\text{sgn}(Av) : v \in \mathbb{R}^\ell, (Av)_i \neq 0 \forall i\} \) by at most \( \ell \) entries. Thus,

\[
\|\{\text{sgn}(Av) : v \in \mathbb{R}^\ell\} \| \leq \left(\frac{m}{\ell}\right)\|\{\text{sgn}(Av) : v \in \mathbb{R}^\ell, (Av)_i \neq 0 \forall i\} \| \leq m^\ell 10m^\ell = 10m^{2\ell}.
\]

□

Proof of Lemma 3.9. For any \( \Omega \subset [m] \), let \( A_\Omega \) denote the submatrix of \( A \) with rows \( a^T_\ell \) where \( \ell \in \Omega \). We claim that it suffices to show

\[
\|A_\Omega u\| \leq \sqrt{\varepsilon}\|u\| \quad \forall u \in T \quad \forall \Omega \subset [m] \text{ satisfying } |\Omega| \leq 2k \leq \delta \varepsilon m. \tag{3.21}
\]

To see this, observe that for any \( w_0 \in W_0, y_0 \in Y_0 \), and \( u, v \in T \) and \( \Omega \subset [m] \), we have that

\[
|\langle \tilde{A}_{y_0} A_{w_0} u, v \rangle| = |\langle \text{diag}(\text{sgn}(A_{\Omega} y_0)) \odot \text{sgn}(A_{\Omega} w_0))A_{\Omega} u, A_{\Omega} v \rangle| \\
\leq \| \text{diag}(\text{sgn}(A_{\Omega} y_0)) \odot \text{sgn}(A_{\Omega} w_0)) \| \|A_{\Omega} u\| \|A_{\Omega} v\| \\
\leq \|A_{\Omega} u\| \|A_{\Omega} v\|
\]

where we used the Cauchy-Schwarz inequality in the first inequality. Hence establishing (3.21) will imply the desired conclusion.

By the rotational invariance of the Gaussian distribution, we may take \( T \) to be in the span of the first \( k \) standard basis vectors. We may further without loss of generality assume \( A \in \mathbb{R}^{m \times k} \) so it suffices to establish \( \|A_{\Omega}\| \leq \sqrt{\varepsilon} \). Fix \( \Omega \subset [m] \) satisfying \( |\Omega| \leq 2k \). By Corollary 5.35 in ref. [73],
we have that for any \( t \geq 0 \), it holds with probability \( 1 - 2 \exp(-t^2/2) \) that
\[
\sqrt{m}\|A_{\Omega}\| \leq \sqrt{\Omega} + \sqrt{k} + t.
\]

Taking \( t = \sqrt{\varepsilon m/3} \), we conclude that if \( |\Omega| \leq \varepsilon m/9 \) and \( m \geq 9k/\varepsilon \), then \( \|A_{\Omega}\| \leq \sqrt{\varepsilon} \) with probability \( 1 - 2\exp(-\varepsilon m/18) \).

We now establish that \( \|A_{\Omega}\| \leq \sqrt{\varepsilon} \) holds simultaneously over all subsets \( \Omega \subset [m] \) of a sufficiently small size with a union bound. Observe that since \( \lim_{\delta \to 0} (\varepsilon/\delta)^{\delta} = 1 \), there exists a \( \delta_\ast > 0 \) such that \( (\varepsilon/\delta_\ast)^{\delta_\ast} \leq \exp(\varepsilon/36) \). Put \( \delta \varepsilon := \min\{\varepsilon, \delta_\ast\} \). Let \( E \) be the event that \( \|A_{\Omega}\| \leq \sqrt{\varepsilon} \) for all subsets \( \Omega \subset [m] \) satisfying \( |\Omega| \leq 2k \leq \delta_\ast m \). If \( m \geq 9\varepsilon^{-1}k \), a union bound shows that this event holds with probability at least
\[
1 - 2 \sum_{\ell=1}^{\lfloor \delta_\ast m \rfloor} \left( \frac{m}{\ell} \right) \exp(-\varepsilon m/18) \geq 1 - 2 \delta_\ast m \left( \frac{\varepsilon m}{\delta_\ast m} \right)^{\delta_\ast m} \exp(-\varepsilon m/18) \geq 1 - 2 \delta_\ast m \left( \frac{\varepsilon}{\delta_\ast} \right)^{\delta_\ast m} \exp(-\varepsilon m/36) \geq 1 - 2m \exp(-\varepsilon m/36)
\]
where we used the fact that \( (\varepsilon/\delta_\ast)^{\delta_\ast} \leq \exp(\varepsilon/36) \) in the second to last inequality.

We now prove the continuity of \( \Phi_{y,w} \) for non-zero \( y, w \in \mathbb{R}^n \). Recall that
\[
\Phi_{y,w} := \frac{\pi - 2\Theta_{y,w}}{\pi} I_n + \frac{2\sin\Theta_{y,w}}{\pi} M_{y \leftrightarrow w}
\]
where \( \Theta_{y,w} := \angle(y, w) \) and \( M_{y \leftrightarrow w} \) is the matrix that sends \( \hat{y} \mapsto e_1 \), \( \hat{w} \mapsto \cos\Theta_{y,w}e_1 + \sin\Theta_{y,w}e_2 \), and \( h \mapsto 0 \) for all \( h \in \text{span}\{\{y, w\}^\perp\} \).

**Lemma 3.10** (Continuity of \( \Phi_{y,w} \)). Fix \( 0 < \varepsilon < 1 \) and \( y, w \in S^{n-1} \). If \( ||\hat{y} - y|| \leq \varepsilon \) and \( ||\hat{w} - w|| \leq \varepsilon \) for some \( \hat{y}, \hat{w} \in S^{n-1} \), then
\[
\|\Phi_{\hat{y},\hat{w}} - \Phi_{y,w}\| \leq \frac{88}{\pi} \varepsilon.
\]

**Proof of Lemma 3.10.** In this proof, we will utilize the following three inequalities:
\[
|\Theta_{y_1,\hat{y}} - \Theta_{y_2,\hat{y}}| \leq |\Theta_{y_1,y_2}|, \forall y_1, y_2, \hat{y} \in S^{n-1}
\]
\[
2\sin(\Theta_{y,w}/2) \leq ||y - w||, \forall y, w \in S^{n-1}
\]
\[ \frac{\theta}{4} \leq \sin(\frac{\theta}{2}), \forall \theta \in [0, \pi]. \] (3.24)

Observe that
\[ \|\Phi_{\tilde{y}, \tilde{w}} - \Phi_{y, w}\| \leq \frac{2|\theta_{\tilde{y}, \tilde{w}} - \theta_{y, w}|}{\pi} \|I_n\| + \left\| \frac{2\sin \frac{\theta_{\tilde{y}, \tilde{w}}}{\pi} M_{\tilde{y} \leftrightarrow \tilde{w}} - \frac{2\sin \frac{\theta_{y, w}}{\pi} M_{y \leftrightarrow w}}{\pi} \right\|. \]

First, observe that by (3.22), we have that
\[ |\theta_{\tilde{y}, \tilde{w}} - \theta_{y, w}| \leq |\theta_{\tilde{y}, \tilde{w}} - \theta_{\tilde{y}, \tilde{w}}| + |\theta_{y, \tilde{w}} - \theta_{y, w}| \leq |\theta_{\tilde{y}, y}| + |\theta_{\tilde{w}, w}|. \]

Then, by (3.23) and (3.24), we have that
\[ |\theta_{\tilde{y}, y}| \leq 4\sin(\frac{\theta_{\tilde{y}, y}}{2}) \leq 2\|\tilde{y} - y\| \leq 2\varepsilon. \]

The same upper bound holds for $|\theta_{\tilde{w}, w}|$. Thus we attain
\[ |\theta_{\tilde{y}, \tilde{w}} - \theta_{y, w}| \leq |\theta_{\tilde{y}, y}| + |\theta_{\tilde{w}, w}| \leq 4\varepsilon. \] (3.25)

Let $R$ be a rotation matrix that maps $y \mapsto e_1$ and $w \mapsto \cos \theta_{y,w} e_1 + \sin \theta_{y,w} e_2$ where $e_1$ and $e_2$ are the first and second standard basis vectors, respectively. Let $\tilde{R}$ denote the matrix that applies the same rotation to the system $\tilde{y}$ and $\tilde{w}$. Recall that $M_{\tilde{y} \leftrightarrow \tilde{w}} := \tilde{R}^T DR$ and $M_{y \leftrightarrow w} := R^T DR$ where

\[
D := \begin{bmatrix}
\cos \theta_{y,w} & \sin \theta_{y,w} & 0 \\
\sin \theta_{y,w} & -\cos \theta_{y,w} & 0 \\
0 & 0 & 0_{k-2}
\end{bmatrix}
\quad \text{and} \quad
\tilde{D} := \begin{bmatrix}
\cos \theta_{\tilde{y},\tilde{w}} & \sin \theta_{\tilde{y},\tilde{w}} & 0 \\
\sin \theta_{\tilde{y},\tilde{w}} & -\cos \theta_{\tilde{y},\tilde{w}} & 0 \\
0 & 0 & 0_{k-2}
\end{bmatrix}.
\]

An elementary calculation shows that $D$ has 2 pairs of non-zero eigenvalues and eigenvectors $(\lambda_1, d_1)$ and $(\lambda_2, d_2)$ where

\[ \lambda_1 = -1 \quad \text{and} \quad d_1 = (\cos \theta_{y,w} - 1)e_1 + \sin \theta_{y,w} e_2 \]

while

\[ \lambda_2 = 1 \quad \text{and} \quad d_2 = (\cos \theta_{y,w} + 1)e_1 + \sin \theta_{y,w} e_2. \]

Let $D = -d_1 d_1^T + d_2 d_2^T$ be the eigenvalue decomposition for $D$. Then by the definition of $M_{\tilde{y} \leftrightarrow \tilde{w}}$, $M_{\tilde{y} \leftrightarrow \tilde{w}} = \tilde{R}^T DR = -R^T d_1 d_1^T R + R^T d_2 d_2^T R =: -v_1 v_1^T + v_2 v_2^T$

so $v_1 := R^T d_1$ and $v_2 := R^T d_2$ are the eigenvectors of $M_{\tilde{y} \leftrightarrow \tilde{w}}$ with corresponding eigenvalues $-1$ and $1$, respectively. Then, recall that $R y = e_1$ while $R w = \cos \theta_{y,w} e_1 + \sin \theta_{y,w} e_2$. Thus the eigenvectors $d_1$ and $d_2$ can be written as

\[ d_1 = Rw - Ry \quad \text{and} \quad d_2 = Rw + Ry. \]
Thus the eigenvectors of $M_{y \leftrightarrow w}$ are precisely
\[ v_1 = w - y \text{ and } v_2 = w + y. \]

By the same argument, the eigenvectors of $\tilde{M}_{\tilde{y} \leftrightarrow \tilde{w}}$ are
\[ \tilde{v}_1 = \tilde{w} - \tilde{y} \text{ and } \tilde{v}_2 = \tilde{w} + \tilde{y} \]
with corresponding eigenvalues $-1$ and $1$, respectively. Hence, we have that
\[
\frac{2 \sin \theta_{y,w}}{\pi} M_{y \leftrightarrow w} = \frac{2 \sin \theta_{\tilde{y},\tilde{w}}}{\pi} \left(-v_1 v_1^T + v_2 v_2^T\right)
\]
\[= \frac{2 \sin \theta_{y,w}}{\pi} \left(-(w - y)(w - y)^T + (w + y)(w + y)^T\right)\]
and likewise
\[
\frac{2 \sin \theta_{\tilde{y},\tilde{w}}}{\pi} \tilde{M}_{\tilde{y} \leftrightarrow \tilde{w}} = \frac{2 \sin \theta_{\tilde{y},\tilde{w}}}{\pi} \left(-(\tilde{w} - \tilde{y})(\tilde{w} - \tilde{y})^T + (\tilde{w} + \tilde{y})(\tilde{w} + \tilde{y})^T\right).
\]

For simplicity of notation, let $h = w - y$, $\tilde{h} = \tilde{w} - \tilde{y}$, $g = w + y$, and $\tilde{g} = \tilde{w} + \tilde{y}$. Then
\[
\left\| \frac{2 \sin \theta_{y,w}}{\pi} M_{y \leftrightarrow w} - \frac{2 \sin \theta_{\tilde{y},\tilde{w}}}{\pi} \tilde{M}_{\tilde{y} \leftrightarrow \tilde{w}} \right\| = \frac{2}{\pi} \left| \sin \theta_{y,w} (hh^T + gg^T) + \sin \theta_{\tilde{y},\tilde{w}} (\tilde{h} \tilde{h}^T - \tilde{g} \tilde{g}^T) \right|
\]
\[\leq \frac{2}{\pi} \left( \left| \sin \theta_{y,w} hh^T - \sin \theta_{\tilde{y},\tilde{w}} \tilde{h} \tilde{h}^T \right| + \left| \sin \theta_{y,w} gg^T - \sin \theta_{\tilde{y},\tilde{w}} \tilde{g} \tilde{g}^T \right| \right).
\]

Note that since $y, w, \tilde{y}, \tilde{w} \in S^{n-1}$, $\|h\|, \|\tilde{h}\|, \|g\|, \|\tilde{g}\| \leq 2$. In addition,
\[\|h - \tilde{h}\| \leq \|y - \tilde{y}\| + \|w - \tilde{w}\| \leq 2\epsilon\]
and (3.25) implies
\[|\sin \theta_{y,w} - \sin \theta_{\tilde{y},\tilde{w}}| \leq |\theta_{y,w} - \theta_{\tilde{y},\tilde{w}}| \leq 4\epsilon.
\]

Hence
\[
\| \sin \theta_{y,w} hh^T - \sin \theta_{\tilde{y},\tilde{w}} \tilde{h} \tilde{h}^T \| \leq \| \sin \theta_{y,w} hh^T - \sin \theta_{y,w} hh^T \| + \| \sin \theta_{y,w} hh^T - \sin \theta_{y,w} \tilde{h} \tilde{h}^T \|
\]
\[+ \| \sin \theta_{\tilde{y},\tilde{w}} \tilde{h} \tilde{h}^T - \sin \theta_{\tilde{y},\tilde{w}} \tilde{h} \tilde{h}^T \|\]
\[\leq |\sin \theta_{y,w}| \|h\| \|h - \tilde{h}\| + |\sin \theta_{y,w}| \|\tilde{h}\| \|h - \tilde{h}\|
\]
\[+ \|\tilde{h} \tilde{h}^T\| |\sin \theta_{y,w} - \sin \theta_{\tilde{y},\tilde{w}}| \leq 20\epsilon.
\]
The same bound holds for $\| \sin \theta_{y,w} g g^T - \sin \theta_{\tilde{y},\tilde{w}} \tilde{g} \tilde{g}^T \|$.

Hence we attain

$$\left\| \frac{2 \sin \theta_{y,w}}{\pi} M_{y \leftrightarrow w} - \frac{2 \sin \theta_{\tilde{y},\tilde{w}}}{\pi} M_{\tilde{y} \leftrightarrow \tilde{w}} \right\| \leq \frac{80}{\pi} \varepsilon. \quad (3.26)$$

Combining (3.25) and (3.26), we see that

$$\| \Phi_{\tilde{y},\tilde{w}} - \Phi_{y,w} \| \leq \frac{2|\theta_{\tilde{y},\tilde{w}} - \theta_{y,w}|}{\pi} \| I_n \| + \left\| \frac{2 \sin \theta_{\tilde{y},\tilde{w}}}{\pi} M_{\tilde{y} \leftrightarrow \tilde{w}} - \frac{2 \sin \theta_{y,w}}{\pi} M_{y \leftrightarrow w} \right\| \leq \frac{88}{\pi} \varepsilon.$$

We now prove the inequalities used in the proof of Lemma 3.10.

**Proof of Equations (3.22), (3.23), and (3.24).** For (3.22), observe that we can write

$$y_1 = \cos \theta_{y_1,y} y + \sin \theta_{y_1,y} y_1^\perp$$

and

$$y_2 = \cos \theta_{y_2,y} y + \sin \theta_{y_2,y} y_2^\perp$$

where $y_1^\perp$ and $y_2^\perp$ are unit vectors that are orthogonal to $y$. Then observe that

$$\langle y_1, y_2 \rangle = \langle \cos \theta_{y_1,y} y + \sin \theta_{y_1,y} y_1^\perp, \cos \theta_{y_2,y} y + \sin \theta_{y_2,y} y_2^\perp \rangle$$

$$= \cos \theta_{y_1,y} \cos \theta_{y_2,y} + \sin \theta_{y_1,y} \sin \theta_{y_2,y} \langle y_1^\perp, y_2^\perp \rangle.$$

Since $\theta_{y_1,y}, \theta_{y_2,y} \in [0, \pi]$, we have that $\sin \theta_{y_1,y} \sin \theta_{y_2,y} \geq 0$. In addition, $\langle y_1^\perp, y_2^\perp \rangle \leq \| y_1^\perp \| \| y_2^\perp \| = 1$ so we attain

$$\langle y_1, y_2 \rangle \leq \cos \theta_{y_1,y} \cos \theta_{y_2,y} + \sin \theta_{y_1,y} \sin \theta_{y_2,y} = \cos (\theta_{y_1,y} - \theta_{y_2,y})$$

by the trigonometric identity $\cos (\alpha \pm \beta) = \cos \alpha \cos \beta \pm \sin \alpha \sin \beta$. Since the function $\cos^{-1}$ is decreasing on $[-1, 1]$, we see that

$$\theta_{y_1,y} - \theta_{y_2,y} \leq \cos^{-1}(\langle y_1, y_2 \rangle) = \theta_{y_1,y_2}.$$

Similarly, $\theta_{y_2,y} - \theta_{y_1,y} \leq \theta_{y_1,y_2}$ so we attain $|\theta_{y_1,y} - \theta_{y_2,y}| \leq |\theta_{y_1,y_2}|$.

For (3.23), observe that for $y, w \in S^{n-1}$,

$$\| y - w \|^2 = \| y \|^2 + \| w \|^2 - 2 \langle y, w \rangle$$

$$= \| y \|^2 + \| w \|^2 - 2 \| y \| \| w \| \cos \theta_{y,w}$$

$$= 2(1 - \cos \theta_{y,w}).$$
Thus, using the half angle formula
\[
\sin \frac{\vartheta}{2} = \text{sgn} \left( 2\pi - \vartheta + 4\pi \left\lfloor \frac{\vartheta}{4\pi} \right\rfloor \right) \sqrt{1 - \cos \vartheta}
\]
we see that
\[
\|y - w\| = \sqrt{2(1 - \cos \vartheta_{y,w})} = 2 \sqrt{1 - \cos \frac{\vartheta_{y,w}}{2}} \geq 2 \sin \frac{\vartheta_{y,w}}{2}.
\]

For (3.24), one can note that the function \(\psi(\vartheta) := 4 \sin \frac{\vartheta}{2} - \vartheta\) is positive for all \(\vartheta \in [0, \pi]\).

ACKNOWLEDGMENTS

Paul Hand acknowledges support by NSF awards DMS-1848087, DMS-2022205, and DMS-2053448. Oscar Leong acknowledges support by the NSF Graduate Research Fellowship under Grant No. DGE-1450681.

REFERENCES

1. A. Aberdam, D. Simon, and M. Elad, When and how can deep generative models be inverted? (2020). Preprint, ArXiv:2006.15555.
2. S. Arora, Y. Liang, and T. Ma, Why are deep nets reversible: a simple theory with implications for training, (2015). Preprint, CoRR abs/1511.05653.
3. M. Asim, F. Shamshad, and A. Ahmed, Blind image deconvolution using deep generative priors, IEEE Trans. Comput. Imaging 6 (2018), 1493–1506.
4. B. Aubin, B. Loureiro, A. Baker, F. Krzakala, and L. Zdeborová, Exact asymptotics for phase retrieval and compressed sensing with random generative priors, Proceedings of The First Mathematical and Scientific Machine Learning Conference, vol. 107, PMLR, Princeton, (2020, pp. 55–73.
5. B. Aubin, B. Loureiro, A. Maillard, F. Krzakala, and L. Zdeborová, The spiked matrix model with generative priors, IEEE Trans. Inform. Theory 67 (2021), no. 2, 1156–1181.
6. S. Bahmani and J. Romberg, Efficient compressive phase retrieval with constrained sensing vectors, Adv. Neural. Inf. Process. Syst. (NIPS) 28 (2015), 523–531.
7. S. Bahmani and J. Romberg, A flexible convex relaxation for phase retrieval, Electron. J. Stat. 11 (2017), 5254–5281.
8. A. S. Bandeira and D. G. Mixon, Near-optimal phase retrieval of sparse vectors, Wavelets and Sparsity XV, Proceedings of SPIE Optics+Photonics, (2013).
9. B. Barak, S. B. Hopkins, J. Kelner, P. K. Kothari, A. Moitra, and A. Potechin, A nearly tight sum-of-squares lower bound for the planted clique problem, SIAM J. Comput. 48 (2019), no. 2, 687–735.
10. R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, A simple proof of the restricted isometry property for random matrices, Constr. Approx. 28 (2008), no. 3, 253–263.
11. A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Sci. 2 (2009), no. 1, 183–202.
12. Q. Berthet and P. Rigollet, Complexity theoretic lower bounds for sparse PCA, J. Mach. Learn. Res. 30 (2013), 1046–1066.
13. A. Bora, A. G. Dimakis, A. Jalal, and E. Price, Compressed sensing using generative models, Proceedings of the 34th International Conference on Machine Learning, vol. 70, PMLR, Sydney, (2017, pp. 537–546.
14. T. Cai, X. Li, and Z. Ma, Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow, Ann. Statist. 44 (2016), no. 5, 2221–2251.
15. E. J. Candès and X. Li, Solving quadratic equations via phaselift when there are about as many equations as unknowns, Found. Comput. Math. 14 (2014), 1017–1026.
16. E. J. Candès, X. Li, and M. Soltanolkotabi, Phase retrieval via wirtinger flow: theory and applications, IEEE Trans. Inform. Theory 61 (2017), no. 4, 195–2007.
17. E. J. Candès, J. K. Romberg, and T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Comm. Pure Appl. Math. 59 (2006), no. 8, 1207–1223.
18. E. J. Candès, T. Strohmer, and V. Voroninski, Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming, Comm. Pure Appl. Math. 66 (2013), no. 8, 1241–1274.
19. C. Clason, Nonsmooth analysis and optimization, (2017). Preprint, ArXiv:1708.04180.
20. J. Cocola, P. Hand, and V. Voroninski, Nonasymptotic guarantees for low-rank matrix recovery with generative priors, Advances in Neural Information Processing Systems (NeurIPS), Vancouver, (2020).
21. C. Clason, Nonsmooth analysis and optimization, (2017). Preprint, ArXiv:1708.04180.
22. J. Cocola, P. Hand, and V. Voroninski, Nonasymptotic guarantees for low-rank matrix recovery with generative priors, Advances in Neural Information Processing Systems (NeurIPS), Vancouver, (2020).
23. Y. Deshpande and A. Montanari, Sparse PCA via covariance thresholding, Advances in Neural Information Processing Systems (NeurIPS), Vancouver, (2014). pp. 334–342.
24. D. Donoho, For most large underdetermined systems of linear equations the minimal L1-norm solution is also the sparsest solution, Comm. Pure Appl. Math. 59 (2006), no. 6, 797–829.
25. D. Donoho, M. Lustig, and J. M. Pauly, Sparse MRI: the application of compressed sensing for rapid mr imaging, Magn. Reson. Med. 58 (2007), no. 6, 1182–1195.
26. Y. C. Eldar, G. B. Giannakis, and G. Wang, Solving systems of random quadratic equations via truncated amplitude flow, IEEE Trans. Inform. Theory 23 (2017), no. 26, 773–794.
27. J. J. Engelsma, S. A. Grosz, and A. K. Jain, Printsgan: synthetic fingerprint generator, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2022), 1–14.
28. A. Fannjiang and T. Strohmer, The numerics of phase retrieval, Acta Numer. 29 (2020), 125–228.
29. J. Fienup, Phase retrieval algorithms: a comparison, Appl. Opt. 21 (1982), 2758–2768.
30. R. Gerchberg and W. Saxton, A practical algorithm for the determination of phase from image and diffraction plane pictures, Optik 35 (1972), 237–246.
31. T. Goldstein and C. Struder, Phasemax: convex phase retrieval via basis pursuit, IEEE Trans. Inform. Theory 64 (2018), no. 4, 2675–2689.
32. F. L. Gómez, A. Eftekhari, and V. Cevher, Fast and provable ADMM for learning with generative priors, Advances in Neural Information Processing Systems (NeurIPS), Vancouver, (2019).
33. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, Advances in Neural Information Processing Systems, Montreal (2014).
34. H. Günaydin, D. Tend, Y. Rivenson, A. Ozcan, and Y. Zhang, Phase recovery and holographic image reconstruction using deep learning in neural networks, Light Sci. Appl. 7 (2018), no. 2, 17141–17141.
35. P. Hand and B. Joshi, Global guarantees for blind demodulation with generative priors, Advances in Neural Information Processing Systems (NeurIPS), Vancouver, (2019).
36. P. Hand, O. Leong, and V. Voroninski, Phase retrieval under a generative prior, Advances in Neural Information Processing Systems (NeurIPS), Vancouver, (2018).
37. P. Hand and V. Voroninski, Compressed sensing from phaseless gaussian measurements via linear programming in the natural parameter space, (2016). Preprint, ArXiv:1611.05985.
38. P. Hand and V. Voroninski, Global guarantees for enforcing deep generative priors by empirical risk, IEEE Trans. Inform. Theory 66 (2019), no. 1, 401–418.
39. R. Heckel, W. Huang, P. Hand, and V. Voroninski, Rate-optimal denoising with deep neural networks, Inf. Inference 10 (2021), no. 4, 1251–1285.
40. W. Huang, P. Hand, R. Heckel, and V. Voroninski, A provably convergent scheme for compressive sensing under random generative priors, J. Fourier Anal. Appl. 27, (2021), 19.
41. R. Hyder, V. Shah, C. Hedge, and M. S. Asif, Alternating phase projected gradient descent with generative priors for solving compressive phase retrieval, Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP) (2019).
42. M. Iwen, A. Viswanathan, and Y. Wang, Robust sparse phase retrieval made easy, Appl. Comput. Harmon. Anal. 42 (2017), no. 1, 135–142.
43. K. Jaganathan, S. Oymak, and B. Hassibi, Sparse phase retrieval: convex algorithms and limitations, 2013 IEEE Int. Symp. Inf. Theory, Istanbul, Turkey, (2013), pp. 1022–1026.
44. G. Jagatap and C. Hegde, *Sample-efficient algorithms for recovering structured signals from magnitude-only measurements*, IEEE Trans. Inform. Theory 65 (2019), no. 7, 4434–4456.

45. P. Jain, P. Netrapalli, and S. Sanghavi, *Phase retrieval using alternating minimization*, IEEE Trans. Signal Process. 63 (2015), no. 18, 4814–4826.

46. A. Kappeler, S. Ghosh, J. Holloway, O. Cossairt, and A. Katsaggelos, *Ptychnet: CNN based fourier ptychography*, 2017 IEEE International Conference Image Processing (ICIP), Beijing, China, (2017), pp. 1712–1716.

47. T. Karras, S. Laine, and T. Aila, *A style-based generator architecture for generative adversarial networks*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2019.

48. D. Kingma and J. Ba, *Adam: a method for stochastic optimization*, International Conference on Learning Representations (ICLR), Banff, (2014).

49. D. P. Kingma and M. Welling, *Auo-encoding variational bayes*, International Conference on Learning Representations (ICLR), Banff, (2014).

50. R. Krauthgamer, B. Nadler, D. Vilenchik, and et al, *Do semidefinite relaxations solve sparse pca up to the information limit?*, Ann. Statist. 43 (2015), no. 3, 1300–1322.

51. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*, Proc. IEEE 86 (1998), no. 11, 2278–2324.

52. X. Li and V. Voroninski, *Sparse signal recovery from quadratic measurements via convex programming*, SIAM J. Math. Anal. 45 (2013), no. 5, 3019–3033.

53. Z. Liu, S. Gomes, A. Tiwari, and J. Scarlett, *Sample complexity bounds for 1-bit compressive sensing and binary stable embeddings with generative priors*, Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR, Vienna, (2020).

54. C. A. Metzler, A. Maleki, and R. G. Baraniuk, *BM3D-PRGAMP: compressive phase retrieval based on BM3D denoising*, IEEE International Conference on Image Processing (ICIP), Phoenix, (2016), pp. 2504–2508.

55. C. A. Metzler, P. Schniter, A. Veeraraghavan, and R. G. Baraniuk, *prDeep: robust phase retrieval with a flexible deep network*, Proceedings of the 35th International Conference on Machine Learning (ICML), PMLR, Sweden, (2018).

56. D. G. Mixon and S. Villar, *Sunlayer: stable denoising with generative networks*, (2018). Preprint, ArXiv:1803.09319.

57. H. Nyquist, *Certain topics in telegraph transmission theory*, Trans. AIEE. 47 (1928), no. 2, 617–644.

58. A. Osokin, A. Chessel, R. E. C. Salas, and F. Vaggi, *Gans for biological image synthesis*, IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017.

59. S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi, *Simultaneously structured models with application to sparse and low-rank matrices*, IEEE Trans. Inform. Theory 61 (2015), no. 5, 2886–2908.

60. Y. Plan and R. Vershynin, *One-bit compressed sensing by linear programming*, Comm. Pure Appl. Math. 66 (2013), no. 8, 1275–1297.

61. S. Qiu, X. Wei, and Z. Yang, *Robust one-bit recovery via ReLU generative networks: improved statistical rates and global landscape analysis*, Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR, Online, (2020).

62. M. Reiche and P. Jung, *Deepinit phase retrieval*, (2020). Preprint, ArXiv:2007.08214.

63. E. Richard and A. Montanari, *A statistical model for tensor PCA*, Adv. Neural Inf. Process. Syst. 2 (2014), 2897–2905.

64. V. Shah and C. Hedge, *Solving linear inverse problems using gan priors: An algorithm with provable guarantees*, Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP), Calgary, Alberta, (2018).

65. F. Shamshad and A. Ahmed, *Compressed sensing-based robust phase retrieval via deep generative priors*, IEEE Sens. J. 21 (2017), no. 2, 2286–2298.

66. H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. Andriole, and M. Michalski, Medical image synthesis for data augmentation and anonymization using generative adversarial networks, Gooya, A., Goksel, O., Oguz, I., Burgos, N. (eds.) Simulation and Synthesis in Medical Imaging. SASHIMI 2018. Lecture Notes in Computer Science (2018).

67. M. Soltanolkotabi, *Structured signal recovery from quadratic measurements: breaking sample complexity barriers via nonconvex optimization*, IEEE Trans. Inform. Theory 65 (2019), no. 4, 2374–2400.

68. G. Song, Z. Fan, and J. Lafferty, *Surfing: iterative optimization over incrementally trained deep networks*, Adv. Neural Inf. Process. Syst. 32 (2019), 15034–15043.
69. J. Sun, Q. Qu, and J. Wright, A geometric analysis of phase retrieval, Found. Comput. Math. 18 (2018), 1131–1198.
70. K. Tayal, C.-H. Lai, R. Manekar, Z. Zhuang, V. Kumar, and J. Sun, Deep learning initialized phase retrieval, NeurIPS 2020 Workshop on Deep Learning and Inverse Problems (2020).
71. A. N. Tikhonov, On stability of inverse problems, Dokl. Akad. Nauk SSSR 39 (1943), no. 5, 176–179.
72. T. Uelwer, A. Oberstraß, and S. Harmeling, Phase retrieval using conditional generative adversarial networks, 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 731–738.
73. R. Vershynin, Introduction to the non-asymptotic analysis of random matrices, chap. Compressed Sensing: Theory and Applications, Cambridge University Press, (2012).
74. V. Voroninski and Z. Xu, A strong restricted isometry property with an application to phaseless compressed sensing, Appl. Comput. Harmon. Anal. 40 (2016), no. 2, 386–395.
75. I. Waldspurger, Phase retrieval with random gaussian sensing vectors by alternating projections, IEEE Trans. Inform. Theory 64 (2018), no. 5, 3301–3312.
76. I. Waldspurger, A. d'Aspremont, and S. Mallat, Phase recovery, maxcut and complex semidefinite programming, Math. Program. 149 (2015), no. 1-2, 47–81.
77. G. Wang, L. Zhang, G. B. Giannakis, M. Akçakaya, and J. Chen, Sparse phase retrieval via truncated amplitude flow, IEEE Trans. Signal Process. 66 (2018), 479–491.
78. Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (2004), no. 4, 600–612.
79. J. G. Wendel, A problem in geometric probability, Math. Scand. 11 (1962), no. 1, 109–111.
80. G. Zhang, T. Guan, Z. Shen, X. Wang, T. Hu, D. Wang, Y. He, and N. Xie, Fast phase retrieval in off-axis digital holographic microscopy through deep learning, Opt. Express 26 (2018), no. 15, 19388–19405.
81. H. Zhang, Y. Liang, and Y. Chi, A nonconvex approach for phase retrieval: reshaped Wirtinger flow and incremental algorithms, J. Mach. Learn. Res. 18 (2017), no. 141, 1–35.