Robust One-Bit Recovery via ReLU Generative Networks: Near-Optimal Statistical Rate and Global Landscape Analysis

Shuang Qiu [*1] **Xiaohan Wei** [*2] **Zhuoran Yang** [3]

Abstract

We study the robust one-bit compressed sensing problem whose goal is to design an algorithm that faithfully recovers any sparse target vector $\theta_0 \in \mathbb{R}^d$ *uniformly* via m quantized noisy measurements. Specifically, we consider a new framework for this problem where the sparsity is implicitly enforced via mapping a low dimensional representation $x_0 \in \mathbb{R}^k$ through a known n-layer ReLU generative network $G : \mathbb{R}^k \to \mathbb{R}^d$ such that $\theta_0 = G(x_0)$. Such a framework poses low-dimensional priors on θ_0 without a known sparsity basis. We propose to recover the target $G(x_0)$ solving an unconstrained empirical risk minimization (ERM). Under a weak *sub-exponential measurement assumption*, we establish a joint statistical and computational analysis. In particular, we prove that the ERM estimator in this new framework achieves a statistical rate of $m = \widetilde{\mathcal{O}}(kn \log d/\varepsilon^2)$ recovering any $G(x_0)$ uniformly up to an error ε. When the network is shallow (i.e., n is small), we show this rate matches the information-theoretic lower bound up to logarithm factors on ε^{-1}. From the lens of computation, we prove that under proper conditions on the network weights, our proposed empirical risk, despite non-convexity, has no stationary point outside of small neighborhoods around the true representation x_0 and its negative multiple; furthermore, we show that the global minimizer of the empirical risk stays within the neighborhood around x_0 rather than its negative multiple under further assumptions on weights.

1. Introduction

Quantized compressed sensing investigates how to design the sensing procedure, quantizer, and reconstruction algorithm so as to recover a high dimensional vector from a limited number of quantized measurements. The problem of one-bit compressed sensing, which aims at recovering a target vector $\theta_0 \in \mathbb{R}^d$ from single-bit observations $y_i = \text{sign}(\langle a_i, \theta_0\rangle),\ i \in \{1, 2, \cdots, m\},\ m \ll d$ and random sensing vectors $a_i \in \mathbb{R}^d$, is particularly challenging. Previous theoretical successes on this problem (e.g. Jacques et al. (2013); Plan & Vershynin (2013); Zhu & Gu (2015)) mainly rely on two key assumptions: (1) The Gaussianity of the sensing vector a_i. (2) The sparsity of the vector θ_0 on a given basis. However, the practical significance of these assumptions are rather limited in the sense that it is difficult to generate Gaussian vectors and high dimensional targets in practice are often distributed near a low-dimensional manifold rather than sparse on some given basis. The goal of this work is to make steps towards addressing these limitations.

1.1. Sub-Gaussian One-Bit Compressed Sensing

As investigated in Ai et al. (2014), sub-Gaussian one-bit compressed sensing can easily fail regardless of the recovery algorithms. More specifically, consider two sparse vectors: $\theta_1 = [1,\ 0,\ 0,\ \cdots,\ 0]$, $\theta_2 = [1,\ -1/2,\ 0,\ \cdots,\ 0]$, and i.i.d. Bernoulli sensing vectors a_i, where each entry takes $+1$ and -1 with equal probabilities. Such sensing vectors are known to perform optimally in the ordinary linear compressed sensing scenario, but cannot distinguish between θ_1 and θ_2 in the current one-bit scenario regardless of algorithms. Moreover, Ai et al. (2014); Goldstein & Wei (2018) further propose *non-consistent* estimators whose discrepancies are measured in terms of certain distances between the Gaussian distribution and the distribution of the sensing vectors.

A major step towards consistent non-Gaussian one-bit compressed sensing is called *dithering*, which has been considered in several recent works (Xu & Jacques, 2018; Dirksen & Mendelson, 2018a). The key idea is that instead of $y_i = \text{sign}(\langle a_i, \theta_0\rangle)$, one considers a new procedure by adding artificial random noise τ_i before quantization: $y_i = \text{sign}(\langle a_i, \theta_0\rangle + \tau_i),\ i \in \{1, 2, \cdots, m\}$. In

[*] Equal contribution [1] University of Michigan, MI, USA. [2] University of Southern California, CA, USA. [3] Princeton University, NJ, USA. Correspondence to: Shuang Qiu <qiush@umich.edu>, Xiaohan Wei <xiaohanw@usc.edu>, Zhuoran Yang <zy6@princeton.edu>.

Proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

addition, Dirksen & Mendelson (2018a) proposes a new computationally-efficient convex recovery algorithm and shows that under the new quantization procedure and the sub-Gaussian assumption on a_i, one can achieve the best known statistical rate[1] $m = \widetilde{\mathcal{O}}(k \log d/\varepsilon^4)$ estimating *any k sparse $\theta_0 \in \mathbb{R}^d$ within radius R uniformly up to error ε with high probability.* Dirksen & Mendelson (2018b) further shows that the same algorithm can achieve the rate $m = \widetilde{\mathcal{O}}(k \log d/\varepsilon^2)$ for vectors a_i sampled from a specific circulant matrix. Without computation tractability, Jacques et al. (2013); Plan & Vershynin (2013); Dirksen & Mendelson (2018a) also show that one can achieve the near-optimal rate solving a non-convex constrained program with Gaussian and sub-Gaussian sensing vectors, respectively. It is not known though if the optimal rate is achievable via *computationally tractable algorithms*, not to mention more general measurements than Gaussian/sub-Gaussian vectors.

It is also worth emphasizing that the aforementioned works Plan & Vershynin (2013); Xu & Jacques (2018); Dirksen & Mendelson (2018a;b) obtain uniform recovery results which hold with high probability for all k sparse $\theta_0 \in \mathbb{R}^d$ within radius R. The ability of performing uniform recovery potentially allows θ_0 to be adversarially chosen with the knowledge of the algorithm. It is a characterization of "robustness" not inherited in the non-uniform recovery results (Plan & Vershynin, 2013; Zhang et al., 2014; Goldstein et al., 2018; Thrampoulidis & Rawat, 2018), which provide guarantees *recovering an arbitrary but fixed sparse vector* θ_0. However, with the better result comes the greater technical difficulty unique to one-bit compressed sensing known as the *random hyperplane tessellation problem*. Simply put, uniform recoverability is, in some sense, equivalent to the possibility of constructing a binary embedding of a sparse set into the Euclidean space via random hyperplanes. See Plan & Vershynin (2014); Dirksen & Mendelson (2018a) .

1.2. Generative Models and Compressed Sensing

Deep generative models have been applied to a variety of modern machine learning areas. In this work, we focus on using deep generative models to solve inverse problems, which has find extensive empirical successes in image reconstructions such as super-resolution (Sønderby et al., 2016; Ledig et al., 2017), image impainting (Yeh et al., 2017) and medical imaging (Hammernik et al., 2018; Yang et al., 2018). In particular, these generative model based methods have been shown to produce comparable results to the classical sparsity based methods with much fewer (sometimes 5-10x fewer) measurements, which will greatly benefit application areas such as magnetic resonance imaging (MRI) and computed tomography (CT), where the measurements are usually quite expensive to obtain. In contrast to widely recognized empirical results, theoretical understanding of generative models remains limited.

In a recent work, Bora et al. (2017) considers a linear model $\mathbf{y} = \mathbf{A}G(x_0) + \eta$, where $\mathbf{A}$ is a Gaussian measurement matrix, η is a bounded noise term and $G(\cdot)$ is an L-Lipschitz generative model. By showing that the Gaussian measurement matrix satisfies a restricted eigenvalue condition (REC) over the range of $G(\cdot)$, the authors prove the L_2 empirical risk minimizer

$$\widehat{x} \in \arg\min_{x \in \mathbb{R}^k} \|\mathbf{A}G(x) - \mathbf{y}\|_2^2 \tag{1}$$

satisfies an estimation error bound $\|\eta\|_2 + \varepsilon$ when the number of samples is of order $\mathcal{O}(k \log(L/\varepsilon)/\varepsilon^2)$. They further show that the $\log(1/\varepsilon)$ term in the error bound can be removed when $G(\cdot)$ is a multilayer ReLU network. In addition, Hand & Voroninski (2018); Huang et al. (2018) consider the same linear model with the aforementioned L_2 empirical risk minimizer and an n-layer ReLU network $G(\cdot)$. They show when the noise in the linear model is small enough, the measurement matrix satisfies range restricted concentration, which is stronger than REC, $m \geq \mathcal{O}(kn \log d \operatorname{poly}(\varepsilon^{-1}))$[2], and suitable conditions on the weights of the ReLU function hold, the L_2 empirical risk enjoys a favorable landscape. Specifically, there is no spurious local stationary point outside of small neighborhoods of radius $\mathcal{O}(\varepsilon^{1/4})$ around the true representation x_0 and its negative multiple, and with further assumptions, the point $\widehat{x}$ is guaranteed to be located around x_0 instead of its negative multiple. Moreover, Liu & Scarlett (2019); Kamath et al. (2019) study sample complexity lower bounds for the generative compressed sensing model as (1).

More recently, generative models have been applied to scenarios beyond linear models with theoretical guarantees. Wei et al. (2019) considers a non-linear recovery using a generative model, where the link function is assumed to be differentiable and the recovery guarantee is non-uniform. Hand & Joshi (2019) studies the landscape of L_2 empirical risk for blind demodulation problem with an n-layer ReLU generative prior. Using the same prior, Hand et al. (2018) analyzes the landscape of the amplitude flow risk objective for phase retrieval. Furthermore, Aubin et al. (2019) investigates the spiked matrix model using generative priors with linear activations. Besides these studies, there is another line of work investigating the problem of compressed sensing via generative models by the approximate message passing framework, e.g. Manoel et al. (2017); Pandit et al. (2020).

1.3. Summary of the Main Results

We introduce a new framework for robust *dithered* one-bit compressed sensing where the structure of target vector θ_0

[1] In this paper, we use $\widetilde{\mathcal{O}}(\cdot)$ to hide the logarithm factors.

[2] Here $\operatorname{poly}(\varepsilon^{-1})$ stands for polynomial dependency on ε^{-1}.

is represented via an n-layer ReLU network $G : \mathbb{R}^k \to \mathbb{R}^d$, i.e., $\theta_0 = G(x_0)$ for some $x_0 \in \mathbb{R}^k$ and $k \ll d$. Building upon this framework, we propose a new recovery algorithm by solving an unconstrained ERM. We show this algorithm enjoys the following favorable properties:

- Statistically, when taking measurements a_i to be *sub-exponential* random vectors, with high probability and uniformly for any $G(x_0) \in G(\mathbb{R}^k) \cap \mathbb{B}_2^d(R)$, where $\mathbb{B}_2^d(R)$ is the ball of radius $R > 0$ centered at the origin, the solution $G(\widehat{x}_m)$ to the ERM recovers the true vector $G(x_0)$ up to error ε when the number of samples $m \geq \mathcal{O}(kn \log^4(\varepsilon^{-1})(\log d + \log(\varepsilon^{-1}))/\varepsilon^2)$. In particular, our result does not require REC type assumptions adopted in previous analysis of generative signal recovery works and at the same time weakens the known sub-Gaussian assumption adopted in previous sparse one-bit compressed sensing works. Moreover, we further establish an information-theoretic lower bound for the sample complexity. When the number of layers n is small, we show that the proved statistical rate matches the information-theoretic lower bound up to logarithm factors on ε^{-1}.

- Computationally, building upon the previous methods guaranteeing uniform recovery, we show that solving the ERM and approximate the true representation $x_0 \in \mathbb{R}^k$ can be tractable under further assumptions on ReLU networks. More specifically, we prove with high probability, there always exists a descent direction outside of two small neighborhoods around x_0 and $-\rho_n x_0$ with radius $\mathcal{O}(\varepsilon^{1/4})$ respectively, where $\rho_n > 0$ is a factor depending on n. This holds uniformly for any $x_0 \in \mathbb{B}_2^k(R')$ with $R' = (0.5 + \varepsilon)^{-n/2} R$, when the ReLU network satisfies a weight distribution condition with parameter $\varepsilon > 0$ and $m \geq \mathcal{O}(kn \log^4(\varepsilon^{-1})(\log d + \log(\varepsilon^{-1}))/\varepsilon^2)$. Furthermore, when ε is small enough, one guarantees that the solution $\widehat{x}_m$ stays within the neighborhood around x_0 rather than $-\rho_n x_0$. Our result is achieved under quantization errors and without assuming the REC type conditions, thereby improving upon previously known computational guarantees for ReLU generative signal recovery in linear models with small noise.

From a technical perspective, our proof makes use of the special piecewise linearity property of ReLU network. The merits of such a property in the current scenario are two folds: (1) It allows us to replace the generic chaining type bounds commonly adopted in previous works, e.g. Dirksen & Mendelson (2018a), by novel arguments that are "sub-Gaussian free". (2) From a hyperplane tessellation point of view, we show that for a given accuracy level, a binary embedding of $G(\mathbb{R}^k) \cap \mathbb{B}_2^d(R)$ into Euclidean space is "easier" in that it requires less random hyperplanes than that of a bounded k sparse set, e.g. Plan & Vershynin (2014); Dirksen & Mendelson (2018a).

Notation. Throughout this paper, let $\mathcal{S}^{d-1}$ and $\mathbb{B}_2^d(r)$ be the unit Euclidean sphere and the Euclidean ball of radius r centered at the origin in $\mathbb{R}^d$, respectively. We also use $\mathcal{B}(x, r)$ to denote the Euclidean ball of radius r centered at $x \in \mathbb{R}^k$. For a random variable $X \in \mathbb{R}$, the L_p-norm ($p \geq 1$) is denoted as $\|X\|_{L_p} = \mathbb{E}[|X|^p]^{1/p}$. The Olicz ψ_1-norm is denoted $\|X\|_{\psi_1} := \sup_{p\geq 1} p^{-1}\|X\|_{L_p}$. We say a random variable is sub-exponential if its ψ_1-norm is bounded. A random vector $x \in \mathbb{R}^d$ is sub-exponential if there exists a a constant $C > 0$ such that $\sup_{t\in\mathcal{S}^{d-1}} \|\langle x, t\rangle\|_{\psi_1} \leq C$. We use $\|x\|_{\psi_1}$ to denote the minimal C such that this bound holds. Furthermore, $C, C', c, c_1, c_2, c_3, c_4, c_5$ denote absolute constants, and their actual values can be different per appearance. We let $[n]$ denote the set $\{1, 2, \ldots, n\}$. We denote $\mathbf{I}_p$ and $\mathbf{0}_{p\times d}$ as a $p \times p$ identity matrix and a $p \times d$ all-zero matrix respectively.

2. Model

In this paper, we focus on one-bit recovery model in which one observes quantized measurements of the following form

$$y = \operatorname{sign}(\langle a, G(x_0)\rangle + \xi + \tau), \tag{2}$$

where $a \in \mathbb{R}^d$ is a random measurement vector, $\xi \in \mathbb{R}$ is a random pre-quantization noise with an unknown distribution, τ is a random quantization threshold (i.e., *dithering noise*) which one can choose, and $x_0 \in \mathbb{R}^k$ is the unknown representation to be recovered. We are interested the high-dimensional scenario where the dimension of the representation space k is potentially much less than the ambient dimension d. The function $G : \mathbb{R}^k \to \mathbb{R}^d$ is a fixed ReLU neural network of the form:

$$G(x) = \sigma \circ (W_n \sigma \circ (W_{n-1} \cdots \sigma \circ (W_1 x))), \tag{3}$$

where $\sigma \circ (\cdot)$ denotes the entry-wise application of the ReLU activation function $\sigma(\cdot) = \max\{\cdot, 0\}$ on a vector. We consider a scenario where the number of layers n is smaller than d and the weight matrix of the i-th layer is $W_i \in \mathbb{R}^{d_i \times d_{i-1}}$ with $d_n = d$ and $d_i \leq d, \forall i \in [n]$. Throughout the paper, we assume that $G(x_0)$ is bounded, i.e. there exists an $R \geq 1$ such that $\|G(x_0)\|_2 \leq R$, and we take $\tau \sim \text{Unif}[-\lambda, +\lambda]$, i.e. a uniform distribution bounded by a chosen parameter $\lambda > 0$. Let $\{(a_i, y_i)\}_{i=1}^m$ be i.i.d. copies of (a, y). Our goal is to compute an estimator $G(\widehat{x}_m)$ of $G(x_0)$ such that $\|G(\widehat{x}_m) - G(x_0)\|_2$ is small.

We propose to solve the following ERM for estimator $\widehat{x}_m$:

$$\arg\min_{x\in\mathbb{R}^k} \left\{ L(x) := \|G(x)\|_2^2 - \frac{2\lambda}{m}\sum_{i=1}^m y_i \langle a_i, G(x)\rangle \right\}, \tag{4}$$

where $y_i = \mathrm{sign}(\langle a_i, G(x_0)\rangle + \xi_i + \tau_i)$. It is worth mentioning that, in general, there is no guarantee that the minimizer of $L(x)$ is unique. Nevertheless, in Sections §3.1 and §3.3, we will show that any solution $\widehat{x}_m$ to this problem must satisfy the desired statistical guarantee and stay inside small neighborhoods around the true signal x_0 and its negative multiple with high probability.

3. Main Results

In this section, we establish our main theorems regarding statistical recovery guarantee of $G(x_0)$ and the associated information-theoretic lower bound in Sections §3.1 and §3.2. The global landscape analysis of the empirical risk $L(x)$ is presented in Section §3.3.

3.1. Statistical Guarantee

We start by presenting the statistical guarantee of using ReLU network for one-bit compressed sensing. Our statistical guarantee relies on the following assumption on the measurement vector and noise.

Assumption 3.1. *The measurement vector $a \in \mathbb{R}^d$ is mean 0, isotropic and sub-exponential. The noise ξ is also a sub-exponential random variable.*

Under this assumption, we have the following main statistical performance theorem.

Theorem 3.2. *Suppose Assumption 3.1 holds and consider any $\varepsilon \in (0,1)$. Set $C_{a,\xi,R} = \max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\}$, $\lambda \geq 4C_{a,\xi,R} \cdot \log(64C_{a,\xi,R} \cdot \varepsilon^{-1})$, and*

$$m \geq c_2\|a\|_{\psi_1}^2 \lambda^2 \log^2(\lambda m)\big[kn\log(ed) + k\log(2R) + k\log m + u\big]/\varepsilon^2. \quad (5)$$

Then, with probability at least $1 - c_3\exp(-u)$, $\forall u \geq 0$, any solution $\widehat{x}_m$ to (4) satisfies

$$\|G(\widehat{x}_m) - G(x_0)\|_2 \leq \varepsilon$$

for all x_0 such that $\|G(x_0)\|_2 \leq R$, where $c_1,\ c_2,\ c_3 \geq 1$ are absolute constants.

Remark 3.3 (Sample Complexity)**.** *One can verify that the sample complexity enforced by (5) holds when $m \geq C\log^4(C_{a,\xi,R} \cdot \varepsilon^{-1})(kn\log(ed) + k\log(2R) + k\log(\varepsilon^{-1}) + u)/\varepsilon^2$, where C is a large enough absolute constant. This gives the $\mathcal{O}(kn\log^4(\varepsilon^{-1})(\log d + \log(\varepsilon^{-1}))/\varepsilon^2)$ sample complexity. In particular, when the number of layers n is small, our result meets the optimal rate of sparse recovery (up to a logarithm factor) and demonstrate the effectiveness of recovery via generative models theoretically. The dependence on the number of layers n results from the fact that our bound counts the number of linear pieces split by the ReLU generative network (see Lemma B.2 for details). Measuring certain complexities of a fixed neural network via counting linear pieces arises in several recent works (e.g. Lei et al. (2018)), and the question whether or not the linear dependence on n is tight warrants further studies.*

Note that our result is a *uniform recovery result* in the sense that the bound $\|G(\widehat{x}_m) - G(x_0)\|_2 \leq \varepsilon$ holds with high probability uniformly for any target $x_0 \in \mathbb{R}^k$ such that $\|G(x_0)\|_2 \leq R$. This should be distinguished from known bounds (Plan & Vershynin, 2013; Zhang et al., 2014; Goldstein et al., 2018; Thrampoulidis & Rawat, 2018) on sparse one-bit sensing which hold only for a *fixed* sparse vector. The boundedness of $G(x_0)$ is only assumed for theoretical purpose, which could be removed for practice.

Moreover, apart from the ReLU network, the proof of this theorem can be extended to other networks possessing the piecewise linearity property. Whether the analysis can be applied to networks with a wider class of nonlinear activation functions remains to be further studied. The proof sketch is presented in Section §4.2 with more proof details in Supplement §C.

3.2. Information-Theoretic Lower Bound

In this section, we show that when the network is shallow, i.e., n is small, for any k and d, there exists a ReLU network of the form (3) such that the above rate in Theorem 3.2 is optimal up to some logarithm factors. More specifically, we have the following theorem.

Theorem 3.4. *For any positive k and d large enough such that $k \ll d$ with $k \leq d/4$, there exists a generative network G of the form (3) with a $k+1$ dimensional input, depth $n = 3$ such that for the linear model before quantization: $\check{y} = \langle a, \theta_0\rangle + \xi$, where $\theta_0 \in G(\mathbb{R}^{k+1}) \cap \mathbb{B}_2^d(1)$, $\xi \sim \mathcal{N}(0,1)$, $a \sim \mathcal{N}(0, \mathbf{I}_d)$ and $m \geq c_1 k\log(d/k)$, we have*

$$\inf_{\widehat{\theta}} \sup_{\theta_0 \in G(\mathbb{R}^{k+1}) \cap \mathbb{B}_2^d(1)} \mathbb{E}\|\widehat{\theta} - \theta_0\|_2 \geq c_2\sqrt{\frac{k\log(d/k)}{m}}, \quad (6)$$

where $c_1, c_2 > 0$ are absolute constants and the infimum is taken over all estimators $\widehat{\theta}$ generated by all possible algorithms depending only on m i.i.d. copies of $(a, \check{y})$.

This theorem gives a lower bound of sample complexity over the set of all algorithms $\widetilde{\mathcal{A}}$ recovering θ_0 from the noisy linear model $\check{y} = \langle a, \theta_0\rangle + \xi$ by observing $(a_i, \check{y}_i)_{i=1}^m$. It gets connected to the one-bit dithered observations as follows: We consider a subset $\mathcal{A} \subseteq \widetilde{\mathcal{A}}$ of algorithms, which adds dithering noise τ_i and then uses quantized observations $(a_i, y_i)_{i=1}^m$ to recover θ_0, where $y_i = \mathrm{sign}(\check{y}_i + \tau_i)$. The corresponding estimators generated by any algorithm in $\mathcal{A}$ will also satisfy (6). Thefore, we have the following corollary of Theorem 3.4, which gives the lower bound of sample complexity for one-bit recovery via ReLU network.

Corollary 3.5. *For any positive k and d large enough such that $k \ll d$ with $k \leq d/4$, there exists a generative network G of the form (3) with a $k+1$ dimensional input, depth $n = 3$ such that for the quantizd linear model: $y = \operatorname{sign}(\langle a, \theta_0 \rangle + \xi + \tau)$, where $\theta_0 \in G(\mathbb{R}^{k+1}) \cap \mathbb{B}_2^d(1)$, $\xi \sim \mathcal{N}(0,1)$, $a \sim \mathcal{N}(0, \mathbf{I}_d)$ and $m \geq c_1 k \log(d/k)$, we have*

$$\inf_{\widehat{\theta}} \sup_{\theta_0 \in G(\mathbb{R}^{k+1}) \cap \mathbb{B}_2^d(1)} \mathbb{E}\|\widehat{\theta} - \theta_0\|_2 \geq c_2 \sqrt{\frac{k \log(d/k)}{m}},$$

where $c_1, c_2 > 0$ are absolute constants and the infimum is taken over all estimators $\widehat{\theta}$ generated by all possible algorithms depending on m i.i.d. copies $(a_i, y_i)_{i=1}^m$ of (a, y).

Remark 3.6 (Lower Bound)**.** *This corollary indicates that the sample complexity recovering θ_0 within error ε is at least $\Omega\big(k \log(d/k)/\varepsilon^2\big)$. Thus, when the ReLU network is shallow (the depth n is small) and $k \ll d$, the sample complexity we have obtained in Theorem 3.2 and Remark 3.3 is near-optimal up to logarithm factors of ε^{-1} and k.*

The proof is inspired by an observation in Liu & Scarlett (2019) that for a specifically chosen ReLU network (with offsets), the linear recovery problem considered here is equivalent to a group sparse recovery problem. The main difference here, though, are two folds: first, we need to tackle the scenario where the range of the generative network is restricted to a unit ball; second, our ReLU network (3) has *no offset*. The proof is postponed in Section §4.2 with more proof details in Supplement §C.

3.3. Global Landscape Analysis

In this section, we present the theoretical properties of the global landscape of the proposed empirical risk $L(x)$ in (4). We start by introducing some notations used in the rest of this paper. For any fixed x, we define $W_{+,x} := \operatorname{diag}(Wx > 0)W$, where $\operatorname{diag}(Wx > 0)$ is a diagonal matrix whose i-th diagonal entry is 1 if the product of the i-th row of W and x is positive, and 0 otherwise. Thus, $W_{+,x}$ retains the rows of W which has a positive product with x, and sets other rows to be all zeros. We further define $W_{i,+,x} := \operatorname{diag}(W_i W_{i-1,+,x} \cdots W_{1,+,x} x > 0) W_i$ recursively, where only active rows of W_i are kept, such that the ReLU network $G(x)$ defined in (3) can be equivalently rewritten as $G(x) = (\Pi_{i=1}^n W_{i,+,x})x := W_{n,+,x} W_{n-1,+,x} \cdots W_{1,+,x} x$. Next, we introduce the Weight Distribution Condition, which is widely used in recent works to analyze the landscape of different empirical risk (Hand & Voroninski, 2018; Hand et al., 2018; Huang et al., 2018).

Definition 3.7 (Weight Distribution Condition (WDC))**.** *A matrix W satisfies the Weight Distribution Condition with $\varepsilon_{\mathrm{wdc}} > 0$ if for any nonzero vectors $x, z \in \mathbb{R}^p$,*

$$\big\|W_{+,x}^\top W_{+,z} - Q_{x,z}\big\|_2 \leq \varepsilon_{\mathrm{wdc}},$$

where $Q_{x,z} := \frac{\pi - \angle(x,z)}{2\pi}\mathbf{I}_p + \frac{\sin\angle(x,z)}{2\pi} M_{\widehat{x}\leftrightarrow\widehat{z}}$ with $M_{\widehat{x}\leftrightarrow\widehat{z}}$ being the matrix[3] transforming $\widehat{x}$ to $\widehat{z}$, $\widehat{z}$ to $\widehat{x}$, and ϑ to 0 for any $\vartheta \in \operatorname{span}(\{x, z\})^\perp$. Here we denote $\widehat{x} = \frac{x}{\|x\|_2}$ and $\widehat{z} = \frac{z}{\|z\|_2}$ as normalized x and z.

Intuitively, the WDC characterizes the invertibility of the ReLU network in the sense that the output of each layer of the ReLU network nearly preserves the angle of any two input vectors. As is shown in Hand & Voroninski (2018), for any arbitrarily small $\varepsilon_{\mathrm{wdc}} > 0$, if the network is sufficiently expansive at each layer, namely $d_i \geq c d_{i-1} \log d_{i-1}$ for all $i \in [n]$ with d_i being polynomial on $\varepsilon_{\mathrm{wdc}}^{-1}$, and entries of W_i are i.i.d. $\mathcal{N}(0, 1/d_i)$, then $W_i \in \mathbb{R}^{d_i \times d_{i-1}}$ for all $i \in [n]$ satisfies WDC with constant $\varepsilon_{\mathrm{wdc}}$ with high probability. In particular, it does not require W_i and W_j to be independent for $i \neq j$. The question whether WDC is necessary for analyzing the computational aspect of the generative network remains open and warrants further studies.

Next, we present the Theorems 3.8 and 3.10. We denote the directional derivative along the direction of the non-zero vector z as $D_z L(x) = \lim_{t \to 0^+} \frac{L(x + t\widehat{z}) - L(x)}{t}$ with $\widehat{z} = \frac{z}{\|z\|_2}$. Specifically, $D_z L(x)$ equals $\langle \nabla L(x), \widehat{z}\,\rangle$ if $L(x)$ is differentiable at x and otherwise equals $\lim_{N \to +\infty} \langle \nabla L(x_N), \widehat{z}\,\rangle$. Here $\{x_N\}_{N \geq 0}$ is a sequence such that $x_N \to x$ and $L(x)$ is differentiable at any x_N. The existence of such a sequence is guaranteed by the piecewise linearity of $G(x)$. Particularly, the gradient of $L(x)$ is computed as $\nabla L(x) = 2(\Pi_{j=1}^n W_{j,+,x})^\top (\Pi_{j=1}^n W_{j,+,x})x - \frac{2\lambda}{m}\sum_{i=1}^m y_i (\Pi_{j=1}^n W_{j,+,x})^\top a_i$.

Theorem 3.8. *Suppose that $G(\cdot)$ is a ReLU network with weights W_i satisfying WDC with $\varepsilon_{\mathrm{wdc}}$ for all $i \in [n]$ where $n > 1$. Let $v_x = \lim_{x_N \to x} \nabla L(x_N)$ where $\{x_N\}$ is the sequence such that $\nabla L(x_N)$ exists for all x_N (and $v_x = \nabla L(x)$ if $L(x)$ is differentiable at x). If $\varepsilon_{\mathrm{wdc}}$ sastisfies $c_1 n^8 \varepsilon_{\mathrm{wdc}}^{1/4} \leq 1$, by setting $\lambda \geq 4C_{a,\xi,R} \cdot \log(64 C_{a,\xi,R} \cdot \varepsilon_{\mathrm{wdc}}^{-1})$ and $m \geq c_2 \|a\|_{\psi_1}^2 \lambda^2 \log^2(\lambda m)(kn\log(ed) + k\log(2R) + k\log m + u)/\varepsilon_{\mathrm{wdc}}^2$, then with probability $1 - c_3\exp(-u)$, for any nonzero x_0 satisfying $\|x_0\|_2 \leq R(1/2 + \varepsilon_{\mathrm{wdc}})^{-n/2}$, the directional derivatives satisfy*

***1.** If $\|x_0\|_2 > \check{\delta}$, then*

$$D_{-v_x} L(x) < 0, \ \ \forall x \notin \mathcal{B}(x_0, \delta_1) \cup \mathcal{B}(-\rho_n x_0, \delta_2) \cup \{0\},$$
$$D_w L(0) < 0, \qquad \forall w \neq 0.$$

***2.** If $\|x_0\|_2 \leq \check{\delta}$, then*

$$D_{-v_x} L(x) < 0, \ \forall x \notin \mathcal{B}(x_0, \delta_1) \cup \mathcal{B}(-\rho_n x_0, \delta_2) \cup \mathcal{B}(0, \check{\delta}),$$

where we have $\check{\delta} = 2^{n/2}\varepsilon_{\mathrm{wdc}}^{1/2}$, $\delta_1 = c_4 n^3 \varepsilon_{\mathrm{wdc}}^{1/4}\|x_0\|_2$, $\delta_2 = c_5 n^{14}\varepsilon_{\mathrm{wdc}}^{1/4}\|x_0\|_2$, and $0 < \rho_n \leq 1$ with $\rho_n \to 1$ as $n \to \infty$.

Remark 3.9 (Interpretation of Theorem 3.8)**.** *Note that in the above theorem, Case 1 indicates that the when the*

[3] The detailed definition of $M_{\widehat{x}\leftrightarrow\widehat{z}}$ is shown in Supplement §A.

magnitude of the true representation $\|x_0\|_2^2$ is larger than $\breve{\delta}^2 = \mathcal{O}(\varepsilon_{\text{wdc}})$ (signal x_0 is strong), the global minimum lies in small neighborhoods around x_0 and its scalar multiple $-\rho_n x_0$, while for any point outside the neighborhoods of x_0 and $-\rho_n x_0$, one can always find a direction with a negative directional derivative. Note that $x = 0$ is a local maximum due to $D_w L(0) < 0$ along any non-zero directions w. One the other hand, Case 2 implies that when $\|x_0\|_2^2$ is smaller than $\breve{\delta}^2$, the global minimum lies in the neighborhood around 0 (and thus around x_0). We will see in Theorem 3.10 that one can further pin down the global minimum around the true x_0 for Case 1.

The next theorem shows that in Case 1 of Theorem 3.8, under certain conditions, the true global minimum lies around the true representation x_0 instead of its negative multiple.

Theorem 3.10. *Suppose that $G(\cdot)$ is a ReLU network with wights W_i satisfying WDC with error ε_{wdc} for all $i \in [n]$ where $n > 1$. Assume that $c_1 n^3 \varepsilon_{\text{wdc}}^{1/4} \leq 1$, and x_0 is any nonzero vector satisfying $\|x_0\|_2 \leq R(1/2 + \varepsilon_{\text{wdc}})^{-n/2}$. Then, setting $\lambda \geq 4C_{a,\xi,R} \cdot \log(64 C_{a,\xi,R} \cdot \varepsilon_{\text{wdc}}^{-1})$ and $m \geq c_2 \|a\|_{\psi_1}^2 \lambda^2 \log^2(\lambda m)(kn\log(ed) + k\log(2R) + k\log m + u)/\varepsilon_{\text{wdc}}^2$, with probability $1 - 2c_3\exp(-u)$, for any x_0 such that $\|x_0\|_2 \geq \breve{\delta} = 2^{n/2}\varepsilon_{\text{wdc}}^{1/2}$, the risk $L(\cdot)$ satisfies*

$$L(x) < L(z), \ \ \forall x \in \mathcal{B}(\varphi x_0, \delta_3) \ \text{ and } \ \forall z \in \mathcal{B}(-\zeta x_0, \delta_3),$$

where φ, ζ are any scalars in $[\rho_n, 1]$ and $\delta_3 = c_4 n^{-5}\|x_0\|_2$. Particularly, we have that the radius $\delta_3 < \rho_n\|x_0\|_2, \forall n > 1$, such that $0 \notin \mathcal{B}(\varphi x_0, \delta_3)$ and $0 \notin \mathcal{B}(-\zeta x_0, \delta_3)$.

Remark 3.11 (Interpretation of Theorem 3.10)**.** *The significance of Theorem 3.10 are two folds: first, it shows that the value of the empirical risk $L(x)$ is always smaller around x_0 compared to its negative multiple $-\rho_n x_0$; second, when the network is sufficiently expansive such that ε_{wdc} is small, i.e. $cn^{19}\varepsilon_{\text{wdc}}^{1/4} \leq 1$, along with Case 1 in Theorem 3.8, we have $\mathcal{B}(x_0, \delta_1) \subseteq \mathcal{B}(\varphi x_0, \delta_3)$ and $\mathcal{B}(-\rho_n x_0, \delta_2) \subseteq \mathcal{B}(-\zeta x_0, \delta_3)$ for some φ and ζ, so that one can guarantee that the global minimum of $L(x)$ stays around x_0. Since we do not focus on optimizing the order of n in our results, further improvement of such a dependency will be one of our future works.*

For better understanding of the landscape analyzed in Theorem 3.8 and Theorem 3.10, we present an illustration of the landscape via a simulation in Supplement §A.

3.4. Connections with Invertibility of Neural Network

As a straightforward corollary to Theorems 3.8 and 3.10, we obtain the approximate invertibility of ReLU network under noisy quantized measurements. Specifically, previous results (Hand & Voroninski, 2018; Gilbert et al., 2017; Arora et al., 2015) show that under proper assumptions, one can invert the neural network (NN) and approximate x_0 by observing the outcome $G(x_0)$ and solving $\text{argmin}_{x\in\mathbb{R}^d} \|G(x) - G(x_0)\|_2$. Here, we consider a generalized version of the previous setting in the sense that instead of observing the full $G(x_0)$, we only observe the randomly probed and quantized information $\frac{\lambda}{m}\sum_{i=1}^m \text{sign}(\langle a_i, G(x_0)\rangle + \tau_i)a_i$. Theorems 3.8 and 3.10 essentially show that by solving following minimization problem: $\text{argmin}_{x\in\mathbb{R}^k} \left\|G(x) - \frac{\lambda}{m}\sum_{i=1}^m \text{sign}(\langle a_i, G(x_0)\rangle + \tau)a_i\right\|_2$, one can still invert the NN and approximate the true representation x_0.

On the other hand, without this random sensing vector a_i, it is not always possible to approximate x_0 via directly quantized measurements $\text{sign}([G(x_0)]_i + \tau_i), \forall i \in [d]$. A simple example would be a $G(x_0)$ which is exactly sparse (e.g. $G(x_0) = \sigma \circ ([\mathbf{I}_{k\times k} \ \mathbf{0}_{k\times(d-k)}]^\top x_0)$) and x_0 is entrywise positive. Then, $G(x_0)$ corresponds to a vector with first k entries being x_0 and other entries 0. In this case, the observations $\text{sign}([G(x_0)]_i + \tau_i), \forall i \in [d]$, are just $\text{sign}(x_{0,i} + \tau_i), \forall i \in [k]$, and 0 otherwise. It is then obvious to see that any estimation procedure would incur a constant error estimating x_0 regardless of the choices τ_i.

4. Proofs of Main Results

4.1. Proof of Theorem 3.2

Consider the excessive risk $L(x) - L(x_0)$ for any $x \in \mathbb{R}^k$. Our goal is to show that under the conditions that m is sufficiently large and λ is set properly, with high probability, for any $x \in \mathbb{R}^k$ and any $x_0 \in \mathbb{R}^k$ satisfying $\|G(x_0)\|_2 \leq R$, if $\|G(x) - G(x_0)\|_2 > \varepsilon$, then $L(x) - L(x_0) > 0$ holds. By proving this claim, we can get that for $\widehat{x}_m$, i.e. the solution to (4), satisfying $L(\widehat{x}_m) \leq L(x_0)$, then $\|G(\widehat{x}_m) - G(x_0)\|_2 \leq \varepsilon$ holds with high probability.

Recall that $\{(y_i, a_i)\}_{i=1}^m$ are m i.i.d. copies of (y, a) defined in (2). For abbreviation, across this section, we let

$$\Delta_{x,x_0}^G := G(x) - G(x_0). \tag{7}$$

Then, we have the following decomposition

$$\begin{aligned}
&L(x) - L(x_0)\\
&= \|G(x)\|_2^2 - \|G(x_0)\|_2^2 - \frac{2\lambda}{m}\sum_{i=1}^m y_i\langle a_i, \Delta_{x,x_0}^G\rangle\\
&= \underbrace{\|G(x)\|_2^2 - \|G(x_0)\|_2^2 - 2\lambda\mathbb{E}\big[y_i\langle a_i, \Delta_{x,x_0}^G\rangle\big]}_{\text{(I)}}\\
&\quad \underbrace{- \frac{2\lambda}{m}\sum_{i=1}^m \big(y_i\langle a_i, \Delta_{x,x_0}^G\rangle - \mathbb{E}\big[y_i\langle a_i, \Delta_{x,x_0}^G\rangle\big]\big)}_{\text{(II)}}.
\end{aligned}$$

The term (I) is the bias of the expected risk, and the term (II) is the variance resulting from the empirical risk. Thus, to see whether $L(x) - L(x_0) > 0$ when $\|\Delta_{x,x_0}^G\|_2 > \varepsilon$, we focus on showing the lower bound of term (I) and the upper bound of term (II). For term (I), we give its lower bound according to the following lemma.

Lemma 4.1. *Letting $K_{a,\xi,R} = \|a\|_{\psi_1} R + \|\xi\|_{\psi_1}$, there exists an absolute constant $c_1 > 0$ such that*

$$\begin{aligned}&\big|\mathbb{E}\big[y_i\langle a_i, \Delta^G_{x,x_0}\rangle\big] - \lambda^{-1}\langle G(x_0), \Delta^G_{x,x_0}\rangle\big| \\ &\quad\leq \sqrt{c_1 K_{a,\xi,R}}\big(\sqrt{2(\lambda+1)}+2\big)e^{-\lambda/(2K_{a,\xi,R})}\|\Delta^G_{x,x_0}\|_2.\end{aligned}$$

Moreover, $\forall \varepsilon \in (0,1)$, if $\lambda \geq 4C_{a,\xi,R}\cdot\log(64C_{a,\xi,R}\cdot\varepsilon^{-1})$ with $C_{a,\xi,R} = \max\{c_1K_{a,\xi,R}, 1\}$, and $\|\Delta^G_{x,x_0}\|_2 > \varepsilon$, then

$$\|G(x)\|_2^2 - \|G(x_0)\|_2^2 - 2\lambda\mathbb{E}\big[y_i\left\langle a_i, \Delta^G_{x,x_0}\right\rangle\big] \geq \frac{1}{2}\|\Delta^G_{x,x_0}\|_2^2.$$

It shows that term (I) $\geq \frac{1}{2}\|\Delta^G_{x,x_0}\|_2^2$ when $\|\Delta^G_{x,x_0}\|_2 > \varepsilon$. This lemma is proved via the ingredient of dithering, i.e., artificially adding the noise smooths the $\operatorname{sign}(\cdot)$ function. To see this, for a fixed V, it holds that

$$\mathbb{E}_\tau[\operatorname{sign}(V+\tau)] = \frac{V}{\lambda}\mathbf{1}_{\{|V|\leq\lambda\}} + \mathbf{1}_{\{V>\lambda\}} - \mathbf{1}_{\{V<-\lambda\}},$$

where the dithering noise $\tau \sim \operatorname{Unif}[-\lambda, +\lambda]$, and $\mathbf{1}_{\{\cdot\}}$ is an indicator function. As a consequence, $\mathbb{E}[y_i|a_i,\xi_i] = (\langle a_i, G(x_0)\rangle + \xi_i)/\lambda$ given that $|\langle a_i, G(x_0)\rangle + \xi_i|$ is not too large, and then Lemma 4.1 follows. Detailed proof can be found in Supplement §B.1.

Next, we present the analysis for showing the upper bound of the term (II), which is the key to proving Theorem 3.2. To give the upper bound of term (II), it suffices to bound the following supremum over all $x \in \mathbb{R}^k$ and all x_0 satisfying $x_0 \in \mathbb{R}^k$, $\|G(x_0)\|_2 \leq R$:

$$\sup \frac{\big|\frac{1}{m}\sum_{i=1}^m y_i\langle a_i, \Delta^G_{x,x_0}\rangle - \mathbb{E}\big[y_i\langle a_i, \Delta^G_{x,x_0}\rangle\big]\big|}{\|\Delta^G_{x,x_0}\|_2}. \tag{8}$$

Recall that $y_i = \operatorname{sign}(\langle a_i, G(x_0)\rangle + \xi_i + \tau_i)$. By symmetrization inequality (Lemma B.7 in the supplement), the following lemma readily implies the similar bound for (8).

Lemma 4.2. *Suppose Assumption 3.1 holds and the number of samples $m \geq c_2\|a\|^2_{\psi_1}\lambda^2\log^2(\lambda m)[kn\log(ed) + k\log(2R) + k\log m + u]/\varepsilon^2$ for some absolute constant c_2 large enough, then, with probability at least $1 - c\exp(-u)$,*

$$\sup_{x_0\in\mathbb{R}^k,\ \|G(x_0)\|_2\leq R, x\in\mathbb{R}^k} \frac{\big|\frac{1}{m}\sum_{i=1}^m \varepsilon_i y_i\langle a_i, \Delta^G_{x,x_0}\rangle\big|}{\|\Delta^G_{x,x_0}\|_2} \leq \frac{\varepsilon}{16\lambda},$$

where $\{\varepsilon_i\}_{i=1}^m$ are i.i.d. Rademacher random variables and $c > 0$ is an absolute constant.

We provide a proof sketch for Lemma 4.2 as below. Details can be found in Supplement §B.2. The main difficulty is the simultaneous supremum over both x_0 and x, whereas in ordinary uniform concentration bounds (e.g. in non-uniform recovery), one only requires to bound a supremum over x. The idea is to consider a δ-covering net over the set $G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R)$, namely $\mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)$, and bounding the supremum over each individual covering ball. The δ value has to be carefully chosen so as to achieve the following goals:

- We replace each $\operatorname{sign}(\langle a_i, G(x_0)\rangle + \xi_i + \tau_i)$ by $\operatorname{sign}(\langle a_i, G(v)\rangle + \xi_i + \tau_i)$, where $G(v)$ is the nearest point to $G(x_0)$ in the δ-net, and show that this supremum when fixing $G(v)$ is small. This is done via a "one-step chaining" argument making use of the piecewise linearity structure of G.

- We consider the gap of such a replacement, i.e., the sign changes when replacing $G(x_0)$ by $G(v)$, and show that the fraction of sign changes, namely $d_H(G(x_0), G(v)) := \frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\{\operatorname{sign}(\langle a_i, G(x_0)\rangle+\xi_i+\tau_i)\neq\operatorname{sign}(\langle a_i, G(v)\rangle+\xi_i+\tau_i)\}}$, is uniformly small for all $G(x_0)$ and $G(v)$ pairs. This can be rephrased as the *uniform hyperplane tessellation problem*: Given an accuracy level $\varepsilon > 0$, for any two points $\theta_1, \theta_2 \in G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R)$ such that $\|\theta_1 - \theta_2\|_2 \leq \delta$, what is the condition on m and δ such that $d_H(\theta_1,\theta_2) \leq \|\theta_1-\theta_2\|_2 + \varepsilon$? We answer this question with a tight sample bound on m in terms of ε by counting the number of linear pieces in $G(\cdot)$ with a VC dimension argument.

- We bound the error regarding a small portion of the indices $\{1, 2, \cdots, m\}$ for which the signs do change in the previous replacement, and take a union bound over the δ-net.

Proof of Theorem 3.2. By Lemma 4.2 and symmetrization inequality (Lemma B.7 in the supplement), one readily gets that (8) is bounded by $\varepsilon/8\lambda$ with probability at least $1 - c_3\exp(-u)$. This further implies the following bound

$$\frac{2\lambda}{m}\sum_{i=1}^m \big(y_i\langle a_i, \Delta^G_{x,x_0}\rangle - \mathbb{E}\big[y_i\langle a_i, \Delta^G_{x,x_0}\rangle\big]\big) \leq \frac{\varepsilon}{4}\|\Delta^G_{x,x_0}\|_2.$$

Thus, when $\|\Delta^G_{x,x_0}\|_2 > \varepsilon$, the left-hand side of the above inequality is further bounded by $\|\Delta^G_{x,x_0}\|_2^2/4$. Combining with Lemma 4.1, we finally obtain $L(x) - L(x_0) = \text{(I)} - \text{(II)} \geq \|\Delta^G_{x,x_0}\|_2^2/4 > 0$, if $\|\Delta^G_{x,x_0}\|_2 > \varepsilon$. Note that with high probability, this inequality holds for any $x \in \mathbb{R}^k$ and $x_0 \in \mathbb{R}^k$ satisfying $\|G(x_0)\|_2 \leq R$. This further implies $\|G(\widehat{x}_m) - G(x_0)\|_2 \leq \varepsilon$, which finishes the proof. □

4.2. Proof of Theorem 3.4

The key to proving Theorem 3.4 is to build a connection between our problem and the sparse recovery problem. Then we can further analyze the lower bound by employing tools from the area of sparse recovery. The detailed proofs for this subsection are presented in Supplement §C.

Definition 4.3. *A vector $v \in \mathbb{R}^d$ is k-group sparse if, when dividing v into k blocks of sub-vectors of size d/k,[4] each block has exactly one non-zero entry.*

We establish the following proposition to build a connection between the ReLU network and the group sparse vector.

Proposition 4.4. *Any nonnegative k-group sparse vector in $\mathbb{B}_2^d(1)$ can be generated by a ReLU network of the form (3) with a $k+1$ dimensional input and depth $n = 3$.*

The idea is to map each of the first k input entries into one block in $\mathbb{R}^d$ of length d/k respectively, and use the remaining one entry to construct proper offsets. We construct this mapping via a ReLU network *with no offset* as follows:

Consider a three-layer ReLU network, which has $k+1$ dimensional input of the form: $[x_1, \cdots, x_k, z]^\top \in \mathbb{R}^{k+1}$. The first hidden layer has the width of $(k + 2d/k)$ whose first k nodes outputs $\sigma(x_i)$, $\forall i \in [k]$, and the next $2d/k$ nodes output $\sigma(r \cdot z), \forall r \in [2d/k]$, which become the offset terms for the second layer. Then, with $\sigma(x_i)$ and $\sigma(r \cdot z)$ from the first layer, the second hidden layer will output the values of $\Upsilon_r(x_i, z) = \sigma(\sigma(x_i) - 2\sigma(r \cdot z))$ and $\Upsilon'_r(x_i, z) = \sigma(\sigma(x_i) - 2\sigma(r \cdot z) - \sigma(z))$, $\forall i \in [k]$ and $\forall r \in [d/k]$. Finally, by constructing the third layer, we have the following mapping: $\forall i \in [k]$ and $\forall r \in [d/k]$, $\Gamma_r(x_i, z) := \sigma\big(\Upsilon_r(x_i, z) - 2\Upsilon'_r(x_i, z)\big)$.

Note that $\Gamma_r(x_i, z)$ fires only when $x_i \geq 0$, in which case we have $\sigma(x_i) = x_i$. Letting z always equal to 1, we can observe that $\{\Gamma_r(x_i, 1)\}_{r=1}^{d/k}$ is a sequence of d/k non-overlapping triangle functions on the positive real line with width 2 and height 1. Therefore, the function $\Gamma_r(x_i, 1)$ can generate the value of the r-th entry in the i-th block of a nonnegative k-group sparse vector in $\mathbb{B}_2^d(1)$.

The above proposition implies that the set of nonnegative k-group sparse vectors in $\mathbb{B}_2^d(1)$ is the subset of $G(\mathbb{R}^{k+1}) \cap \mathbb{B}_2^d(1)$ where $G(\cdot)$ is defined by the mapping Γ.

Lemma 4.5. *Assume that $\theta_0 \in K \subseteq \mathbb{B}_2^d(1)$ where K is a set containing any k-group sparse vectors in $\mathbb{B}_2^d(1)$, and K satisfies that $\forall v \in K$ then $\lambda v \in K, \forall \lambda \in [0, 1)$. Assume that $\check{y} = \langle a, \theta_0 \rangle + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2)$ and $a \sim \mathcal{N}(0, \mathbf{I}_d)$. Then, there exist absolute constants $c_1, c_2 > 0$ such that any estimator $\widehat{\theta}$ which depends only on m observations of $(a, \check{y})$ satisfies that when $m \geq c_1 k \log(d/k)$, there is*

$$\sup_{\theta_0 \in K} \mathbb{E}\|\widehat{\theta} - \theta_0\|_2 \geq c_2 \sqrt{\frac{k \log(d/k)}{m}}.$$

Then, we are ready to show the proof of Theorem 3.4.

Proof of Theorem 3.4. According to Proposition 4.4, let $G(\cdot)$ be defined by the mapping Γ. One can verify that $G(\lambda x) = \lambda G(x)$, $\forall \lambda \geq 0$, by the positive homogeneity of ReLU network with no offsets. Letting $K = G(\mathbb{R}^{k+1}) \cap \mathbb{B}_2^d(1)$ and then by Lemma 4.5, we can obtain Theorem 3.4, which completes the proof. □

4.3. Proof Outline of Theorem 3.8 and Theorem 3.10

The key to proving Theorems 3.8 and 3.10 lies in understanding the concentration of $L(x)$ and $\nabla L(x)$. We prove two critical lemmas, Lemmas D.1 and D.2 in Supplement §D, to show that, with high probability, when λ and m are sufficiently large, for any x, z and x_0 such that $|G(x_0)| \leq R$, the following holds

$$\left| \left\langle \frac{\lambda}{m} \sum_{i=1}^m y_i a_i - G(x_0), H_x(z) \right\rangle \right| \leq \varepsilon \|H_x(z)\|_2, \qquad (9)$$

where $H_x(z) := \prod_{j=1}^n W_{j,+,x} z$ and $G(x) = H_x(x)$. In particular, this replaces the range restricted isometry condition (RRIC) adopted in previous works (Hand & Voroninski, 2018).

Under the conditions of Theorems 3.8 and 3.10, the inequality (9) essentially implies $\lambda/m \sum_{i=1}^m y_i \langle a_i, H_x(z) \rangle \approx \langle G(x_0), H_x(z) \rangle, \forall x, z$. Therefore, by definition of $L(x)$ in (4), we can approximate $\nabla L(x)$ and $L(x)$ as follows:

$$\langle \nabla L(x), z \rangle \approx 2\langle G(x), H_x(z) \rangle - 2\langle G(x_0), H_x(z) \rangle, \qquad (10)$$

$$L(x) \approx \|G(x_0)\|_2^2 - 2\langle G(x_0), G(x) \rangle. \qquad (11)$$

We give a sketch proof of Theorem 3.8 as follows. Please see Supplement §D for proof details.

- We show that $\forall x$, z, $\langle G(x), H_x(z) \rangle - \langle G(x_0), H_x(z) \rangle \approx \langle h_{x,x_0}, z \rangle$, where we define a certain approximation function $h_{x,x_0} := 2^{-n} x - 2^{-n} \big[\big(\prod_{i=0}^{n-1} \frac{\pi - \overline{\varrho}_i}{\pi} \big) x_0 + \sum_{i=0}^{n-1} \frac{\sin \overline{\varrho}_i}{\pi} \big(\prod_{j=i+1}^{d-1} \frac{\pi - \overline{\varrho}_j}{\pi} \big) \frac{\|x_0\|_2}{\|x\|_2} x \big]$ with $\overline{\varrho}_i = g(\overline{\varrho}_{i-1})$, $\overline{\varrho}_0 = \angle(x, x_0)$, and $g(\varrho) := \cos^{-1} \left(\frac{(\pi - \varrho) \cos \varrho + \sin \varrho}{\pi} \right)$ as shown in Lemmas D.3 and D.4. Combining with (10), we obtain $\langle \nabla L(x), z \rangle \approx 2\langle h_{x,x_0}, z \rangle$.
- With v_x being defined in Theorem 3.8, the directional derivative along the direction v_x is approximated as $D_{-v_x} L(x) \|v_x\|_2 \approx -4\|h_{x,x_0}\|_2^2$ following the previous step. Particularly, $\|h_{x,x_0}\|_2$ being small implies x is close to x_0 or $-\rho_n x_0$ by Lemma D.3 and $\|h_{x,x_0}\|_2$ gets small as $\|x_0\|_2$ approaches 0.
- We consider the error of approximating $D_{-v_x} L(x) \|v_x\|_2$ by $-4\|h_{x,x_0}\|_2^2$. When $\|x_0\|_2$ is not small, and $x \neq 0$, one can show the error is negligible compared to $-4\|h_{x,x_0}\|_2^2$, so that by the previous step, one finishes the proof of Case 1 when $x \neq 0$. On the other hand, for Case 2, when $\|x_0\|_2$ approaches 0, such an error is decaying slower than $-4\|h_{x,x_0}\|_2^2$ itself and eventually dominates it. As

[4] We assume WLOG that d/k is an integer.

a consequence, one can only conclude that $\widehat{x}_m$ is around the origin.

- For Case 1 when $x = 0$, we can show $D_w L(0) \cdot \|w\|_2 \leq |\langle G(x_0), H_{x_N}(w)\rangle - \lambda/m \sum_{i=1}^m y_i \langle a_i, H_{x_N}(w)\rangle| - \langle G(x_0), H_{x_N}(w)\rangle$ with $x_N \to 0$. By giving the upper bound of the first term and the lower bound of the second term according to (9) and Lemma D.4, we obtain $D_w L(0) < 0, \forall w \neq 0$.

Theorem 3.10 is proved in Supplement §E. We have the following proof sketch. We show by (11) that $L(x) \approx 2\langle h_{x,x_0}, x\rangle - \|G(x)\|_2^2$. With such approximation, by Lemmas E.1, E.2 in the supplement, under certain conditions, we have that if x and z are around x_0 and $-\rho_n x_0$ respectively, $L(x) < L(z)$ holds.

5. Conclusion

We consider the problem of one-bit compressed sensing via ReLU generative networks, in which $G : \mathbb{R}^k \to \mathbb{R}^d$ is an n-layer ReLU generative network with a low dimensional representation x_0 to $G(x_0)$. We propose to recover the target $G(x_0)$ solving an unconstrained empirical risk minimization problem. Under a weak sub-exponential measurement assumption, we establish a joint statistical and computational analysis. We prove that the ERM estimator in this new framework achieves a statistical rate of $m = \widetilde{\mathcal{O}}(kn \log d/\varepsilon^2)$ recovering any $G(x_0)$ uniformly up to an error ε. When the network is shallow, this rate matches the information-theoretic lower bound up to logarithm factors on ε^{-1}. Computationally, we prove that under proper conditions on the network weights, the proposed empirical risk has no stationary point outside of small neighborhoods around the true representation x_0 and its negative multiple. Under further assumptions on weights, we show that the global minimizer of the empirical risk stays within the neighborhood around x_0 rather than its negative multiple.

References

Ai, A., Lapanowski, A., Plan, Y., and Vershynin, R. One-bit compressed sensing with non-gaussian measurements. *Linear Algebra and its Applications*, 441:222–239, 2014.

Angluin, D. and Valiant, L. G. Fast probabilistic algorithms for Hamiltonian circuits and matchings. *Journal of Computer and system Sciences*, 18(2):155–193, 1979.

Arora, S., Liang, Y., and Ma, T. Why are deep nets reversible: A simple theory, with implications for training. *arXiv preprint arXiv:1511.05653*, 2015.

Aubin, B., Loureiro, B., Maillard, A., Krzakala, F., and Zdeborová, L. The spiked matrix model with generative priors. *arXiv preprint arXiv:1905.12385*, 2019.

Bora, A., Jalal, A., Price, E., and Dimakis, A. G. Compressed sensing using generative models. *arXiv preprint arXiv:1703.03208*, 2017.

Dirksen, S. and Mendelson, S. Non-gaussian hyperplane tessellations and robust one-bit compressed sensing. *arXiv preprint arXiv:1805.09409*, 2018a.

Dirksen, S. and Mendelson, S. Robust one-bit compressed sensing with partial circulant matrices. *arXiv preprint arXiv:1812.06719*, 2018b.

Gilbert, A. C., Zhang, Y., Lee, K., Zhang, Y., and Lee, H. Towards understanding the invertibility of convolutional neural networks. *arXiv preprint arXiv:1705.08664*, 2017.

Goldstein, L. and Wei, X. Non-Gaussian observations in nonlinear compressed sensing via Stein discrepancies. *Information and Inference: A Journal of the IMA*, 8(1): 125–159, 05 2018. ISSN 2049-8772. doi: 10.1093/imaiai/iay006. URL https://doi.org/10.1093/imaiai/iay006.

Goldstein, L., Minsker, S., and Wei, X. Structured signal recovery from non-linear and heavy-tailed measurements. *IEEE Transactions on Information Theory*, 64(8):5513–5530, 2018.

Hammernik, K., Klatzer, T., Kobler, E., Recht, M. P., Sodickson, D. K., Pock, T., and Knoll, F. Learning a variational network for reconstruction of accelerated mri data. *Magnetic resonance in medicine*, 79(6):3055–3071, 2018.

Hand, P. and Joshi, B. Global guarantees for blind demodulation with generative priors. *arXiv preprint arXiv:1905.12576*, 2019.

Hand, P. and Voroninski, V. Global guarantees for enforcing deep generative priors by empirical risk. In *Conference On Learning Theory*, pp. 970–978, 2018.

Hand, P., Leong, O., and Voroninski, V. Phase retrieval under a generative prior. In *Advances in Neural Information Processing Systems*, pp. 9136–9146, 2018.

Huang, W., Hand, P., Heckel, R., and Voroninski, V. A provably convergent scheme for compressive sensing under random generative priors. *arXiv preprint arXiv:1812.04176*, 2018.

Jacques, L., Laska, J. N., Boufounos, P. T., and Baraniuk, R. G. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Transactions on Information Theory*, 59(4):2082–2102, 2013.

Kamath, A., Karmalkar, S., and Price, E. Lower bounds for compressed sensing with generative models. 2019.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114. IEEE, 2017.

Lei, N., Luo, Z., Yau, S.-T., and Gu, D. X. Geometric understanding of deep learning. *arXiv preprint arXiv:1805.10451*, 2018.

Liu, Z. and Scarlett, J. Information-theoretic lower bounds for compressive sensing with generative models. *arXiv preprint arXiv:1908.10744*, 2019.

Manoel, A., Krzakala, F., Mézard, M., and Zdeborová, L. Multi-layer generalized linear estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 2098–2102. IEEE, 2017.

Pandit, P., Sahraee-Ardakan, M., Rangan, S., Schniter, P., and Fletcher, A. K. Inference with deep generative priors in high dimensions. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Plan, Y. and Vershynin, R. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59 (1):482–494, 2013.

Plan, Y. and Vershynin, R. Dimension reduction by random hyperplane tessellations. *Discrete & Computational Geometry*, 51(2):438–461, 2014.

Plan, Y., Vershynin, R., and Yudovina, E. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2016.

Sønderby, C. K., Caballero, J., Theis, L., Shi, W., and Huszár, F. Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, 2016.

Thrampoulidis, C. and Rawat, A. S. The generalized lasso for sub-gaussian measurements with dithered quantization. *arXiv preprint arXiv:1807.06976*, 2018.

Wei, X., Yang, Z., and Wang, Z. On the statistical rate of nonlinear recovery in generative models with heavy-tailed data. In *International Conference on Machine Learning*, pp. 6697–6706, 2019.

Wellner, J. et al. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.

Winder, R. Partitions of n-space by hyperplanes. *SIAM Journal on Applied Mathematics*, 14(4):811–818, 1966.

Xu, C. and Jacques, L. Quantized compressive sensing with rip matrices: The benefit of dithering. *arXiv preprint arXiv:1801.05870*, 2018.

Yang, G., Yu, S., Dong, H., Slabaugh, G., Dragotti, P. L., Ye, X., Liu, F., Arridge, S., Keegan, J., Guo, Y., et al. Dagan: Deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction. *IEEE transactions on medical imaging*, 37(6):1310–1321, 2018.

Yeh, R. A., Chen, C., Yian Lim, T., Schwing, A. G., Hasegawa-Johnson, M., and Do, M. N. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5485–5493, 2017.

Zhang, L., Yi, J., and Jin, R. Efficient algorithms for robust one-bit compressive sensing. In *International Conference on Machine Learning*, pp. 820–828, 2014.

Zhu, R. and Gu, Q. Towards a lower sample complexity for robust one-bit compressed sensing. In *International Conference on Machine Learning*, pp. 739–747, 2015.