Sample Complexity Bounds for 1-bit Compressive Sensing and Binary Stable Embeddings with Generative Priors

Zhaoqiang Liu [1] **Selwyn Gomes** [1] **Avtansh Tiwari** [2] **Jonathan Scarlett** [1]

Abstract

The goal of standard 1-bit compressive sensing is to accurately recover an unknown sparse vector from binary-valued measurements, each indicating the sign of a linear function of the vector. Motivated by recent advances in compressive sensing with generative models, where a generative modeling assumption replaces the usual sparsity assumption, we study the problem of 1-bit compressive sensing with generative models. We first consider noiseless 1-bit measurements, and provide sample complexity bounds for approximate recovery under i.i.d. Gaussian measurements and a Lipschitz continuous generative prior, as well as a near-matching algorithm-independent lower bound. Moreover, we demonstrate that the Binary ϵ-Stable Embedding property, which characterizes the robustness of the reconstruction to measurement errors and noise, also holds for 1-bit compressive sensing with Lipschitz continuous generative models with sufficiently many Gaussian measurements. In addition, we apply our results to neural network generative models, and provide a proof-of-concept numerical experiment demonstrating significant improvements over sparsity-based approaches.

1. Introduction

The compressive sensing (CS) problem (Foucart & Rauhut, 2013; Wainwright, 2019), which aims to recover a *sparse* signal from a small number of linear measurements, is fundamental in machine learning, signal processing and statistics. It has been popular over the past 1–2 decades and has become increasingly well-understood, with theoretical guarantees including sharp performance bounds for both practical algorithms (Amelunxen et al., 2014; Donoho et al., 2013; Wainwright, 2009b; Wen et al., 2016) and potentially intractable information-theoretically optimal algorithms (Arias-Castro et al., 2013; Candes & Davenport, 2013; Scarlett & Cevher, 2017; Wainwright, 2009a).

Unlike conventional compressive sensing, which assumes infinite-precision real-valued measurements, in *1-bit* compressive sensing (Boufounos & Baraniuk, 2008), each measurement is quantized to a single bit, namely its sign. Considerable research effort has been placed to 1-bit compressive sensing (Ai et al., 2014; Awasthi et al., 2016; Gopi et al., 2013; Gupta et al., 2010; Zhang et al., 2014; Zhu & Gu, 2015), one motivation being that 1-bit quantization can be implemented in hardware with low cost and is robust to certain nonlinear distortions (Boufounos, 2010).

In addition, motivated by recent advances in deep generative models (Foster, 2019), a new perspective has recently emerged in CS, in which the sparsity assumption is replaced by the assumption that the underlying signal lies near the range of a suitably-chosen generative model, typically corresponding to a deep neural network (Bora et al., 2017). Along with several theoretical developments, it has been numerically verified that generative priors can reduce the number of measurements required for a given accuracy by large factors such as 5 to 10 (Bora et al., 2017).

In this paper, following the developments in both 1-bit CS and CS with generative priors, we establish a variety of fundamental theoretical guarantees for 1-bit compressive sensing using generative models.

1.1. Related Work

Sparsity-based 1-bit compressive sensing: The framework of 1-bit compressive sensing (CS) was introduced and studied in (Boufounos & Baraniuk, 2008). Subsequently, various numerical algorithms were designed (Boufounos, 2009; 2010; Boufounos & Baraniuk, 2008; Laska et al., 2011; Zymnis et al., 2009), often with convergence guarantees. In addition, several works have developed theoretical guarantees for support recovery and approximate vector recovery in 1-bit CS (Acharya et al., 2017; Gopi et al., 2013; Plan & Vershynin, 2012; 2013; Zhang et al.,

[1]National University of Singapore, Singapore [2]Indian Institute of Technology (IIT) Kanpur, India. Correspondence to: Jonathan Scarlett <scarlett@comp.nus.edu.sg>, Zhaoqiang Liu <dcslizha@nus.edu.sg>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

2014; Zhu & Gu, 2015). In these works, it is usually assumed that the measurement matrix contains i.i.d. Gaussian entries. Such an assumption is generalized to allow sub-Gaussian (Ai et al., 2014; Dirksen & Mendelson, 2018) and log-concave (Awasthi et al., 2016) measurement matrices. A survey on 1-bit CS can be found in (Li et al., 2018).

To address the fact that standard 1-bit measurements give no information about the norm of the underlying signal vector, the so-called dithering technique, which adds artificial random noise before quantization, has been considered (Dirksen & Mendelson, 2018; Jacques & Cambareri, 2017; Knudson et al., 2016; Xu & Jacques, 2019) to also enable the estimation of the norm.

Perhaps most relevant to the present paper, (Jacques et al., 2013) studies the robustness of 1-bit CS by considering binary stable embeddings of sparse vectors. We seek to provide analogous theoretical guarantees to those in (Jacques et al., 2013), but with a generative prior in place of the sparsity assumption. We adopt similar high-level proof steps, but with significantly different details.

Compressive sensing with generative models: Bora *et al.* (Bora et al., 2017) show that roughly $O(k \log L)$ random Gaussian linear measurements suffice for accurate recovery when the generative model is an L-Lipschitz function with bounded k-dimensional inputs. The analysis in (Bora et al., 2017) is based on minimizing an empirical loss function. In practice, such a task may be hard, and the authors propose to use a simple gradient descent algorithm in the latent space. The theoretical analysis is based on showing that Gaussian random matrices satisfy a natural counterpart to the Restricted Eigenvalue Condition (REC) termed the Set-REC. Follow-up works of (Bora et al., 2017) provide various additional algorithmic guarantees for compressive sensing with generative models (Dhar et al., 2018; Hand & Voroninski, 2018; Latorre et al., 2019; Liu & Scarlett, 2020a; Peng et al., 2020; Shah & Hegde, 2018), as well as information-theoretic lower bounds (Kamath et al., 2019; Liu & Scarlett, 2020b).

In a recent work, the authors of (Qiu et al., 2019) study robust 1-bit compressive sensing with ReLU-based generative models. In particular, the authors design an empirical risk minimization algorithm, and prove that it is able to faithfully recover bounded target vectors produced by the model from quantized noisy measurements. Our results and those of (Qiu et al., 2019) are complementary to each other, with several differences in the setup:

- In (Qiu et al., 2019), the dithering technique is used, adding artificial random noise before quantization to enable the recovery the norm of the signal vector, whereas we do not consider the use of dithering. Both settings are of interest depending on whether dithering is feasible to implement in the application at hand.
- In (Qiu et al., 2019), the focus is on ReLU networks without offset terms, whereas we consider general L-Lipschitz generative models.
- The pre-quantization noise in (Qiu et al., 2019) is assumed to be sub-exponential, whereas we allow for general and possibly adversarial noise.
- The theoretical analysis in (Qiu et al., 2019) focuses on a particular recovery algorithm, whereas our results are information-theoretic in nature.

1.2. Contributions

In this paper, we establish a variety of fundamental theoretical guarantees for 1-bit compressive sensing using generative models. Our main results are outlined as follows:

- In Section 2.1, for noiseless measurements, we characterize the number of i.i.d. Gaussian measurements sufficient (i.e., an upper bound on the sample complexity) to attain approximate recovery of the underlying signal under a Lipschitz continuous generative prior.
- In Section 2.2, for noiseless measurements, we show that our upper bound is nearly tight by giving a near-matching algorithm-independent lower bound for a particular Lipschitz continuous generative model.
- In Section 3, we establish the Binary ϵ-Stable Embedding (BϵSE) property, which characterizes the reconstruction robustness to measurement errors and noise. Specifically, we characterize the number of i.i.d. Gaussian measurements sufficient to ensure that this property holds. In Section 4, we specialize these results to feedforward neural network generative models.
- In Section 5, we propose a practical iterative algorithm for 1-bit CS with generative priors, and demonstrate its effectiveness in a simple numerical example.

1.3. Notation

We use upper and lower case boldface letters to denote matrices and vectors respectively. We write $[N] = \{1, 2, \cdots, N\}$ for a positive integer N. A *generative model* is a function $G : \mathcal{D} \to \mathbb{R}^n$, with latent dimension k, ambient dimension n, and input domain $\mathcal{D} \subseteq \mathbb{R}^k$. $\mathcal{S}^{n-1} := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = 1\}$ represents the unit sphere in $\mathbb{R}^n$. For $\mathbf{x}, \mathbf{s} \in \mathcal{S}^{n-1}$, $\mathrm{d_S}(\mathbf{x}, \mathbf{s}) := \frac{1}{\pi} \arccos\langle \mathbf{x}, \mathbf{s} \rangle$ denotes the geodesic distance, which is the normalized angle between vectors $\mathbf{x}$ and $\mathbf{s}$. For $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^m$, $\mathrm{d_H}(\mathbf{v}, \mathbf{v}') := \frac{1}{m} \sum_{i=1}^m \mathbf{1}\{v_i \neq v_i'\}$ denotes the Hamming distance. We use $\|\mathbf{X}\|_{2\to 2}$ to denote the spectral norm of a matrix $\mathbf{X}$. We define the ℓ_2-ball $B_2^k(r) := \{\mathbf{z} \in \mathbb{R}^k : \|\mathbf{z}\|_2 \leq r\}$, and the ℓ_∞-ball $B_\infty^k(r) := \{\mathbf{z} \in \mathbb{R}^k : \|\mathbf{z}\|_\infty \leq r\}$. For a set

$B \subseteq \mathbb{R}^k$ and a generative model $G : \mathbb{R}^k \to \mathbb{R}^n$, we write $G(B) = \{G(\mathbf{z}) : \mathbf{z} \in B\}$.

2. Noiseless Measurements

In this section, we derive near-matching upper and lower bounds on the sample complexity in the noiseless setting, in which the measurements take the form

$$\mathbf{b} = \Phi(\mathbf{x}) := \mathrm{sign}(\mathbf{A}\mathbf{x}) \tag{1}$$

for some measurement matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and unknown underlying signal $\mathbf{x} \in G(B_2^k(r))$, where $G : B_2^k(r) \to \mathcal{S}^{n-1}$ is the generative model.

Remark 1 *In the following, for clarity, we will assume that the range of the generative model is contained in the unit sphere, i.e., $G(B_2^k(r)) \subseteq \mathcal{S}^{n-1}$, and provide guarantees of the form $\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \le \epsilon$ for some estimate $\hat{\mathbf{x}}$. While the preceding assumption may appear restrictive, these results readily transfer to any general (unnormalized) generative model $\tilde{G}$ with $\tilde{G}(B_2^k(r)) \subseteq \mathbb{R}^n$ when the recovery guarantee is modified to $\left\|\frac{\mathbf{x}}{\|\mathbf{x}\|_2} - \frac{\hat{\mathbf{x}}}{\|\hat{\mathbf{x}}\|_2}\right\|_2 \le \epsilon$ (note that norm estimation is impossible under 1-bit measurements of the form (1)). The idea is to apply our results to $G(\mathbf{x}) = \frac{\tilde{G}(\mathbf{x})}{\|\tilde{G}(\mathbf{x})\|_2}$; see Section 4 for an example and further discussion.*

In addition, we consider spherical domains with radius r. Similar to that in (Bora et al., 2017), the assumption of a bounded domain is mild, since the dependence on r in the sample complexity will only be logarithmic. In addition, our lower bound will show that such a dependence on r is unavoidable.

2.1. Upper Bound

Our first main result shows that with sufficiently many independent Gaussian measurements, with high probability, any two signals separated by some specified distance ϵ produce distinct measurements. This amounts to an upper bound on the sample complexity for noiseless 1-bit recovery.

Theorem 1 *Fix $r > 0$ and $\epsilon \in (0, 1)$, and let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be generated as $A_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Suppose that the generative model $G : \mathbb{R}^k \to \mathbb{R}^n$ is L-Lipschitz and $G(B_2^k(r)) \subseteq \mathcal{S}^{n-1}$. For $m = \Omega\left(\frac{k}{\epsilon}\log\frac{Lr}{\epsilon^2}\right)$,[1] with probability at least $1 - e^{-\Omega(\epsilon m)}$, we have for all $\mathbf{x}, \mathbf{s} \in G(B_2^k(r))$ that*

$$\|\mathbf{x} - \mathbf{s}\|_2 > \epsilon \Rightarrow \mathrm{d_H}(\Phi(\mathbf{x}), \Phi(\mathbf{s})) = \Omega(\epsilon). \tag{2}$$

In particular, if $\|\mathbf{x} - \mathbf{s}\|_2 > \epsilon$, then $\Phi(x) \ne \Phi(\mathbf{s})$.

The proof is outlined below, with the full details given in the supplementary material. From Theorem 1, we immediately obtain the following corollary giving a recovery guarantee for noiseless 1-bit compressive sensing.

Corollary 1 *Let $\mathbf{A}$ and G follow the same assumptions as those given in Theorem 1. Then, for a fixed $\epsilon \in (0, 1)$, when $m = \Omega\left(\frac{k}{\epsilon}\log\frac{Lr}{\epsilon^2}\right)$, the following holds with probability at least $1 - e^{-\Omega(\epsilon m)}$: For any $\mathbf{x} \in G(B_2^k(r))$ and its noiseless measurements $\mathbf{b} = \Phi(\mathbf{x})$, any estimate $\hat{\mathbf{x}} \in G(B_2^k(r))$ such that $\Phi(\hat{\mathbf{x}}) = \mathbf{b}$ satisfies*

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \le \epsilon. \tag{3}$$

In addition, following similar ideas to those in the proof of Theorem 1, we obtain the following corollary, which provides a supplementary guarantee to that of Theorem 1. The proof can be found in the supplementary material.

Corollary 2 *Let $\mathbf{A}$ and G follow the same assumptions as those given in Theorem 1. Then, for a fixed $\epsilon \in (0, 1)$, if $m = \Omega\left(\frac{k}{\epsilon}\log\frac{Lr}{\epsilon^2}\right)$, with probability at least $1 - e^{-\Omega(\epsilon m)}$, for all $\mathbf{x}, \mathbf{s} \in G(B_2^k(r))$, it holds that*

$$\|\mathbf{x} - \mathbf{s}\|_2 \le \epsilon \Rightarrow \mathrm{d_H}(\Phi(\mathbf{x}), \Phi(\mathbf{s})) \le O(\epsilon). \tag{4}$$

Remark 2 *Combining the results of Theorem 1 and Corollary 2, we arrive at the so-called Local Binary Embedding property. This property is of independent interest, e.g., see (Oymak & Recht, 2015), and will also be used as a stepping stone to a stronger binary embedding property in Section 3.1. Briefly, the distinction is that the local binary embedding property can be interpreted as "If $\mathbf{x}$ is close to $\mathbf{s}$ then $\Phi(\mathbf{x})$ is close to $\Phi(\mathbf{s})$ (and vice versa)", whereas in Section 3.1 we seek a stronger statement of the form "The distance between $\mathbf{x}$ and $\mathbf{s}$ always approximately equals the distance between $\Phi(\mathbf{x})$ and $\Phi(\mathbf{s})$".*

2.1.1. Proof Outline for Theorem 1

To prove Theorem 1, we follow the technique used in (Bora et al., 2017) to construct a chain of nets for $G(B_2^k(r))$, and approximate $\mathbf{x}$ using a point $\mathbf{x}_0$ in one of the ϵ-nets (and similarly, approximating $\mathbf{s}$ using $\mathbf{s}_0$). We can control various terms consisting of points in the ϵ-nets using probabilistic arguments and the union bound. Before providing a more detailed outline, we state some useful auxiliary results.

Lemma 1 (Plan & Vershynin, 2013, Lemma 4.4) *Let $\mathbf{x}, \mathbf{s} \in \mathcal{S}^{n-1}$ and assume that $\|\mathbf{x} - \mathbf{s}\|_2 \ge \epsilon$ for some $\epsilon > 0$. Let $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. Then for $\epsilon_0 = \frac{\epsilon}{12}$, we have*

$$\mathbb{P}(\langle\mathbf{a}, \mathbf{x}\rangle > \epsilon_0, \langle\mathbf{a}, \mathbf{s}\rangle < -\epsilon_0) \ge \epsilon_0. \tag{5}$$

This result essentially states that if two unit vectors are far apart, then for a random hyperplane, the probability of a certain level of separation can be lower bounded. In addition, we will use the following concentration inequality.

[1] In all statements of the form $m = \Omega(\cdot)$ in our upper bounds, the implied constant is implicitly assumed to be sufficiently large.

Lemma 2 (Vempala, 2005, Lemma 1.3) *Let $\mathbf{x} \in \mathbb{R}^n$, and assume that the entries in $\mathbf{A} \in \mathbb{R}^{m \times n}$ are sampled independently from $\mathcal{N}(0,1)$. Then, for any $\epsilon \in (0,1)$, we have*

$$\mathbb{P}\left((1-\epsilon)\|\mathbf{x}\|_2^2 \le \left\|\frac{1}{\sqrt{m}}\mathbf{A}\mathbf{x}\right\|_2^2 \le (1+\epsilon)\|\mathbf{x}\|_2^2\right) \ge 1 - 2e^{-\epsilon^2(1-\epsilon)m/4}. \quad (6)$$

The following definition formally introduces the notion of an ϵ-net, also known as a covering set.

Definition 1 *Let $(\mathcal{X}, d)$ be a metric space, and fix $\epsilon > 0$. A subset $S \subseteq \mathcal{X}$ is said be an* ϵ-net *of $\mathcal{X}$ if, for all $\mathbf{x} \in \mathcal{X}$, there exists some $\mathbf{s} \in S$ such that $d(\mathbf{x}, \mathbf{s}) \le \epsilon$.*

With the above auxiliary results in place, the proof of Theorem 1 is outlined as follows:

1. For a fixed $\delta > 0$ and a positive integer l, let $M = M_0 \subseteq M_1 \subseteq \ldots \subseteq M_l$ be a chain of nets of $B_2^k(r)$ such that M_i is a $\frac{\delta_i}{L}$-net with $\delta_i = \frac{\delta}{2^i}$. There exists such a chain of nets with (Vershynin, 2010, Lemma 5.2)

$$\log |M_i| \le k \log \frac{4Lr}{\delta_i}. \quad (7)$$

   Then, by the L-Lipschitz assumption on G, we have for any $i \in [l]$ that $G(M_i)$ is a δ_i-net of $G(B_2^k(r))$.

2. For $\mathbf{x} \in G(B_2^k(r))$, we write $\mathbf{x} = (\mathbf{x} - \mathbf{x}_l) + (\mathbf{x}_l - \mathbf{x}_{l-1}) + \ldots + (\mathbf{x}_1 - \mathbf{x}_0) + \mathbf{x}_0$ with $\mathbf{x}_i \in G(M_i)$ and $\|\mathbf{x} - \mathbf{x}_l\|_2 \le \frac{\delta}{2^l}$, $\|\mathbf{x}_i - \mathbf{x}_{i-1}\|_2 \le \frac{\delta}{2^{i-1}}$, $i \in [l]$. The triangle inequality yields $\|\mathbf{x} - \mathbf{x}_0\|_2 \le 2\delta$. We apply similar reasoning to a second signal $\mathbf{s} \in G(B_2^k(r))$ with $\|\mathbf{x} - \mathbf{s}\|_2 > \epsilon$, and choose $\delta = O(\epsilon^2)$ sufficiently small so that $\|\mathbf{x}_0 - \mathbf{s}_0\|_2 > \frac{\epsilon}{2}$. This allows us to apply Lemma 1 to get

$$\mathbb{P}\left(\langle \underline{\mathbf{a}}_i, \mathbf{x}_0\rangle > \frac{\epsilon}{24}, \langle \underline{\mathbf{a}}_i, \mathbf{s}_0\rangle < -\frac{\epsilon}{24}\right) \ge \frac{\epsilon}{24} \quad (8)$$

   for any $i \in [m]$, where $\underline{\mathbf{a}}_i$ is the i-th row of $\mathbf{A}$. Since the tests are independent, we can use binomial concentration to deduce that at least an $\Omega(\epsilon)$ fraction of the measurements satisfy the condition in (8), with probability $1 - e^{-\Omega(\epsilon m)}$. Then, by (7) and a union bound, the same holds simultaneously for *all* $(\mathbf{x}', \mathbf{s}') \in G(M) \times G(M)$ with high probability.

3. We use the Cauchy-Schwartz inequality and triangle inequality to obtain the following decomposition:

$$\frac{1}{m}\sum_{i=1}^m |\langle \underline{\mathbf{a}}_i, \mathbf{x} - \mathbf{x}_0\rangle| \le \sum_{i=1}^l \left\|\frac{1}{\sqrt{m}}\mathbf{A}(\mathbf{x}_i - \mathbf{x}_{i-1})\right\|_2 + \left\|\frac{1}{\sqrt{m}}\mathbf{A}(\mathbf{x} - \mathbf{x}_l)\right\|_2, \quad (9)$$

   and upper bound the two terms as follows:

   (a) For the first term, we use Lemma 2 and a union bound over the signals in the i-th and $(i-1)$-th nets to upper bound each summand by $\big(1 + \frac{\epsilon_i}{2}\big)\frac{\delta}{2^{i-1}}$ with high probability, for some $\epsilon_1, \ldots, \epsilon_l$. We show that a choice of the form $\epsilon_i^2 = O\big(\epsilon + \frac{ik}{m}\big)$ suffices to take the overall term down to $O(\delta)$.

   (b) For the second term, we upper bound the spectral norm of $\mathbf{A}$ by $2 + \sqrt{\frac{n}{m}}$ with high probability, and show that when this bound holds, $l = O(\log n)$ suffices to bring the overall term down to $O(\delta)$.

   This argument holds uniformly in $\mathbf{x}$, and we apply the resulting bound to both signals $\mathbf{x}, \mathbf{s}$ under consideration. The choice $\delta = O(\epsilon^2)$ allows us to deduce that a fraction $1 - O(\epsilon)$ of the measurements satisfy $|\langle \underline{\mathbf{a}}_i, \mathbf{x} - \mathbf{x}_0\rangle| + |\langle \underline{\mathbf{a}}_i, \mathbf{s} - \mathbf{s}_0\rangle| \le O(\epsilon)$. The implied constant in this fraction of measurements is carefully designed to be smaller than that in the $\Omega(\epsilon)$ fraction of Step 2.

4. We combine Steps 2 and 3 to show that a fraction $\Omega(\epsilon)$ of the measurements satisfy *both* of the conditions therein, and we show that $\mathrm{sign}(\langle \underline{\mathbf{a}}_i, \mathbf{x}\rangle) \ne \mathrm{sign}(\langle \underline{\mathbf{a}}_i, \mathbf{s}\rangle)$ for every such measurement. As a result, we find that $\mathrm{d_H}(\Phi(\mathbf{x}), \Phi(\mathbf{s})) \ge \Omega(\epsilon)$, as desired.

2.2. Lower Bound

In this subsection, we address the question of whether the upper bound in Theorem 1 can be improved. To do this, following the approach of (Liu & Scarlett, 2020b), we consider a specific L-Lipschitz generative model, and derive an algorithm-independent lower bound on the number of samples required to accurately recover signals from this model. This result is formally stated as follows.

Theorem 2 *Fix $r > 0$ and $L = \Omega\big(\frac{1}{r}\big)$ with a sufficiently large implied constant, and $\epsilon \in \big(0, \frac{\sqrt{3}}{4\sqrt{2}}\big)$. Then, there exists an L-Lipschitz generative model with input domain $B_2^k(r)$ such that for any measurement matrix $\mathbf{A}$ and decoder producing an estimate $\hat{\mathbf{x}}$ such that $\sup_{\mathbf{x} \in G(B_2^k(r))} \|\mathbf{x} - \hat{\mathbf{x}}\|_2 \le \epsilon$, it must be the case that $m = \Omega\left(k \log(Lr) + \frac{k}{\epsilon}\right)$.*

The proof is given in the supplementary material, and is briefly outlined as follows. We follow the high-level approach from (Liu & Scarlett, 2020b) of choosing a generative model that can produce group-sparse signals, with suitable normalization to ensure that all signals lie on the unit sphere. Both the $\Omega\left(\frac{k}{\epsilon}\right)$ and $\Omega\left(k\log(Lr)\right)$ lower bounds are established by choosing a *hard subset* of signals, and comparing its size to the number of possible output sequences:

- For the $\Omega\left(\frac{k}{\epsilon}\right)$ bound, following (Acharya et al., 2017), we consider packing as many signals as possible onto a unit sphere corresponding to the subspace of an arbitrary single sparsity pattern, and we bound the number of output sequences using a result from (Jacques et al., 2013) on the number of orthants of $\mathbb{R}^m$ intersected by a single lower-dimensional subspace.

- For the $\Omega\left(k\log(Lr)\right)$ bound, we use the Gilbert-Varshamov bound to show that there exist $e^{\Omega(k\log\frac{n}{k})}$ sequences separated by a constant distance, and trivially upper bound the number of output sequences by 2^m. This gives an $m = \Omega\big(k\log\frac{n}{k}\big)$ lower bound, which reduces to $\Omega(k\log(Lr))$ upon calculating the Lipschitz constant of our chosen generative model.

Remark 3 *In Theorem 1, the sample complexity derived is $\Omega\left(\frac{k}{\epsilon}\log\frac{Lr}{\epsilon^2}\right)$. Comparing with the lower bound provided in Theorem 2, we observe that when $\epsilon = \Theta(1)$ the upper and lower bounds match, and when $\epsilon = o(1)$, they match up to a logarithmic factor in $\frac{Lr}{\epsilon^2}$.*

Remark 4 *A recent result in (Flodin et al., 2019) suggests that the presence of separate $\frac{k}{\epsilon}$ and $k\log(Lr)$ terms (as opposed to a combined term such as $\frac{k}{\epsilon}\log(Lr)$) is the correct behavior in certain cases. Specifically, it is shown that in the case of sparse signals, one can indeed achieve $m = O\big(\frac{k}{\epsilon} + k\log\frac{n}{k}\big)$ by moving beyond i.i.d. Gaussian measurement matrices. However, the technique is based on first identifying a superset of the sparse support, and it is unclear what a suitable counterpart would be in the case of general generative models.*

3. Binary Embeddings and Noisy Measurements

Thus far, we have considered recovery guarantees under noiseless measurements. In this section, we turn to the *Binary ϵ-Stable Embedding* (BϵSE) property (defined below), which roughly requires the binary measurements to preserve the geometry of signals produced by the generative model. Similarly to the case of sparse signals (Jacques et al., 2013), we will see that this permits 1-bit CS recovery guarantees even in the presence of random or adversarial noise.

Definition 2 *Let $\epsilon \in (0,1)$. A mapping $\Phi(\cdot) : \mathbb{R}^n \to \{-1,1\}^m$ is a Binary ϵ-Stable Embedding (BϵSE) for vectors in $G(B_2^k(r)) \subseteq \mathcal{S}^{n-1}$ if*

$$\mathrm{d_S}(\mathbf{x},\mathbf{s}) - \epsilon \le \mathrm{d_H}(\Phi(\mathbf{x}),\Phi(\mathbf{s})) \le \mathrm{d_S}(\mathbf{x},\mathbf{s}) + \epsilon \tag{10}$$

for all $\mathbf{x},\mathbf{s} \in G(B_2^k(r))$, where $\mathrm{d_S}$ is the geodesic distance (cf., Section 1.3).

3.1. Establishing the BϵSE Property

Our main goal in this section is to prove the following theorem, which gives the BϵSE property.

Theorem 3 *Let $\mathbf{A}$ and G follow the same assumptions as those given in Theorem 1. For a fixed $\epsilon \in (0,1)$, if $m = \Omega\left(\frac{k}{\epsilon^2}\log\frac{Lr}{\epsilon}\right)$, then with probability at least $1 - e^{-\Omega(\epsilon^2 m)}$, we have for all $\mathbf{x},\mathbf{s} \in G(B_2^k(r))$ that*

$$|\mathrm{d_S}(\mathbf{x},\mathbf{s}) - \mathrm{d_H}(\Phi(\mathbf{x}),\Phi(\mathbf{s}))| \le \epsilon. \tag{11}$$

In the proof of Theorem 3, we construct an ϵ-net and use $\mathbf{x}_0,\mathbf{s}_0$ in the net to approximate $\mathbf{x},\mathbf{s}$. We use the triangle inequality to decompose $|\mathrm{d_S}(\mathbf{x},\mathbf{s}) - \mathrm{d_H}(\Phi(\mathbf{x}),\Phi(\mathbf{s}))|$ into three terms: $|\mathrm{d_H}(\Phi(\mathbf{x}),\Phi(\mathbf{s})) - \mathrm{d_H}(\Phi(\mathbf{x}_0),\Phi(\mathbf{s}_0))|$, $|\mathrm{d_H}(\Phi(\mathbf{x}_0),\Phi(\mathbf{s}_0)) - \mathrm{d_S}(\mathbf{x}_0,\mathbf{s}_0)|$ and $|\mathrm{d_S}(\mathbf{x},\mathbf{s}) - \mathrm{d_S}(\mathbf{x}_0,\mathbf{s}_0)|$. We derive an upper bound for the first term by using Corollary 2 to bound $\mathrm{d_H}(\Phi(\mathbf{x}),\Phi(\mathbf{x}_0))$ and $\mathrm{d_H}(\Phi(\mathbf{s}),\Phi(\mathbf{s}_0))$. The second term is upper bounded using a concentration bound from (Jacques et al., 2013) and a union bound for all $(\mathbf{x}_0,\mathbf{s}_0)$ pairs in the ϵ-net. The third term is directly upper bounded via the definition of an ϵ-net.

Before formalizing this outline, we introduce the following useful lemmas.

Lemma 3 (Goemans & Williamson, 1995, Lemma 3.2) *Suppose that $\mathbf{r}$ is drawn uniformly from the unit sphere $\mathcal{S}^{n-1}$. Then for any two fixed vectors $\mathbf{x},\mathbf{s} \in \mathcal{S}^{n-1}$, we have*

$$\mathbb{P}\left(\mathrm{sign}(\langle\mathbf{x},\mathbf{r}\rangle) \ne \mathrm{sign}(\langle\mathbf{s},\mathbf{r}\rangle)\right) = \mathrm{d_S}(\mathbf{x},\mathbf{s}). \tag{12}$$

Based on this lemma, the following lemma concerning the geodesic distance and the Hamming distance follows via a concentration argument.

Lemma 4 (Jacques et al., 2013, Lemma 2) *For fixed $\mathbf{x},\mathbf{s} \in \mathcal{S}^{n-1}$ and $\epsilon > 0$, we have*

$$\mathbb{P}\left(|\mathrm{d_H}(\Phi(\mathbf{x}),\Phi(\mathbf{s})) - \mathrm{d_S}(\mathbf{x},\mathbf{s})| \le \epsilon\right) \ge 1 - 2e^{-\epsilon^2 m}, \tag{13}$$

where the probability is with respect to the generation of $\mathbf{A}$ with i.i.d. standard normal entries.

Note that in Lemma 4 the vectors $\mathbf{x},\mathbf{s}$ are fixed in advance, before the sample matrix $\mathbf{A}$ is drawn. In contrast, for Theorem 3, we need to consider drawing $\mathbf{A}$ first and then choosing $\mathbf{x},\mathbf{s}$ arbitrarily.

In addition, we have the following simple lemma, which states that the Euclidean norm distance and geodesic distance are almost equivalent for vectors on the unit sphere.

Lemma 5 *For any $\mathbf{x},\mathbf{s} \in \mathcal{S}^{n-1}$, we have*

$$\frac{1}{\pi}\|\mathbf{x}-\mathbf{s}\|_2 \le \mathrm{d_S}(\mathbf{x},\mathbf{s}) \le \frac{1}{2}\|\mathbf{x}-\mathbf{s}\|_2. \tag{14}$$

Proof Let $d = \|\mathbf{x}-\mathbf{s}\|_2$. Using the definition of geodesic distance and $\langle\mathbf{x},\mathbf{s}\rangle = \frac{1}{2}\big(\|\mathbf{x}\|_2^2 + \|\mathbf{s}\|_2^2 - \|\mathbf{x}-\mathbf{s}\|_2^2\big)$, we have

$\mathrm{d_S}(\mathbf{x},\mathbf{s}) = \frac{1}{\pi}\arccos\left(1-\frac{d^2}{2}\right)$. It is straightforward to show that $1-\frac{2}{\pi^2}x^2 \ge \cos x \ge 1-\frac{x^2}{2}$ for any $x \in [0,\pi]$. In addition, letting $a = \arccos\left(1-\frac{d^2}{2}\right)$, we have $1-\frac{a^2}{2} \le \cos a = 1-\frac{d^2}{2} \le 1-\frac{2}{\pi^2}a^2$, which implies $d \le a \le \frac{\pi}{2}d$. Therefore, $\frac{d}{\pi} \le \mathrm{d_S}(\mathbf{x},\mathbf{s}) = \frac{a}{\pi} \le \frac{d}{2}$. ■

We now proceed with the proof of Theorem 3.

Proof of Theorem 3. Let M be an $\frac{\epsilon}{L}$-net of $B_2^k(r)$ such that $\log|M| \le k\log\frac{4Lr}{\epsilon}$. By the L-Lipschitz property of G, $G(M)$ is a ϵ-net of $G(B_2^k(r))$. Hence, for any $\mathbf{x},\mathbf{s} \in G(B_2^k(r))$, there exist $\mathbf{x}_0,\mathbf{s}_0 \in G(M)$ such that

$$\|\mathbf{x}-\mathbf{x}_0\|_2 \le \epsilon, \quad \|\mathbf{s}-\mathbf{s}_0\|_2 \le \epsilon. \tag{15}$$

By Lemma 5, we have

$$\mathrm{d_S}(\mathbf{x},\mathbf{x}_0) \le \frac{\epsilon}{2}, \quad \mathrm{d_S}(\mathbf{s},\mathbf{s}_0) \le \frac{\epsilon}{2}, \tag{16}$$

and hence, by the triangle inequality,

$$|\mathrm{d_S}(\mathbf{x},\mathbf{s}) - \mathrm{d_S}(\mathbf{x}_0,\mathbf{s}_0)| \le \epsilon. \tag{17}$$

In addition, by Lemma 4 and the union bound, if we set $m = \Omega\left(\frac{k}{\epsilon^2}\log\frac{Lr}{\epsilon}\right)$, with probability at least $1-|M|^2 e^{-\epsilon^2 m} = 1-e^{-\Omega(\epsilon^2 m)}$, we have

$$|\mathrm{d_H}(\Phi(\mathbf{u}),\Phi(\mathbf{v})) - \mathrm{d_S}(\mathbf{u},\mathbf{v})| \le \epsilon \tag{18}$$

for all $(\mathbf{u},\mathbf{v}) \in G(M)\times G(M)$. Furthermore, by Corollary 2 and (15), if $m = \Omega\left(\frac{k}{\epsilon}\log\frac{Lr}{\epsilon^2}\right)$, then with probability at least $1-e^{-\Omega(\epsilon m)}$, we have

$$\mathrm{d_H}(\Phi(\mathbf{x}),\Phi(\mathbf{x}_0)) \le C\epsilon, \quad \mathrm{d_H}(\Phi(\mathbf{s}),\Phi(\mathbf{s}_0)) \le C\epsilon, \tag{19}$$

where C is a positive constant. (Note that the result of Corollary 2 holds *uniformly* for signals in $G(B_2^k(r))$.) Using the two upper bounds in (19) and applying the triangle inequality in the same way as (17), we obtain

$$|\mathrm{d_H}(\Phi(\mathbf{x}),\Phi(\mathbf{s})) - \mathrm{d_H}(\Phi(\mathbf{x}_0),\Phi(\mathbf{s}_0))| \le 2C\epsilon. \tag{20}$$

Combining (17), (18) and (20), we obtain that if $m = \Omega\left(\frac{k}{\epsilon^2}\log\frac{Lr}{\epsilon} + \frac{k}{\epsilon}\log\frac{Lr}{\epsilon^2}\right)$, with probability at least $1 - e^{-\Omega(\epsilon^2 m)} - e^{-\Omega(\epsilon m)}$, the following holds uniformly in $\mathbf{x},\mathbf{s} \in G(B_2^k(r))$:

$$\begin{aligned}
&|\mathrm{d_H}(\Phi(\mathbf{x}),\Phi(\mathbf{s})) - \mathrm{d_S}(\mathbf{x},\mathbf{s})| \\
&\quad\le |\mathrm{d_H}(\Phi(\mathbf{x}),\Phi(\mathbf{s})) - \mathrm{d_H}(\Phi(\mathbf{x}_0),\Phi(\mathbf{s}_0))| \\
&\qquad + |\mathrm{d_H}(\Phi(\mathbf{x}_0),\Phi(\mathbf{s}_0)) - \mathrm{d_S}(\mathbf{x}_0,\mathbf{s}_0)| \\
&\qquad + |\mathrm{d_S}(\mathbf{x},\mathbf{s}) - \mathrm{d_S}(\mathbf{x}_0,\mathbf{s}_0)| \\
&\quad\le 2C\epsilon + \epsilon + \epsilon \\
&= 2(C+1)\epsilon.
\end{aligned} \tag{21}$$

Then, recalling that $Lr = \Omega(1)$, and scaling ϵ by $2(C+1)$, we deduce that when $m = \Omega\left(\frac{k}{\epsilon^2}\log\frac{Lr}{\epsilon}\right)$, we have with probability at least $1-e^{-\Omega(\epsilon^2 m)}$ that

$$|\mathrm{d_H}(\Phi(\mathbf{x}),\Phi(\mathbf{s})) - \mathrm{d_S}(\mathbf{x},\mathbf{s})| \le \epsilon. \tag{22}$$

■

3.2. Implications for Noisy 1-bit CS

Here we demonstrate that Theorem 3 implies recovery guarantees for 1-bit CS in the case of noisy measurements. In particular, we have the following corollary.

Corollary 3 *Let* $\mathbf{A}$ *and* G *follow the same assumptions as those given in Theorem 1. For an* $\epsilon \in (0,1)$*, if* $m = \Omega\left(\frac{k}{\epsilon^2}\log\frac{Lr}{\epsilon}\right)$*, then with probability at least* $1-e^{-\Omega(\epsilon^2 m)}$*, we have the following: For any* $\mathbf{x} \in G(B_2^k(r))$*, if* $\tilde{\mathbf{b}} := \mathrm{sign}(\mathbf{A}\mathbf{x})$ *and* $\mathbf{b}$ *is any vector of corrupted measurements satisfying* $\mathrm{d_H}(\mathbf{b},\tilde{\mathbf{b}}) \le \tau_1$*, then any* $\hat{\mathbf{x}} \in G(B_2^k(r))$ *with* $\mathrm{d_H}(\mathrm{sign}(\mathbf{A}\hat{\mathbf{x}}),\mathbf{b}) \le \tau_2$ *satisfies*

$$\mathrm{d_S}(\mathbf{x},\hat{\mathbf{x}}) \le \epsilon + \tau_1 + \tau_2. \tag{23}$$

Proof By Theorem 3, we have

$$|\mathrm{d_S}(\mathbf{x},\hat{\mathbf{x}}) - \mathrm{d_H}(\Phi(\mathbf{x}),\Phi(\hat{\mathbf{x}}))| \le \epsilon. \tag{24}$$

In addition, by the triangle inequality and $\tilde{\mathbf{b}} = \Phi(\mathbf{s})$,

$$\mathrm{d_H}(\Phi(\mathbf{x}),\Phi(\hat{\mathbf{x}})) \le \mathrm{d_H}(\mathbf{b},\tilde{\mathbf{b}}) + \mathrm{d_H}(\mathrm{sign}(\mathbf{A}\hat{\mathbf{x}}),\mathbf{b}) \le \tau_1+\tau_2. \tag{25}$$

Combining (24)–(25) gives (23). ■

Corollary 3 gives a guarantee for arbitrary (possibly adversarial) perturbations of $\tilde{\mathbf{b}}$ to produce $\mathbf{b}$. Naturally, this directly implies high-probability bounds on the recovery error in the case of random noise. For instance, if $\mathbf{b} = \mathrm{sign}(\mathbf{A}\mathbf{x}+\boldsymbol{\xi})$ with $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_m)$, then by (Jacques et al., 2013, Lemma 4), we have for any $\gamma > 0$ that

$$\mathbb{P}\left[\mathrm{d_H}(\tilde{\mathbf{b}},\mathbf{b}) > \frac{\sigma}{2} + \gamma\right] \le e^{-2m\gamma^2}. \tag{26}$$

Analogous results can be derived for other noise distributions such as Poisson noise, random sign flips, and so on. In addition, we may derive upper bounds on $\mathrm{d_H}(\mathrm{sign}(\mathbf{A}\hat{\mathbf{x}}),\mathbf{b})$ for specific algorithms. For example, for algorithms with consistent sign constraints, we have $\mathrm{d_H}(\mathrm{sign}(\mathbf{A}\hat{\mathbf{x}}),\mathbf{b}) = 0$, which corresponds to $\tau_2 = 0$ in Corollary 3.

4. Neural Network Generative Models

In this section, we consider feedforward neural network generative models. Such a model $\tilde{G}: \mathbb{R}^k \to \mathbb{R}^n$ with d layers can be written as

$$\tilde{G}(\mathbf{z}) = \phi_d\left(\phi_{d-1}\left(\cdots\phi_2(\phi_1(\mathbf{z},\boldsymbol{\theta}_1),\boldsymbol{\theta}_2)\cdots,\boldsymbol{\theta}_{d-1}\right),\boldsymbol{\theta}_d\right), \tag{27}$$

where $\mathbf{z} \in B_2^k(r)$, $\phi_i(\cdot)$ is the functional mapping corresponding to the i-th layer, and $\boldsymbol{\theta}_i = (\mathbf{W}_i, \mathbf{b}_i)$ is the parameter pair for the i-th layer: $\mathbf{W}_i \in \mathbb{R}^{n_i \times n_{i-1}}$ is the matrix of weights, and $\mathbf{b}_i \in \mathbb{R}^{n_i}$ is the vector of offsets, where n_i is the number of neurons in the i-th layer. Note that $n_0 = k$ and $n_d = n$. Defining $\mathbf{z}^0 = \mathbf{z}$ and $\mathbf{z}^i = \phi_i(\mathbf{z}^{i-1}, \boldsymbol{\theta}_i)$, we set $\phi_i(\mathbf{z}^{i-1}, \boldsymbol{\theta}_i) = \phi_i(\mathbf{W}_i \mathbf{z}^{i-1} + \mathbf{b}_i)$, $i = 1, 2, \dots, d$, for some operation $\phi_i(\cdot)$ applied element-wise.

The function $\phi_i(\cdot)$ is referred to as the activation function for the i-th layer, with popular choices including (i) the ReLU function, $\phi_i(x) = \max(x, 0)$; (ii) the Sigmoid function, $\phi_i(x) = \frac{1}{1+e^{-x}}$; and (iii) the Hyperbolic tangent function with $\phi_i(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. Note that for each of these examples, $\phi_i(\cdot)$ is 1-Lipschitz.

To establish Lipschitz continuity of the entire network, we can utilize the following standard result.

Lemma 6 *Consider any two functions f and g. If f is L_f-Lipschitz and g is L_g-Lipschitz, then their composition $f \circ g$ is $L_f L_g$-Lipschitz.*

Suppose that $\tilde{G}$ is defined as in (27) with at most w nodes per layer. We assume that all weights are upper bounded by $W_{\max}$ in absolute value, and that the activation functions are 1-Lipschitz. Then, from Lemma 6, we obtain that $\tilde{G}$ is $\tilde{L}$-Lipschitz with $\tilde{L} = (wW_{\max})^d$ (*cf.* (Bora et al., 2017, Lemma 8.5)). Since we consider normalized signals having unit norm, we limit our attention to signals in $\mathrm{Range}(\tilde{G})$ with norm at least $R_{\min}$, for some small $R_{\min} > 0$, so as to control the Lipschitz constant of $\frac{\tilde{G}(\mathbf{z})}{\|\tilde{G}(\mathbf{z})\|_2}$. We obtain the following from Theorem 3.

Theorem 4 *Suppose that $\mathbf{A} \in \mathbb{R}^{m \times n}$ is generated with $A_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and the generative model $\tilde{G} : B_2^k(r) \to \mathbb{R}^n$ is defined as in (27) with at most w nodes per layer. Suppose that all weights are upper bounded by $W_{\max}$ in absolute value, and that the activation function is 1-Lipschitz. Then, for fixed $\epsilon \in (0,1)$ and $R_{\min} > 0$, if $m = \Omega\big(\frac{k}{\epsilon^2} \log \frac{r(wW_{\max})^d}{\epsilon R_{\min}}\big)$, with probability at least $1 - e^{-\Omega(\epsilon^2 m)}$, we have the following: For any $\mathbf{x} \in \tilde{G}(B_2^k(r)) \setminus B_2^n(R_{\min})$, let $\tilde{\mathbf{b}} := \mathrm{sign}(\mathbf{A}\mathbf{x})$ be its uncorrupted measurements, and let $\mathbf{b}$ be any corrupted measurements satisfying $\mathrm{d_H}(\mathbf{b}, \tilde{\mathbf{b}}) \le \tau_1$. Then, any $\hat{\mathbf{x}} \in \tilde{G}(B_2^k(r)) \setminus B_2^n(R_{\min})$ with $\mathrm{d_H}(\mathrm{sign}(\mathbf{A}\hat{\mathbf{x}}), \mathbf{b}) \le \tau_2$ satisfies*

$$\mathrm{d_S}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|_2}, \frac{\hat{\mathbf{x}}}{\|\hat{\mathbf{x}}\|_2}\right) \le \epsilon + \tau_1 + \tau_2. \tag{28}$$

Proof Let $\tilde{\mathcal{D}} = \{\mathbf{z} \in B_2^k(r) : \|\tilde{G}(\mathbf{z})\|_2 > R_{\min}\}$, and define $G(\mathbf{z}) := \frac{\tilde{G}(\mathbf{z})}{\|\tilde{G}(\mathbf{z})\|_2}$ for $\mathbf{z} \in \tilde{\mathcal{D}}$. Observe that $G(\tilde{\mathcal{D}}) \subseteq \mathcal{S}^{n-1}$. In addition, by Lemma 6 and the assumption $\|\tilde{G}(\mathbf{z})\|_2 \ge R_{\min}$, we have that G is L-Lipschitz on $\tilde{\mathcal{D}}$ with $L = \frac{(wW_{\max})^d}{R_{\min}}$. Recall that Theorem 3 and Corollary 3 are proved by forming an ϵ-net of $B_2^k(r)$. Since an ϵ-net for a given set implies a 2ϵ-net of the same size (or smaller) for any subset, the same results hold (with a near-identical proof) when the domain of the generative model is restricted to $\tilde{\mathcal{D}} \subseteq B_2^k(r)$. Thus, the desired result follows by applying Corollary 3 to G with the restricted domain $\tilde{\mathcal{D}}$. ■

Remark 5 *The dependence on $R_{\min}$ in the sample complexity is very mild; for instance, under the typical scaling of $(wW_{\max})^d = n^{O(d)}$ (Bora et al., 2017), the scaling laws remain unchanged even with $R_{\min} = \frac{1}{n^{O(d)}}$. In addition, for common types of data such as images and audio, vectors with a very low norm are not of significant practical interest (e.g., a flat audio signal or an image of all black pixels).*

5. Efficient Algorithm & Numerical Example

In (Jacques et al., 2013), an algorithm termed Binary Iterative Hard Thresholding (BIHT) was proposed for 1-bit compressive sensing of sparse signals. In the case of a generative prior, we can adapt the BIHT algorithm by replacing the hard thresholding step by a projection onto the generative model. This gives the following iterative procedure:

$$\mathbf{x}^{(t+1)} = \mathcal{P}_G\left(\mathbf{x}^{(t)} + \lambda \mathbf{A}^T(\mathbf{b} - \mathrm{sign}(\mathbf{A}\mathbf{x}^{(t)}))\right), \tag{29}$$

where $\mathcal{P}_G(\cdot)$ is the projection function onto $G(B_2^k(r))$, $\mathbf{b}$ is the observed vector, $\mathbf{x}^{(0)} = \mathbf{0}$, and $\lambda > 0$ is a parameter. A counterpart to (29) for the linear model was also recently proposed in (Shah & Hegde, 2018).

It has been shown in (Jacques et al., 2013) that the quantity $\mathbf{A}^T(\mathrm{sign}(\mathbf{A}\mathbf{x}) - \mathbf{b})$ is a subgradient of the convex one-sided ℓ_1-norm $2\|[\mathbf{b} \odot (\mathbf{A}\mathbf{x})]_-\|_1$, where "$\odot$" denotes the element-wise product and $[x]_- = \min\{x, 0\}$. Therefore, the BIHT algorithm can be viewed as a projected gradient descent (PGD) algorithm that attempts to minimize $\|[\mathbf{b} \odot (\mathbf{A}\mathbf{x})]_-\|_1$. In addition, as argued in (Jacques & Vleeschouwer, 2013), there exist certain promising properties suggesting the stability and convergence of the BIHT algorithm.

Numerical example. While our main contributions are theoretical, in the following we present a simple proof-of-concept experiment for the MNIST dataset. The dataset consists of 60,000 handwritten images, each of size 28x28 pixels. The variational autoencoder (VAE) model uses a pre-trained VAE with a latent dimension of $k = 20$. The encoder and decoder both have the structure of a fully connected neural network with two hidden layers.

The projection step in (29) is approximated using gradient descent, performed using the Adam optimizer with 200 steps and a learning rate of 0.1. The update of $\mathbf{x}^{(t)}$ in (29) is done with a step size of $\lambda = 1.25$, with a total of 15 iterations. To reduce the impact of local minima, we choose the best

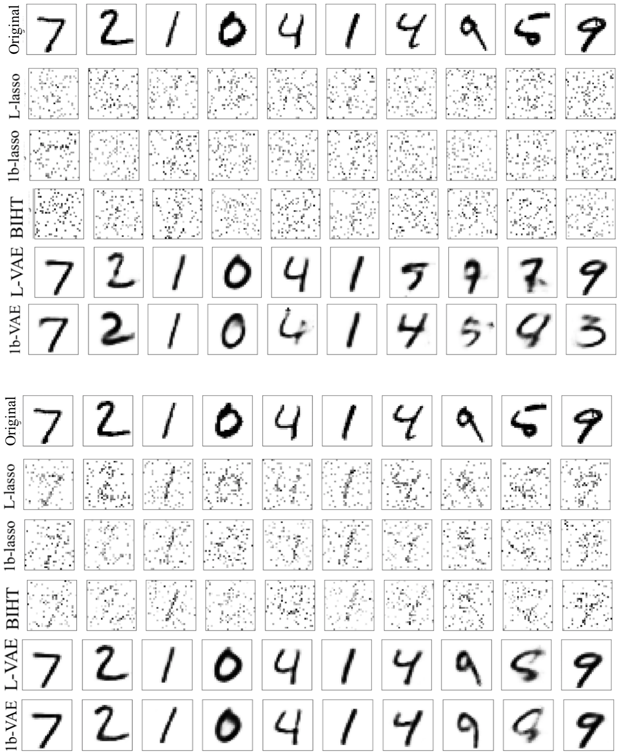

Figure 1. Examples of reconstructed images with 50 measurements (top) and reconstruction with 200 measurements (bottom) on the MNIST dataset.

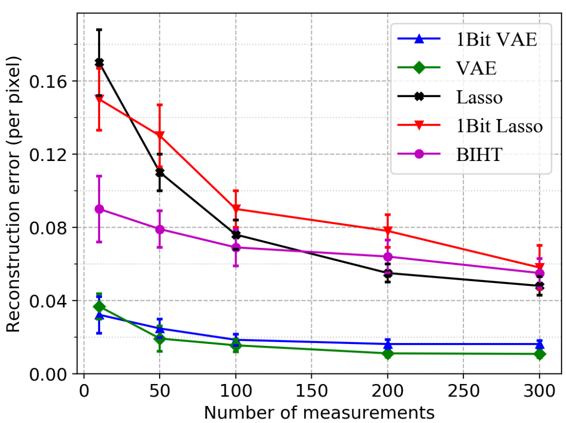

Figure 2. Average reconstruction error (per pixel) of the images from the MNIST dataset shown in Figure 1. The error bars indicate half of a standard deviation.

estimate among 4 random restarts. The reconstruction error is calculated over 10 images by averaging the per-pixel error in terms of the ℓ_2-norm. In accordance with our theoretical results, we focus on i.i.d. Gaussian measurements.

In Figure 1, we provide some examples of reconstructed images under both linear measurements and 1-bit measurements,[2] using sparsity based algorithms (Lasso (Tibshirani, 1996), 1-bit Lasso (Plan & Vershynin, 2013), and BIHT (Jacques et al., 2013)) and generative prior based algorithms (linear PGD (Shah & Hegde, 2018) and 1-bit PGD as per (29)). For convenience, we re-state the Lasso and 1-bit Lasso optimization problems here: We solve

$$\text{minimize} \ \ \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{b} = \mathbf{A}\mathbf{x} \tag{30}$$

for linear measurements, and we solve

$$\text{minimize} \ \ \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{b} = \text{sign}(\mathbf{A}\mathbf{x}), \|\mathbf{A}\mathbf{x}\|_1 = m \tag{31}$$

for 1-bit measurements. As discussed in (Plan & Vershynin, 2013), the second constraint can be viewed as normalization that prevents a zero or near-zero solution.

We observe from Figure 1 that all three sparsity-based methods attain poor reconstructions even when $m = 200$. In contrast, the generative prior based methods attain mostly accurate reconstructions even when $m = 50$, and highly accurate constructions when $m = 200$.

In this experiment, the loss due to the 1-bit quantization appears to be mild, and this is corroborated in Figure 2, where we plot the average per-pixel reconstruction error as a function of the number of measurements, averaged over the images show in Figure 1. For both generative model based methods, the reconstruction error eventually saturates, most likely due to *representation error* (i.e., the generative model being unable to perfectly produce the signal) (Bora et al., 2017). In addition, the curve for 1-bit measurements saturates to a slightly higher value than that of linear measurements, most likely due to the impossibility of estimating the norm. However, at least for this particular dataset, the gap between the two remains small.

6. Conclusion

We have established sample complexity bounds for both noiseless and noisy 1-bit compressive sensing with generative models. In the noiseless case, we showed that the sample complexity for ϵ-accurate recovery is between $\Omega\left(k \log(Lr) + \frac{k}{\epsilon}\right)$ and $O\left(\frac{k}{\epsilon} \log \frac{Lr}{\epsilon^2}\right)$. For noisy measurements, we showed that the binary ϵ-stable embedding property can be attained with $m = O\left(\frac{k}{\epsilon^2} \log \frac{Lr}{\epsilon}\right)$. An immediate direction for further research is to establish a matching lower bound for the latter result, though we are unaware of any such result even for the simpler case of sparse signals.

Acknowledgment. This work was supported by the Singapore National Research Foundation (NRF) under grant number R-252-000-A74-281.

[2] We used the PGD implementation of (Shah & Hegde, 2018) available at `https://github.com/shahviraj/pgdgan`, along with the pre-trained generative model and Lasso implementation of (Bora et al., 2017) available at `https://github.com/AshishBora/csgm`, and adapted these to their 1-bit variants.

References

Acharya, J., Bhattacharyya, A., and Kamath, P. Improved bounds for universal one-bit compressive sensing. In *Int. Symp. Inf. Theory (ISIT)*, pp. 2353–2357, 2017.

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

Amelunxen, D., Lotz, M., McCoy, M. B., and Tropp, J. A. Living on the edge: Phase transitions in convex programs with random data. *Inf. Inference*, 3(3):224–294, 2014.

Arias-Castro, E., Candes, E. J., and Davenport, M. A. On the fundamental limits of adaptive sensing. *IEEE Trans. Inf. Theory*, 59(1):472–481, Jan. 2013.

Awasthi, P., Balcan, M. F., Haghtalab, N., and Zhang, H. Learning and 1-bit compressed sensing under asymmetric noise. *J. Mach. Learn. Res.*, 49(June):152–192, 2016.

Bora, A., Jalal, A., Price, E., and Dimakis, A. G. Compressed sensing using generative models. In *Int. Conf. Mach. Learn. (ICML)*, pp. 537–546, 2017.

Boufounos, P. T. Greedy sparse signal reconstruction from sign measurements. In *Conf. Rec. Asilomar. Conf. Sig. Syst. Comput. (ACSSC)*, pp. 1305–1309. IEEE, 2009.

Boufounos, P. T. Reconstruction of sparse signals from distorted randomized measurements. In *Int. Conf. Acoust. Sp. Sig. Proc. (ICASSP)*, pp. 3998–4001, 2010.

Boufounos, P. T. and Baraniuk, R. G. 1-bit compressive sensing. In *Conf. Inf. Sci. Syst. (CISS)*, pp. 16–21. IEEE, 2008.

Candes, E. J. and Davenport, M. A. How well can we estimate a sparse vector? *Appl. Comp. Harm. Analysis*, 34(2):317–323, 2013.

Dhar, M., Grover, A., and Ermon, S. Modeling sparse deviations for compressed sensing using generative models. In *Int. Conf. Mach. Learn. (ICML)*, 2018.

Dirksen, S. and Mendelson, S. Non-Gaussian hyperplane tessellations and robust one-bit compressed sensing. *https://arxiv.org/abs/1805.09409*, 2018.

Donoho, D. L., Javanmard, A., and Montanari, A. Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Trans. Inf. Theory*, 59(11):7434–7464, Nov. 2013. ISSN 0018-9448.

Flodin, L., Gandikota, V., and Mazumdar, A. Superset technique for approximate recovery in one-bit compressed sensing. In *Conf. Neur. Inf. Proc. Sys. (NeurIPS)*, pp. 10387–10396, 2019.

Foster, D. *Generative Deep Learning : Teaching Machines to Paint, Write, Compose and Play*. O'Reilly Media, Inc, USA, 2019.

Foucart, S. and Rauhut, H. *A Mathematical Introduction to Compressive Sensing*. Springer New York, 2013.

Goemans, M. X. and Williamson, D. P. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42 (6):1115–1145, 1995.

Gopi, S., Netrapalli, P., Jain, P., and Nori, A. One-bit compressed sensing: Provable support and vector recovery. In *Int. Conf. Mach. Learn. (ICML)*, pp. 154–162, 2013.

Gupta, A., Nowak, R., and Recht, B. Sample complexity for 1-bit compressed sensing and sparse classification. In *Int. Symp. Inf. Theory (ISIT)*, pp. 1553–1557. IEEE, 2010.

Hand, P. and Voroninski, V. Global guarantees for enforcing deep generative priors by empirical risk. In *Conf. Learn. Theory (COLT)*, 2018.

Jacques, L. and Cambareri, V. Time for dithering: Fast and quantized random embeddings via the restricted isometry property. *Inf. Inference*, 6(4):441–476, 2017.

Jacques, L. and Vleeschouwer, C. D. Quantized iterative hard thresholding: Bridging 1-bit and high-resolution quantized compressed sensing. In *Int. Conf. Samp. Theory Apps. (SampTA)*, 2013.

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 Trans. Inf. Theory*, 59(4):2082–2102, 2013.

Kamath, A., Karmalkar, S., and Price, E. Lower bounds for compressed sensing with generative models. *https://arxiv.org/abs/1912.02938*, 2019.

Knudson, K., Saab, R., and Ward, R. One-bit compressive sensing with norm estimation. *IEEE Trans. Inf. Theory*, 62(5):2748–2758, 2016.

Laska, J. N., Wen, Z., Yin, W., and Baraniuk, R. G. Trust, but verify: Fast and accurate signal recovery from 1-bit compressive measurements. *IEEE Trans. Sig. Proc.*, 59 (11):5289–5301, 2011.

Latorre, F., Eftekhari, A., and Cevher, V. Fast and provable ADMM for learning with generative priors. In *Conf. Neur. Inf. Proc. Sys. (NeurIPS)*, pp. 12027–12039, 2019.

Li, Z., Xu, W., Zhang, X., and Lin, J. A survey on one-bit compressed sensing: Theory and applications. *Front. Comput. Sci.*, 12(2):217–230, 2018.

Liu, Z. and Scarlett, J. The generalized lasso with nonlinear observations and generative priors. *https://arxiv.org/abs/2006.12415*, 2020a.

Liu, Z. and Scarlett, J. Information-theoretic lower bounds for compressive sensing with generative models. *IEEE J. Sel. Areas Inf. Theory*, 1(1):292–303, 2020b.

Oymak, S. and Recht, B. Near-optimal bounds for binary embeddings of arbitrary sets. *https://arxiv.org/abs/1512.04433*, 2015.

Peng, P., Jalali, S., and Yuan, X. Solving inverse problems via auto-encoders. *IEEE J. Sel. Areas Inf. Theory*, 1(1): 312–323, 2020.

Plan, Y. and Vershynin, R. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Trans. Inf. Theory*, 59(1):482–494, 2012.

Plan, Y. and Vershynin, R. One-bit compressed sensing by linear programming. *Comm. Pure Appl. Math.*, 66(8): 1275–1297, 2013.

Qiu, S., Wei, X., and Yang, Z. Robust one-bit recovery via ReLU generative networks: Improved statistical rates and global landscape analysis. *https://arxiv.org/abs/1908.05368*, 2019.

Scarlett, J. and Cevher, V. Limits on support recovery with probabilistic models: An information-theoretic framework. *IEEE Trans. Inf. Theory*, 63(1):593–620, 2017.

Shah, V. and Hegde, C. Solving linear inverse problems using GAN priors: An algorithm with provable guarantees. In *IEEE Int. Conf. Acoust. Sp. Sig. Proc. (ICASSP)*, pp. 4609–4613, 2018.

Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. Series B*, pp. 267–288, 1996.

Vempala, S. S. *The random projection method*, volume 65. American Mathematical Soc., 2005.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *https://arxiv.org/abs/1011.3027*, 2010.

Wainwright, M. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inf. Theory*, 55(12):5728–5741, Dec. 2009a. ISSN 0018-9448.

Wainwright, M. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory*, 55(5): 2183–2202, May 2009b.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wen, J., Zhou, Z., Wang, J., Tang, X., and Mo, Q. A sharp condition for exact support recovery with orthogonal matching pursuit. *IEEE Trans. Sig. Proc.*, 65(6): 1370–1382, 2016.

Xu, C. and Jacques, L. Quantized compressive sensing with RIP matrices: The benefit of dithering. *Inf. Inference*, 2019.

Zhang, L., Yi, J., and Jin, R. Efficient algorithms for robust one-bit compressive sensing. In *Int. Conf. Mach. Learn. (ICML)*, pp. 820–828, 2014.

Zhu, R. and Gu, Q. Towards a lower sample complexity for robust one-bit compressed sensing. In *Int. Conf. Mach. Learn. (ICML)*, pp. 739–747, 2015.

Zymnis, A., Boyd, S., and Candes, E. Compressed sensing with quantized measurements. *IEEE Sig. Proc. Lett.*, 17 (2):149–152, 2009.